

Inter-Model Agreement in AI-Augmented Process Classification: A Multi-LLM Empirical Study

JIANFENG OU

March 2026

Chapter 1

On the Limits of Large Language Model Agreement in Structured Domain Assessment: A Multi-Model Study of AI Impact on Business Process Taxonomies

March 2026

Author: JIANFENG OU

Copyright 2026 JIANFENG OU. All rights reserved.

1.1 Abstract

Large language models (LLMs) are increasingly deployed as automated assessors for structured evaluation tasks, yet the reliability of multi-model assessment—particularly whether different frontier models reach concordant judgments on the same domain—remains poorly understood. This dissertation presents the first large-scale, controlled study of inter-model agreement in structured domain assessment. Four frontier LLMs—Gemini 2.5 Flash (Google DeepMind), DeepSeek V3.2 (DeepSeek AI), Qwen3 235B (Alibaba Cloud), and GPT-5 mini (OpenAI)—independently assessed 2,325 business process nodes drawn from four classification frameworks (APQC PCF 7.4, ITIL 4, SCOR 12.0, and AI-era extensions) across nine structured dimensions yielding 23 assessment fields (5 categorical, 18 numerical), producing 213,900 individual judgments.

Agreement is alarmingly low by conventional standards. Mean pairwise Cohen’s kappa = 0.078 (bootstrap 95% CI [0.047, 0.110]), indicating agreement barely above

chance. Four-rater Fleiss' kappa = 0.032, consistent with "slight" to "none" on the Landis and Koch (1977) scale. Yet this conclusion is metric-dependent: Gwet's AC1 = 0.587, a value in the "moderate" range, and majority consensus (at least 3 of 4 models agreeing) resolves 80.3% of field-judgments, though Monte Carlo simulation reveals that excess agreement over the random baseline is only 3.1 percentage points. A pronounced Kappa Paradox emerges: for boundary_current_type, AC1 = 0.889 ("almost perfect") while Fleiss' kappa = 0.041 ("slight"), demonstrating that the answer to "do LLMs agree?" depends fundamentally on the metric chosen (Observation 1: Agreement Indeterminacy).

A mixed effects variance decomposition via the Structured Disagreement Analysis Framework (SDAF) partitions total disagreement into three components: model bias ($D_{\text{bias}} = 17.2\%$), node-level ambiguity ($D_{\text{ambiguity}} = 9.1\%$), and residual model-by-node interaction ($D_{\text{residual}} = 73.6\%$). The dominance of residual variance establishes that most disagreement is idiosyncratic rather than systematically correctable—a finding with immediate practical consequences for organizations deploying LLMs as assessors. Each model exhibits a distinctive "bias fingerprint" that persists across all domains: Gemini assigns the most "changed" labels (23.7%); GPT-5 mini collapses assessments to a single dominant category (97.5% "will change"); DeepSeek produces the most differentiated distributions.

Extended analyses strengthen the robustness of these findings. Cluster bootstrap, resampling at the L2 process-group level to respect hierarchical dependence, yields a tighter confidence interval for mean kappa: [0.064, 0.091]. A cross-validated logistic regression predicting four-way agreement achieves AUC = 0.877, eliminating circularity concerns from the in-sample AUC = 0.923 reported in the primary analysis. Model-by-framework interaction analysis reveals that D_{bias} is not uniform: it ranges from 4.3 percentage points (SCOR) to 23.6 percentage points (AI-era extensions), indicating that model calibration differences are amplified in novel domains where training data is sparse.

This dissertation makes three primary contributions. First, it establishes the empirical magnitude and structure of multi-model disagreement at unprecedented scale, demonstrating that agreement is simultaneously low (by kappa), moderate (by AC1), and recoverable (by majority vote)—a phenomenon we term the Assessment Reliability Boundary. Second, it applies standard mixed effects ANOVA in a novel context—multi-model LLM agreement—via the SDAF organizational framework, decomposing disagreement into actionable components (correctable bias vs. inherent ambiguity vs. irreducible residual) and establishing field-specific reliability ceilings. Third, it demonstrates that model disagreement, properly analyzed, serves as an informative signal revealing inherent ambiguity in the assessment domain itself—each model's "dissent" on a given node carries diagnostic value proportional to its deviation from that model's

baseline tendency (Counter-Bias Credibility).

Keywords: large language models, inter-rater reliability, business process classification, AI impact assessment, multi-model consensus, structured disagreement analysis framework

Chapter 2

Chapter 1: Introduction

2.1 1.1 Research Background

Large language models are increasingly deployed as automated assessors in structured domains—classifying risks, scoring compliance, rating impact levels, and prioritizing strategic interventions. In business process management, organizations have begun using frontier LLMs to evaluate how artificial intelligence will transform individual processes, judgments that inform investment decisions and workforce planning at enterprise scale. The implicit assumption underlying such deployments is that LLM assessments are sufficiently reliable to serve as decision inputs. Yet this assumption has received remarkably little empirical scrutiny, particularly when multiple models are involved.

The reliability question becomes acute when one considers the diversity of the current LLM landscape. Frontier models differ in training data, architecture, alignment procedures, and organizational origin. If these differences produce systematically divergent assessments on the same structured task, then the choice of which model to deploy may be more consequential than the underlying data itself. This is not a hypothetical concern: the study presented in this dissertation reveals that four frontier LLMs, given identical inputs and instructions, achieve a mean Cohen’s κ of only 0.078 (95% CI [0.047, 0.110]) on structured process assessment—a level conventionally classified as “slight” agreement, barely distinguishable from chance.

The domain of business process classification offers a particularly suitable testbed for studying multi-model reliability. Established frameworks—the APQC Process Classification Framework (PCF) version 7.4, ITIL 4, and SCOR 12.0—provide thousands of well-defined process nodes organized in hierarchical taxonomies. These nodes span a wide range of domains, abstraction levels, and AI susceptibility profiles, creating natural variation in assessment difficulty. The structured, multi-dimensional nature of the assessment task (5 categorical fields, 18 numerical scores per node) generates sufficient

data to decompose disagreement into identifiable sources rather than reporting it as a single aggregate statistic.

This dissertation presents a large-scale, controlled study of inter-model assessment reliability. Four frontier LLMs—Gemini 2.5 Flash, DeepSeek V3.2, Qwen3 235B, and GPT-5 mini—independently assessed 2,325 business process nodes across 23 assessment fields, producing 213,900 individual judgments. The study employs a fully independent assessment design in which no model has access to any other model’s responses, eliminating inter-model contamination as a confound. The resulting dataset enables not only the measurement of agreement but also the systematic decomposition of disagreement into model-level bias, node-level ambiguity, and residual interaction effects.

2.2 1.2 Research Questions

This dissertation addresses four primary research questions:

RQ1 (Agreement Level): What level of inter-rater agreement do frontier LLMs achieve on structured process assessment?

Prior work on LLM-as-Judge tasks has reported agreement levels ranging from near-chance to near-perfect, but these studies overwhelmingly examine simple classification tasks (binary or ternary) with well-defined categories. No prior study has measured four-way inter-model agreement on a multi-dimensional structured assessment task at scale. This question asks whether the optimistic agreement levels reported for simple tasks generalize to complex, real-world assessment scenarios.

RQ2 (Variation Across Dimensions): How does agreement vary across assessment dimensions, process domains, and hierarchy levels?

If agreement were uniformly low across all dimensions and domains, the finding would be straightforward. Preliminary analysis suggests otherwise: some dimensions yield near-perfect agreement under certain metrics while simultaneously showing near-zero agreement under others (the Kappa Paradox). This question investigates the structural determinants of agreement variation, including the distinction between observable and subjective assessment dimensions, the effect of domain novelty, and the role of hierarchy depth.

RQ3 (Bias vs. Ambiguity): Can systematic model bias be separated from task-inherent ambiguity?

Low agreement, by itself, is diagnostically uninformative—it could reflect model incompetence, task ambiguity, or both. This question asks whether the sources of disagreement can be decomposed into identifiable components. If most disagreement arises from systematic model biases, it is potentially correctable through calibration. If it arises from inherent task ambiguity, it represents a fundamental limit on assess-

ment reliability regardless of model quality. The Structured Disagreement Analysis Framework (SDAF) developed in this dissertation addresses this decomposition.

RQ4 (Consensus Utility): Can multi-model consensus recover practical utility despite low pairwise agreement?

If individual model pairs agree only slightly above chance, a natural question is whether aggregation across multiple models can nonetheless produce useful assessments. This question examines whether majority-vote consensus (≥ 3 of 4 models agreeing) yields reliable classifications, how much the observed consensus rate exceeds a random baseline preserving marginal distributions, and whether consensus conditions are themselves predictable from node-level features.

2.3 1.3 Research Scope and Limitations

Scope. The study encompasses 2,325 process nodes drawn from four classification frameworks: APQC PCF 7.4 (1,921 nodes), ITIL 4 (141 nodes), SCOR 12.0 (164 nodes), and AI-era extensions (99 nodes). All four models assessed all nodes using an identical structured prompt (version 2.2) with temperature set to 0.0 for deterministic output. The analysis covers inter-rater reliability (Cohen’s κ , Fleiss’ κ , Gwet’s AC1, ICC, Krippendorff’s α), variance decomposition (mixed effects models), information-theoretic measures (entropy, mutual information), and predictive modeling (logistic regression with cross-validation).

Conflict of interest disclosure. Of the 2,325 process nodes assessed, 99 belong to the “AI-era extensions” category. These nodes—covering AI governance, ML operations, and data ethics—were created by the dissertation author as part of the O’Process Framework development. This creates a potential conflict of interest: the author has domain-specific knowledge and design intent for these nodes that may influence their interpretation. Several mitigations partially address this concern. First, the AI-era nodes constitute only 4.3% of the total dataset; removing them entirely does not materially alter the aggregate findings. Second, the models assessed all nodes blindly with no indication of authorship or provenance. Third, the analysis reports results both for the full dataset and with breakdowns by source framework, allowing readers to evaluate whether AI-era nodes behave differently. Nevertheless, readers should be aware that the author’s dual role as framework designer and dissertation analyst introduces a form of non-independence that cannot be fully eliminated through statistical controls.

Limitations. Several limitations constrain the generalizability of the findings. (1) The study examines a single assessment domain (AI impact on business processes); cross-domain generalization requires independent replication. (2) All assessments used a single prompt version; prompt sensitivity analysis (beyond the v2.0 to v2.2 iteration history) was not systematically conducted. (3) The four models, while represent-

ing diverse providers, share architectural similarities (all employ Mixture of Experts architectures) that may limit the range of observable disagreement. (4) Temperature was fixed at 0.0 to maximize reproducibility, which suppresses intra-model variability that would be present in typical deployment settings. (5) No human expert ground truth was established for the full 2,325-node dataset, limiting the ability to adjudicate which model’s assessment is “correct” in any given case.

2.4 1.4 Dissertation Structure

The dissertation is organized as follows.

Chapter 2 reviews three intersecting research traditions: process classification frameworks, LLM evaluation and inter-rater reliability, and the treatment of annotation disagreement as information rather than noise. The review identifies seven specific research gaps that this dissertation addresses.

Chapter 3 describes the research design: the empirical domain (O’Process Framework integrating APQC PCF 7.4, ITIL 4, SCOR 12.0, and AI-era extensions), the assessment instrument, model selection rationale, the role of SDAF as an analytical lens applying mixed effects variance decomposition to multi-model agreement, and the statistical methods that operationalize the four research questions.

Chapter 4 details the data collection protocol that produced 213,900 individual judgments from four frontier LLMs, including operational parameters, quality controls, missing data handling, and ethical considerations.

Chapter 5 presents the empirical results, including distribution analysis, pairwise and four-way agreement (Cohen’s κ , Fleiss’ κ , Gwet’s AC1), SDAF variance decomposition ($D_{bias} = 17.2\%$, $D_{ambiguity} = 9.1\%$, $D_{residual} = 73.6\%$), consensus analysis, cluster bootstrap, cross-validated AUC, ICC for numerical dimensions, and the reliability corridor.

Chapter 6 discusses implications, situates the findings within prior work, formalizes the Agreement Indeterminacy observation and Reliability Ceiling, examines the bias fingerprint and counter-bias signal, offers practical recommendations for multi-model deployment, and identifies limitations.

Chapter 7 provides concluding answers to each research question, summarizes contributions, and identifies directions for future research.

Chapter 3

Chapter 2: Literature Review

This chapter reviews three intersecting research traditions that collectively motivate and contextualize the dissertation. Section 2.1 surveys the process classification frameworks that define the assessment domain. Section 2.2 reviews the rapidly growing literature on LLMs as automated assessors, with emphasis on inter-model reliability. Section 2.3 examines the Kappa Paradox—the phenomenon whereby standard agreement metrics can yield contradictory conclusions—tracing its intellectual history from Feinstein and Cicchetti (1990) through Gwet (2008). Section 2.4 reviews the computational linguistics tradition that reframes annotation disagreement as information. Section 2.5 discusses multi-model bias and the limits of calibration. Section 2.6 identifies the research gap that this dissertation addresses.

3.1 2.1 Process Classification Frameworks

Business process classification frameworks provide standardized taxonomies for organizing enterprise activities. Three established frameworks and one emerging extension constitute the assessment domain for this study.

APQC PCF 7.4. The American Productivity and Quality Center’s Process Classification Framework, now in version 7.4, defines 13 top-level categories encompassing over 1,900 processes organized in a four-level hierarchy (APQC, 2023). The framework spans operating processes (categories 1–6, covering strategy through delivery) and management/support processes (categories 7–13, covering human capital through risk management). PCF has been adopted by over 600 organizations worldwide and represents the most comprehensive cross-industry process taxonomy available. Its breadth makes it particularly valuable for studying agreement variation across domains: processes range from highly standardized (e.g., “Process Accounts Payable”) to inherently ambiguous (e.g., “Manage Innovation”).

ITIL 4. The Information Technology Infrastructure Library version 4 (AXELOS,

2019) provides 141 practice-level processes focused on IT service management. ITIL processes tend to be more technically specific than PCF processes, with clearer operational definitions—a characteristic that, per the Observability Hypothesis, should yield higher inter-model agreement.

SCOR 12.0. The Supply Chain Operations Reference model version 12.0 (APICS, 2017) defines 164 supply chain processes organized around Plan, Source, Make, Deliver, Return, and Enable categories. SCOR processes are notable for their quantitative orientation, with built-in performance metrics that provide concrete assessment anchors.

AI-era extensions. Ninety-nine additional nodes were developed as part of the O’Process Framework to cover domains absent from the three established frameworks: AI governance, machine learning operations, data ethics, and algorithmic auditing. These nodes represent emerging process categories for which no established consensus taxonomy exists. As disclosed in Section 1.3, these nodes were created by the dissertation author, introducing a potential conflict of interest that readers should consider when interpreting framework-specific results.

Together, these four sources yield 2,325 process nodes spanning a wide range of domain maturity, operational specificity, and AI susceptibility—providing natural variation for studying how these factors influence inter-model assessment agreement.

3.2 2.2 LLM Evaluation and Inter-Rater Reliability

The paradigm of using large language models as automated evaluators crystallized in 2023. Zheng et al. (2023) introduced the foundational LLM-as-a-Judge framework alongside MT-Bench and Chatbot Arena, demonstrating that GPT-4 judgments achieve over 80% agreement with human preferences on open-ended text quality assessment. Gilardi et al. (2023) showed that ChatGPT outperforms crowd workers on text-annotation tasks, achieving higher inter-coder agreement at approximately one-twentieth the cost. These studies established the basic viability of LLM-based evaluation but shared a critical limitation: they evaluated single models against human baselines rather than examining whether different LLMs would agree with each other.

The transition from single-model evaluation to inter-model agreement represents a fundamental shift. Chen et al. (2025) reported $\kappa > 0.80$ across Gemini, GPT-4o, and Claude for thematic analysis of qualitative research data, but this result applies to bounded classification with well-defined coding schemes. Alizadeh et al. (2025) provided the first large-scale comparison of LLMs and human annotators in latent content analysis, finding that inter-model agreement is strongly task-dependent. Chandra et al. (2025) found that Claude approaches perfect reliability in structured rubric-based writing assessment—a result that reinforces the pattern that agreement scales with task

definitional precision.

The contrast between reported agreement levels is instructive. Prior studies examining simple classification (binary, ternary, or preference ranking) with clear category definitions report κ values ranging from 0.60 to above 0.80. The present study, examining multi-dimensional structured assessment with 23 fields across 2,325 items, finds $\kappa = 0.078$. This is not contradictory: it reflects the well-established principle that inter-rater reliability decreases as task complexity and the number of assessment dimensions increase.

Several recent studies have identified systematic biases in LLM evaluation. Li et al. (2025) provide a comprehensive taxonomy of twelve bias types, including self-enhancement bias, position bias, and verbosity bias. Deldjoo et al. (2025) identified “agreeableness bias” in LLM judge panels, where models converge toward consensus when they can see each other’s responses. The fully independent assessment design used in this study eliminates this particular confound, making the observed low agreement more trustworthy precisely because it cannot be inflated by inter-model contamination.

Chehbouni et al. (2025), in a NeurIPS position paper titled “Neither Valid nor Reliable?”, challenged the assumption that LLMs can serve as reliable judges, drawing on measurement theory to identify four core assumptions underlying LLM-as-Judge that may be undermined by inherent model limitations. Their skeptical perspective provides important context for interpreting the low agreement levels reported here: the findings may reflect genuine limitations of LLM-based assessment rather than an idiosyncrasy of the particular task or models studied.

3.3 2.3 The Kappa Paradox

The measurement of inter-rater agreement using chance-corrected statistics has a long history, beginning with Scott’s π (1955) and Cohen’s κ (1960). Fleiss (1971) extended κ to the multi-rater case. Landis and Koch (1977) provided the interpretive benchmarks—“slight” ($\kappa < 0.20$), “fair” (0.21–0.40), “moderate” (0.41–0.60), “substantial” (0.61–0.80), “almost perfect” (> 0.80)—that have dominated agreement research for nearly five decades, despite never being empirically validated against decision-making outcomes.

The intellectual watershed came in 1990, when Feinstein and Cicchetti published their landmark two-part paper identifying two distinct paradoxes. The *First Paradox*: high observed agreement can coexist with low κ when one category dominates the marginal distribution. This occurs because κ ’s denominator ($1 - p_e$) shrinks as marginal prevalence becomes extreme, making it increasingly difficult to exceed chance-level agreement. The *Second Paradox*: unequal marginal distributions between raters can

depress κ even when raw agreement is acceptable. Byrt et al. (1993) proposed the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) as a response, though its aggressive adjustment can obscure genuine distributional information.

Gwet (2008) introduced the AC1 coefficient, modeling chance agreement as arising from a mixture of deterministic and random responses—a formulation robust to prevalence imbalance. When most items fall into a single dominant category, AC1’s expected agreement does not inflate as dramatically as κ ’s, preserving the ability to detect genuine agreement in skewed settings. Gwet’s (2014) comprehensive handbook established the statistical properties of the AC1/AC2 family while acknowledging that AC1 and κ answer fundamentally different questions about agreement.

The data presented in this dissertation provide a dramatic empirical instance of the Kappa Paradox. For the `boundary_current_type` field, Gwet’s $AC1 = 0.889$ (“almost perfect” agreement) while Fleiss’ $\kappa = 0.041$ (“slight” agreement). This 20-fold divergence means that the answer to the question “do these models agree on boundary classification?” depends entirely on which metric one consults. The paradox arises because `boundary_current_type` has an extremely skewed marginal distribution, with one category dominating across all four models. Under κ , this skewness inflates expected chance agreement to a level that nearly consumes observed agreement. Under AC1, the same skewness is accommodated by a different model of chance behavior.

This is not a mere statistical curiosity. If a practitioner were to use Fleiss’ κ alone to evaluate the reliability of multi-model boundary classification, they would conclude that the models are essentially random. If they used AC1 alone, they would conclude that the models are highly reliable. Neither conclusion is wrong; they answer different questions. The formal characterization of the conditions under which this divergence arises, and its severity as a function of distributional skewness, is developed in Chapter 6.

3.4 2.4 Annotation Disagreement as Signal

A parallel tradition in computational linguistics has progressively reframed annotation disagreement from noise to information, providing intellectual foundations for the approach taken in this dissertation.

Artstein and Poesio (2008) developed the definitive treatment of inter-annotator agreement in computational linguistics, demonstrating that disagreement can arise from three sources: ambiguity in the annotation scheme, annotator error, and inherent vagueness in the underlying construct. Their argument that “disagreement can be informative” is a direct intellectual predecessor of the present work, though they applied it qualitatively to human annotators rather than formally to LLM raters.

Plank et al. (2014) challenged the assumption that annotator disagreement always reflects error, arguing that some items are “linguistically debatable” rather than simply miscategorized. Their distinction between genuine ambiguity and annotator incompetence maps onto the SDAF’s separation of $D_{ambiguity}$ from D_{bias} . Pavlick and Kwiatkowski (2019) demonstrated that for many NLP tasks, the distribution of human judgments—not just the majority label—carries meaningful information about item properties such as difficulty, ambiguity, and context-dependence. This perspective motivates treating the full distribution of model assessments, rather than just their modal value, as the primary object of analysis.

Uma et al. (2021), in a comprehensive JAIR survey titled “Learning from Disagreement,” established the “perspectivist” view of annotation as a mainstream position: multiple valid interpretations of the same item are not noise but information about the item’s properties. Basile et al. (2021) argued that evaluation practices predicated on a single correct answer are fundamentally flawed when multiple valid answers exist.

The present study extends this tradition from human annotators to LLM raters. The finding that majority consensus (≥ 3 of 4 models agreeing) reaches 80.3% of nodes, while full consensus covers only 1.8% (42 nodes), suggests that most process nodes admit multiple defensible assessments—a pattern more consistent with genuine task ambiguity than with model failure. However, this interpretation requires caution: it is also consistent with all four models being unreliable in correlated ways. The SDAF decomposition (Section 5.5) attempts to distinguish between these explanations by separating systematic bias ($D_{bias} = 17.2\%$) from node-level ambiguity ($D_{ambiguity} = 9.1\%$) and residual interaction ($D_{residual} = 73.6\%$).

The dominance of the residual component (73.6%) is itself a finding that resists easy interpretation. It could reflect genuine idiosyncratic model-node interactions, measurement noise from the assessment instrument, or latent structure not captured by the two-facet model. The honest conclusion is that most disagreement among frontier LLMs on structured process assessment remains unexplained by the decomposition framework employed here.

3.5 2.5 Multi-Model Bias and Variance Decomposition

Each of the four models in this study exhibits a distinctive “bias fingerprint”—a consistent pattern of assessment tendencies that persists across process domains. Understanding whether these biases are correctable is central to the practical implications of the work.

The calibration literature provides partial guidance. Zhao et al. (2021) demonstrated that systematic biases in GPT-3 can be partially corrected through contextual calibration, but their approach operates at the label distribution level, leaving item-specific

biases uncorrected. Kadavath et al. (2022) found that models can express meaningful uncertainty about their outputs, but this self-knowledge is imperfect. Xiong et al. (2024) showed that significant miscalibration persists, particularly for tasks outside the training distribution—a relevant concern given that AI impact assessment of business processes is unlikely to be well-represented in any model’s training data.

The statistical framework for decomposing disagreement into identifiable sources draws on Generalizability Theory (Cronbach et al., 1972; Brennan, 2001), which decomposes observed score variance into components attributable to items, raters, and their interactions. The mixed effects model used in this dissertation ($Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, where α_i represents node effects and β_j represents model effects) yields a three-component decomposition: $D_{bias} = 17.2\%$ (systematic model differences), $D_{ambiguity} = 9.1\%$ (stable node-level difficulty), and $D_{residual} = 73.6\%$ (idiosyncratic model-node interactions).

The practical implication is sobering. Even if all systematic model bias could be perfectly calibrated away, and all assessment instrument ambiguity could be resolved through improved prompt design, the maximum agreement improvement would be bounded by $D_{bias} + D_{ambiguity} = 26.3\%$. The remaining 73.6% of disagreement appears to be irreducibly idiosyncratic—arising from specific model-node interactions that cannot be predicted from either the model’s general bias profile or the node’s average difficulty. Cross-validated logistic regression predicting four-way agreement from node features achieves $AUC = 0.877 \pm 0.019$ (excluding response diversity as a predictor to avoid circularity), confirming that structural features explain substantial variance in agreement but leave considerable residual unpredictability.

This finding echoes results from other structured assessment domains. Berg et al. (2022), in their study of ESG rating divergence, found that the average pairwise correlation among six major ESG rating providers is approximately $r = 0.54$, with substantial idiosyncratic disagreement that persists after controlling for scope and methodology differences. The parallel is imperfect—ESG raters are human organizations with economic incentives, not LLMs—but the structural similarity (systematic bias + inherent ambiguity + large residual) suggests that the pattern may be a general property of multi-rater structured assessment rather than an artifact specific to LLMs.

3.6 2.6 Research Gap

The intersection of these literatures reveals a specific gap: no large-scale, controlled study of multi-model LLM reliability on structured assessment tasks exists.

Prior LLM-as-Judge studies overwhelmingly examine simple classification tasks (binary or ternary) with well-defined categories and report agreement against human baselines or between pairs of models. No study has examined four-way inter-model

agreement on a multi-dimensional assessment task (23 fields) across thousands of items (2,325 nodes), let alone decomposed the resulting disagreement into identifiable sources.

The Kappa Paradox, first identified by Feinstein and Cicchetti (1990) and addressed through alternative metrics by Gwet (2008), has been extensively documented in clinical and social science settings but never examined in the context of LLM-based assessment. The extreme marginal skewness characteristic of LLM categorical outputs makes the paradox particularly acute in this setting—as the `boundary_current_type` example ($AC1 = 0.889$ vs. Fleiss’ $\kappa = 0.041$) dramatically illustrates.

The annotation disagreement literature has established the principle that disagreement can be informative, but this principle has not been formally applied to multi-model LLM assessment or operationalized through a variance decomposition framework that quantifies the relative contributions of bias, ambiguity, and residual effects.

The variance decomposition literature from Generalizability Theory (Cronbach et al., 1972; Brennan, 2001) provides well-established statistical machinery for partitioning rater disagreement, but this standard methodology has not been applied to LLM inter-model reliability, and no study has derived actionable reliability ceilings from such decompositions in the LLM evaluation context.

This dissertation addresses these gaps through 213,900 judgments from four frontier LLMs, analyzed via a multi-metric reliability framework, mixed effects variance decomposition, and the SDAF organizational protocol that applies standard variance decomposition to the diagnostic interpretation of multi-model disagreement. The core findings—mean $\kappa = 0.078$, cluster bootstrap CI [0.064, 0.091] corrected for nested dependency, Fleiss’ $\kappa = 0.032$, $AC1 = 0.587$, majority consensus at 80.3%, 42 full-consensus nodes (1.8%), 215 hard nodes, and a variance decomposition of $D_{bias} = 17.2\%$, $D_{ambiguity} = 9.1\%$, $D_{residual} = 73.6\%$ —collectively establish that multi-model LLM agreement on structured assessment is low, structurally patterned, and only partially amenable to improvement through calibration or instrument redesign.

Chapter 4

Chapter 3: Research Design

This chapter describes the empirical domain, assessment instrument, model selection rationale, and analytical strategy that collectively operationalize the four research questions posed in Chapter 1. The design pursues a single overarching objective: to measure and characterize the agreement structure among four frontier large language models performing the same structured assessment task under identical conditions. The chapter is organized around five design decisions: the choice of empirical domain (Section 3.1), the construction of the assessment instrument (Section 3.2), the selection and limitations of the model panel (Section 3.3), the role of SDAF as an analytical lens (Section 3.4), and the statistical methods that operationalize the four research questions (Section 3.5).

4.1 3.1 The O’Process Framework (OPF)

The empirical domain is the O’Process Framework (OPF), an AI-native process classification system that unifies four established source taxonomies into a single hierarchical structure of 2,325 process nodes. The framework was selected for this study because it satisfies four requirements of a suitable empirical domain for multi-model agreement research: (a) sufficient scale to support statistical analysis with adequate power, (b) hierarchical structure providing natural stratification variables, (c) multi-source composition enabling cross-domain comparisons, and (d) a mix of well-established and novel process categories enabling tests of domain familiarity effects.

Table 3.1: OPF Source Composition

Source Framework	Nodes	Domain Coverage	Provenance
APQC PCF 7.4	1,921	Cross-industry business processes	APQC (2023)
ITIL 4	141	IT service management	AXELOS (2019)

Source Framework	Nodes	Domain Coverage	Provenance
SCOR 12.0	164	Supply chain operations	ASCM (2017)
AI-era extensions	99	AI governance, ML operations, data ethics	Author-created
Total	2,325		

The framework organizes processes into five hierarchy levels with decreasing granularity:

Table 3.2: Hierarchy Level Distribution

Level	Count	Description	Example
L1	13	Operating/management categories	“Develop and Manage Human Capital”
L2	98	Process groups	“Manage Employee Onboarding”
L3	502	Processes	“Manage Employee Orientation”
L4	1,523	Activities	“Provide Orientation Materials”
L5	189	Tasks (select branches)	“Distribute Employee Handbook”

The hierarchical structure serves a dual analytical purpose. First, it provides a natural stratification variable for examining whether assessment granularity affects inter-model agreement. Second, because nodes at deeper levels describe increasingly specific activities, the hierarchy creates a gradient from abstract (L1–L2) to concrete (L4–L5) process descriptions, enabling tests of the Observability Hypothesis—whether more concrete, observable processes elicit higher model agreement.

The four source frameworks contribute complementary domain perspectives. APQC PCF 7.4 provides the broadest coverage, spanning 13 top-level categories from strategy development through environmental management, and accounts for 82.6% of total nodes. Its well-established taxonomy—adopted by over 600 organizations globally—provides a stable baseline against which model agreement in less standardized domains can be compared. ITIL 4 and SCOR 12.0 contribute domain-specific process vocabularies for IT service management and supply chain operations, respec-

tively. These frameworks use distinct terminological conventions and process decomposition logics, introducing controlled variation in node description style that tests whether models’ agreement patterns depend on the source vocabulary. The AI-era extensions cover processes that emerged after the three established frameworks were finalized, including algorithmic audit, synthetic data governance, and responsible AI deployment—processes for which no industry-standard taxonomy exists.

Each node carries bilingual labels (Chinese and English), a domain classification (operating or management/support), and source provenance metadata. The bilingual labeling is relevant because the assessment prompt (Section 3.2) is written in Chinese; models encounter Chinese process names during assessment, and their familiarity with Chinese business terminology may influence scoring behavior.

Why business process classification? The choice of business process classification as the empirical domain is motivated by three characteristics that make it particularly suitable for studying LLM agreement. First, process classification involves structured judgment—assigning categorical labels and numerical scores according to defined criteria—which falls squarely within the “LLM-as-Judge” paradigm studied in the literature (Section 2.1). Second, the domain requires both factual knowledge (understanding what a process involves) and evaluative judgment (assessing how AI will transform it), creating a natural spectrum from objective to subjective assessment that maps onto the Observability Hypothesis. Third, the domain is large enough (2,325 nodes) to support the statistical analyses required for variance decomposition, yet homogeneous enough that all nodes can be assessed with the same instrument, controlling for task-type variation that would confound cross-domain comparisons.

Conflict of interest disclosure. The 99 AI-era extension nodes were created by the dissertation author to cover emerging AI governance processes (e.g., model bias monitoring, algorithmic audit, synthetic data governance) not represented in the three established frameworks. These nodes constitute 4.3% of the total corpus. While their inclusion extends OPF’s coverage to AI-native processes, it introduces a potential conflict of interest: the author who designed the assessment instrument also created 99 of its subjects. We address this through stratified analysis (Section 5.4.3), reporting agreement metrics separately for AI-era nodes versus established-framework nodes, and flag results where AI-era nodes exhibit anomalous patterns. Readers should interpret AI-era results with appropriate caution.

Framework validity. The three established frameworks (APQC PCF, ITIL 4, SCOR 12.0) have been validated through decades of industry adoption and academic study. APQC PCF, in particular, has been maintained through multiple revision cycles since the 1990s and is used as a benchmarking standard by hundreds of organizations across industries. ITIL 4 is the globally recognized framework for IT service management, endorsed by governmental and private-sector IT organizations. SCOR 12.0 is the stan-

standard reference model for supply chain operations, maintained by the Association for Supply Chain Management (ASCM). The combination of these three validated frameworks with the author-created AI-era extensions creates a natural experiment: we can test whether model agreement patterns differ between independently validated content (96% of nodes) and author-created content (4%), providing an internal validity check on the assessment instrument.

4.2 3.2 Assessment Instrument

Each process node was assessed across nine structured dimensions, yielding 23 assessment fields organized into two groups. The instrument design reflects a deliberate balance between assessment breadth (covering multiple facets of AI impact) and response burden (keeping each assessment tractable for a single model call). The 23-field structure was iteratively refined through three prompt versions (v2.0, v2.1, v2.2); the final v2.2 instrument is used for all data collection reported in this dissertation.

Categorical fields (5 fields). These represent discrete classifications that commit the model to a definitive judgment:

Field	Dimension	Categories	Scale
change_status	D2: Change Status	3	{changed, will change, stable}
penetration_current	D1: Penetration Level	3	{high, medium, low}
uncertainty_offset	D5: Confidence	3	{high, medium, low}
boundary_current_type	D4: Boundary Type	4	{Type 1: purely human, Type 2: human-AI, Type 3: AI-led, Type 4: fully automated}
summary_priority_flag	Priority Flag	3	{high priority, routine, low priority}

Numerical fields (18 fields, all scored on 1–5 integer scales). These are organized into five dimension groups:

- **D1 sub-scores** (3 items): decision replaceability, processing acceleration, tacit knowledge dependency — these measure the degree to which AI can substitute for, accelerate, or replicate tacit human knowledge in a given process
- **D3 sub-scores** (4 items): change nature types A through D — these classify the character of AI-driven change along four orthogonal axes (automation, augmentation, restructuring, and creation)
- **D7 sub-scores** (3 items): rule-driven degree, exception flexibility, feedback loop maturity — these assess the structural properties that determine a process’s amenability to AI intervention
- **D8 sub-scores** (4 items): data intensity, cross-process dependency, data standardization, integration barrier — these characterize the data ecosystem surrounding a process
- **D9 sub-scores** (4 items): data availability, tech maturity, implementation simplicity, value density — these assess implementation readiness from a practical deployment perspective

The 23-field design creates an intentional tension between categorical breadth and numerical depth. Categorical fields require models to commit to discrete classifications (e.g., “changed” vs. “will change”), where disagreement is binary and unambiguous. Numerical fields allow models to express finer gradations on a 1–5 scale, where disagreement may be ordinal (adjacent scores) or extreme (endpoints). This dual structure enables the study to examine whether models that disagree categorically nonetheless agree on relative rankings—a question addressed through weighted kappa and ICC analyses in Chapter 5.

The nine parent dimensions (D1–D5, D7–D9, and Priority) are not orthogonal by design. For example, a process with high AI penetration (D1) is likely to have undergone substantial change (D2) and to exhibit high data intensity (D8). These inter-dimensional correlations are not a design flaw but a feature: they enable analysis of cross-dimensional consistency within and between models. A model that assigns high penetration but low data intensity to the same process may be internally inconsistent—a signal that the model lacks domain coherence for that particular process. The bias fingerprinting analysis (Section 5.5) exploits these correlations to characterize each model’s internal assessment logic.

Prompt design (v2.2). All four models received an identical Chinese-language structured prompt containing: (a) explicit 1–5 anchor descriptions for each numerical sub-dimension, (b) ten calibration rules governing scoring behavior, and (c) a mandatory JSON output schema specifying field names, types, and valid value sets.

The prompt evolved through three versions (v2.0, v2.1, v2.2), with each iteration addressing calibration issues identified in pilot scans. The final v2.2 prompt incorporates ten calibration rules, three of which merit specific mention:

- **Rule 8 (scale utilization):** Instructed models to use the full 1–5 range rather than clustering around middle values. Pilot scans with v2.0 revealed that all models exhibited strong central tendency, assigning scores of 2–4 to over 90% of nodes. Rule 8 explicitly anchored the endpoints: “A score of 1 means the AI has essentially no impact; a score of 5 means the process is fundamentally transformed by AI.”
- **Rule 9 (logical consistency):** Imposed cross-field constraints on `change_status` to prevent logically contradictory assessments. For example, a process rated with high AI penetration (`penetration_overall` = “高”) cannot simultaneously be classified as “stable” (`change_status` = “稳定”), because high AI penetration implies active or imminent transformation.
- **Rule 10 (confidence calibration):** Required models to express genuine assessment uncertainty through the `uncertainty_confidence` field rather than defaulting to “high confidence.” The rule specified that “高” (high confidence) should be reserved for processes where the model has strong domain knowledge and the assessment criteria are unambiguous.

Prompt design limitations. Three design limitations constrain the conclusions that can be drawn from this study:

Single-prompt design. All four models received the same single prompt version (v2.2). This eliminates prompt variation as a confound but also means the study cannot distinguish model-level effects from model-by-prompt interaction effects. A fully crossed prompt-by-model design—using, for example, three prompt variants (minimal, structured, chain-of-thought) crossed with four models—would yield a 3 × 4 factorial design that could decompose variance into model main effects, prompt main effects, and model-by-prompt interactions. The absence of prompt variation means that any observed model differences may partly reflect differential sensitivity to the specific v2.2 prompt structure rather than fundamental model capabilities. This is the single most important limitation of the study design.

Chinese-language prompt. The prompt is written in Chinese, which may differentially affect models trained primarily on English-language corpora. Models with stronger Chinese-language capabilities (Qwen3, DeepSeek) may exhibit systematically different scoring patterns from those with English-primary training (Gemini, GPT-5 mini), confounding language proficiency with assessment reliability. An ideal design would include parallel Chinese and English prompt versions, but this would double the data collection cost and introduce translation-equivalence concerns.

No iterative prompting. Each model received each node exactly once, with no opportunity for clarification, revision, or self-correction. This single-pass design mirrors common deployment patterns but does not capture the potential improvement from multi-turn assessment protocols (e.g., chain-of-thought followed by revision, or multi-

turn dialogue with clarifying questions). Recent work on deliberative alignment suggests that iterative prompting can substantially improve assessment quality (Wei et al., 2022), meaning the single-pass agreement levels reported here may represent a lower bound on achievable multi-model consistency.

These three limitations—single prompt, single language, single pass—collectively mean that the study characterizes agreement under a specific, practically common deployment configuration rather than under optimized conditions. This is a deliberate design choice: the most policy-relevant question is how models perform under realistic deployment constraints, not how they perform under idealized conditions. These limitations are revisited in Section 7.8.

4.3 3.3 Model Selection and MoE Homogeneity Limitation

Four frontier LLMs were selected to span diverse training pipelines, organizational origins, and parameter scales:

Table 3.3: Model Specifications

Model	Provider	Origin	Active Parameters	Architecture	Prompt
Gemini 2.5 Flash	Google	US	~8B (MoE)	Mixture of Experts	v2.2
DeepSeek V3.2	DeepMind	CN	~37B of 671B (MoE)	Mixture of Experts	v2.2
Qwen3 235B	DeepSeek AI	CN	~22B (MoE)	Mixture of Experts	v2.2
GPT-5 mini	Alibaba Cloud	US	Undisclosed (MoE)	Mixture of Experts	v2.2

Selection criteria included: (a) frontier-class capability on standard benchmarks at the time of data collection, (b) geographic and organizational diversity (two US-origin, two CN-origin providers), providing variation in training data composition and cultural context, (c) API availability for programmatic assessment at scale with deterministic temperature settings, and (d) sufficient context window to accommodate the full structured prompt including all anchor descriptions and calibration rules. All models used identical prompt version v2.2, controlling for prompt variation.

The parameter scale variation across the panel is substantial: from Gemini’s approximately 8 billion active parameters to DeepSeek’s 37 billion active parameters drawn from a 671-billion-parameter total. This variation tests whether model scale correlates with assessment behavior—whether larger models produce systematically different (or more internally consistent) assessments than smaller ones.

Architecture homogeneity limitation. A significant constraint on the diversity of this model panel is that all four models employ Mixture-of-Experts (MoE) architectures, characterized by sparse activation patterns where only a subset of parameters is engaged for any given input. In MoE architectures, a gating network routes each input token to a small subset of “expert” sub-networks (typically 2 of 8–64 experts), so that only a fraction of total parameters are active for any given input. While the four models differ substantially in total parameter count (8B to 671B), number of experts, routing strategies, and training data composition, they share this fundamental architectural paradigm.

This homogeneity limits the strength of claims about “diverse model agreement” in three specific ways:

1. **Correlated routing failures:** MoE models may route ambiguous inputs to similar expert subsets, producing correlated errors that would not appear in dense transformer models where all parameters process every input.
2. **Sparse activation biases:** The sparse activation pattern may systematically under-represent certain input features that fall between expert specializations, creating shared blind spots across the panel.
3. **Training objective convergence:** All four models were trained with variants of next-token prediction on web-scale corpora, potentially encoding similar distributional priors about business process terminology regardless of architectural differences.

The observed inter-model agreement (or disagreement) should therefore be interpreted as agreement among MoE models specifically, not among LLMs in general. Future studies should include architecturally diverse model panels—incorporating dense transformers (e.g., Llama-class models), state-space models (e.g., Mamba variants), or hybrid architectures—to test whether the patterns reported here generalize beyond the MoE family. The degree to which MoE-specific agreement patterns differ from architecturally heterogeneous panels remains an open empirical question.

Despite this architectural homogeneity, meaningful variation exists along other dimensions. The four models differ in: (a) total parameter count by nearly two orders of magnitude, (b) training data composition (US-origin models likely overweight English web text; CN-origin models likely overweight Chinese web text and potentially government-curated datasets), (c) post-training alignment procedures (RLHF vs. DPO vs. proprietary methods), and (d) release date (spanning 2025–2026). These differences provide genuine variation in the factors that shape assessment behavior, even if the underlying architecture is shared.

The geographic diversity of the panel (two US-origin, two CN-origin providers) is particularly relevant for this study’s Chinese-language prompt design. If Chinese-language proficiency were a dominant factor in assessment quality, we would expect

systematic differences between the US-origin and CN-origin model pairs—a testable hypothesis examined in the pairwise agreement analysis (Section 5.4.1). The 2-by-2 geographic structure also enables a crude test of “cultural” training bias: if US-origin and CN-origin models disagree systematically on certain process categories, this may reflect differing institutional norms or business practices encoded in their respective training corpora.

Table 3.4: Model Panel Diversity Matrix

Variation Axis	Range Across Panel	Analytical Implication
Parameter count	~8B to 671B total	Tests scale-agreement relationship
Geographic origin	2 US, 2 CN	Tests cultural/linguistic bias
Training data	English-primary vs. Chinese-primary	Interacts with Chinese-language prompt
Release date	Late 2025 to early 2026	Controls for temporal knowledge cutoff
MoE configuration	Varying expert counts and routing	Within-architecture variation
Post-training	RLHF, DPO, proprietary	Tests alignment procedure effects

4.4 3.4 SDAF as Organizational Tool

The Structured Disagreement Analysis Framework (SDAF) provides the primary analytical lens for interpreting multi-model disagreement. It is essential to state clearly what SDAF is and is not.

SDAF is an **organizational tool**—a structured protocol for decomposing observed variance in multi-model assessments into interpretable components. It draws on the well-established tradition of variance decomposition in psychometrics (Cronbach et al., 1972) and applies it to the specific context of multi-LLM assessment. The novelty lies not in the statistical machinery—which is standard mixed effects modeling—but in the application to LLM disagreement and the practical implications drawn from the decomposition results.

SDAF employs a two-way mixed effects model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where Y_{ij} is the assessment score for node i by model j , μ is the grand mean, α_i is the random effect of node i (capturing node-level ambiguity), β_j is the fixed effect of

model j (capturing systematic model bias), and ε_{ij} is the residual interaction term. The variance decomposition partitions total disagreement into three components:

- D_{bias} (model effect): Systematic differences in scoring tendencies across models (17.2% of total variance)
- $D_{\text{ambiguity}}$ (node effect): Variation attributable to inherent difficulty or ambiguity of specific process nodes (9.1%)
- D_{residual} (interaction + error): Idiosyncratic model-by-node interactions that cannot be attributed to either main effect (73.6%)

SDAF is **not** a novel theoretical framework, a predictive model, or a claim about latent causal mechanisms. It is a diagnostic decomposition that answers a practical question: when four models disagree, how much of that disagreement is systematic (and therefore potentially correctable through bias adjustment) versus idiosyncratic (and therefore irreducible without fundamentally changing the assessment approach)? The finding that 73.6% of variance is residual establishes that most disagreement falls in the latter category—a result with direct practical implications for multi-model deployment.

The choice of a mixed effects specification—rather than a purely fixed effects model or a Bayesian hierarchical model—reflects a pragmatic balance between interpretability and statistical appropriateness. Treating nodes as random effects is justified because the 2,325 OPF nodes are a sample from a larger universe of possible business processes; treating models as fixed effects is appropriate because the four specific models were selected deliberately rather than sampled randomly from all possible LLMs. The residual term ε_{ij} absorbs both genuine model-by-node interaction (model j may have unusual expertise or blind spots for specific node i) and measurement noise (stochastic variation within the model’s response generation).

An important caveat is that the three SDAF components are not causally identified. D_{bias} captures systematic scoring differences but does not explain *why* models differ—whether due to training data composition, architectural idiosyncrasies, or language processing capabilities. Similarly, $D_{\text{ambiguity}}$ captures node-level variance but cannot distinguish inherent process ambiguity from ambiguity in the prompt’s description of that process. And D_{residual} , by construction, absorbs everything not captured by the two main effects—genuine model-by-node interactions, measurement noise, and any systematic sources of variation not modeled. These interpretive limitations are inherent to variance decomposition methods and are discussed further in Section 7.8.

The relationship between SDAF and the empirical observations presented in Chapters 5–6 should also be clarified. The Agreement Indeterminacy observation and Reliability Ceiling corollary are analytical results grounded in the data. SDAF provides the empirical framework for *contextualizing* these observations—for example, by using the D_{bias} estimate to compute the reliability ceiling bound for this specific set of four models. The observations provide the analytical insight; SDAF provides the organizational

structure for applying them.

The three SDAF components map onto practical decisions in multi-model deployment. If D_{bias} dominates, organizations can improve agreement by calibrating model outputs (e.g., subtracting each model’s mean bias). If $D_{\text{ambiguity}}$ dominates, the remedy is to improve node descriptions or assessment criteria. If D_{residual} dominates—as it does at 73.6%—neither model calibration nor instrument refinement will substantially improve agreement, and organizations must accept irreducible disagreement as a feature of the assessment domain. This practical interpretation motivates the term “organizational tool”: SDAF organizes disagreement into actionable categories rather than proposing a theory of why models disagree.

The SDAF variance decomposition results are presented in Section 5.5, with their implications discussed in Chapter 6.

4.5 3.5 Statistical Methods Overview

The analytical strategy employs seven complementary metric families, each addressing a different facet of inter-model agreement. This section provides a conceptual overview of each family and its role in the overall analysis; detailed mathematical formulations appear in the results sections of Chapter 5 where each method is first applied.

The multi-metric approach is motivated by the Kappa Paradox (Section 2.3): no single agreement metric can fully characterize the reliability structure, and seemingly contradictory conclusions across metrics are informative rather than erroneous. The seven families are organized in a logical progression: from basic agreement quantification (metrics 1–2), through structural decomposition (metrics 3–4), to ensemble assessment (metric 5), robustness verification (metric 6), and practical interpretation (metric 7).

Agreement metrics. Cohen’s κ quantifies pairwise chance-corrected agreement for all 6 model pairs across 5 categorical fields, producing a total of 30 pairwise κ values. The six model pairs are: Gemini-DeepSeek, Gemini-Qwen3, Gemini-GPT5mini, DeepSeek-Qwen3, DeepSeek-GPT5mini, and Qwen3-GPT5mini. Fleiss’ κ extends agreement measurement to the full four-rater case, providing a single omnibus statistic per field. Gwet’s AC1 serves as a prevalence-robust alternative that enables detection of the Kappa Paradox—cases where κ and AC1 yield contradictory conclusions due to distributional skewness. Quadratic-weighted Cohen’s κ gives partial credit for ordinal near-misses (e.g., “changed” vs. “will change” penalized less than “changed” vs. “stable”), testing whether models that disagree categorically nonetheless place assessments in adjacent categories. For the 18 numerical dimensions, ICC(2,1) measures absolute agreement treating both nodes and models as random effects, which is appropriate because the four models are a sample from the universe of possible frontier LLMs.

Information-theoretic measures. Normalized Mutual Information (NMI) captures shared information between model pairs without assuming a specific agreement model. Unlike κ , NMI does not require a definition of “chance agreement” and is therefore immune to the Kappa Paradox. Per-node entropy quantifies the degree of model disagreement at the individual process level, enabling identification of “easy” nodes (low entropy, high consensus) versus “hard” nodes (high entropy, maximal disagreement).

Structural analysis. Stratified agreement rates by hierarchy level (L1–L5), framework source (PCF, ITIL, SCOR, AI-era), and process domain (operating vs. management/support) test whether agreement depends systematically on node characteristics. These stratifications directly address RQ2 (structural determinants). Kruskal-Wallis H tests assess multi-group significance; Mann-Whitney U tests handle two-group comparisons; Spearman’s ρ measures rank-order agreement on ordinal fields, testing whether models agree on the *relative ordering* of processes even when they disagree on absolute ratings.

Bias characterization. Each model’s systematic assessment tendencies are characterized through four complementary measures: (a) mean ordinal scores across all fields, capturing central tendency (whether a model rates “high” or “low” on average), (b) distributional entropy of category assignments, measuring whether a model uses all categories evenly or concentrates on a few, (c) cross-dimensional consistency, quantifying the strength of association between a model’s assessments across different fields, and (d) domain sensitivity profiles, measuring how a model’s assessment patterns vary across framework sources. Together, these four measures form a “bias fingerprint” that uniquely characterizes each model’s assessment behavior.

Consensus mechanisms. Three aggregation approaches test whether multi-model consensus recovers useful signal despite low pairwise agreement: (a) majority vote (category chosen by at least 3 of 4 models), (b) unanimous agreement (all 4 models must concur), and (c) confidence-weighted vote (each model’s vote weighted by its self-reported `uncertainty_confidence`). The three mechanisms span a precision-recall tradeoff: unanimous agreement has high precision but low coverage (few nodes qualify), majority vote balances both, and confidence-weighted voting leverages model self-awareness (if it exists).

Robustness checks. Bootstrap resampling ($B = 2,000$) provides nonparametric confidence intervals for all point estimates. The bootstrap is preferred over asymptotic approximations because the sampling distribution of κ near zero is non-normal, and the four-rater design violates independence assumptions of standard variance formulas. Cluster bootstrap (111 L2-parent clusters) corrects for hierarchical dependency within the OPF tree structure—nodes sharing the same L2 parent are likely to receive correlated assessments. Monte Carlo simulation ($S = 1,000$) establishes random base-

lines by generating synthetic rating sets that preserve each model’s marginal distributions while destroying inter-model agreement, providing precise null distributions. This marginal-preserving null model is conservative in the sense that shared marginal biases (reflecting shared training data) are already built into the baseline; the 3.1 pp excess over this baseline therefore represents agreement *beyond* what would be expected from shared distributional tendencies alone. Leave-one-out analysis tests sensitivity to individual model removal, assessing whether findings depend critically on any single model. Subsample stability analysis ($N = 50$ to $2,325$) assesses whether agreement metrics converge as sample size increases, testing external validity.

Effect sizes and corrections. Cohen’s d , ANOVA η^2 , and Cramer’s V supplement significance tests with measures of practical effect magnitude, following Cohen’s (1988) conventions. With $N = 2,325$, even trivially small effects can achieve statistical significance; effect sizes distinguish substantively meaningful results from merely significant ones. Bonferroni correction ($\alpha_{\text{adj}} = 0.0045$ for 11 primary hypotheses) controls family-wise error rate for confirmatory tests, while Benjamini-Hochberg FDR ($q = 0.05$) provides better-powered control for the 18 correlated numerical dimension tests. Exploratory analyses are reported as descriptive findings without formal hypothesis testing, following the confirmatory-exploratory distinction advocated by Tukey (1977).

Eleven formal hypotheses (H1–H11) map each research question to testable statistical claims, covering agreement magnitude (H1–H2), structural determinants (H3–H5), model bias (H6), methodological robustness (H7–H9), and metric concordance (H10–H11). The complete hypothesis table and detailed mathematical formulations for all methods appear in the corresponding results sections (Chapter 5) and appendices.

Cross-validated prediction. To assess whether node-level features predict agreement outcomes, a logistic regression model predicts four-way agreement (binary: all four models agree vs. at least one dissenter) from structural features (hierarchy level, framework source, domain type). Five-fold cross-validation with AUC evaluation tests out-of-sample predictive performance, providing a benchmark for how much of the agreement structure is predictable from observable node characteristics alone. An AUC substantially above 0.5 would indicate that agreement is structurally determined; an AUC near 0.5 would suggest that agreement is essentially unpredictable from node metadata.

Rationale for multi-metric design. The choice to employ seven metric families rather than relying on a single “best” metric reflects a deliberate methodological position grounded in the literature review (Section 2.3). No single agreement metric is universally appropriate: κ penalizes prevalence imbalance; AC1 may be overly generous in highly skewed distributions; ICC assumes interval-level measurement; NMI is distribution-free but lacks established interpretive benchmarks. Each metric answers a subtly different question about agreement, and the choice of metric can change the

conclusion from “slight agreement” to “almost perfect agreement” (as demonstrated empirically for `boundary_current_type` in Section 5.10.9).

By computing all seven families and comparing their conclusions, the study can identify cases where metrics converge (increasing confidence in the finding) and cases where they diverge (revealing measurement-theoretic ambiguity that is itself informative). This “reliability corridor” approach—reporting the range of conclusions across metrics rather than selecting a single preferred metric—is the primary defense against the Kappa Paradox documented in Section 2.3. The approach follows He et al.’s (2025) recommendation against relying on single metrics in skewed distributions, and Geijer et al.’s (2025) advocacy for multi-metric reporting in agreement studies.

The transition from research design (this chapter) to data collection (Chapter 4) follows a logical separation: design decisions are made before any data is collected; data collection documents the execution of those decisions and any deviations or anomalies encountered during execution.

Chapter 5

Chapter 4: Data Collection

This chapter describes the operational protocol for collecting 213,900 individual judgments, the quality controls applied during and after collection, and the ethical considerations governing the study. Where Chapter 3 established *what* was measured and *how* it would be analyzed, this chapter documents the *execution* of the data collection campaign—the practical decisions, quality assurance mechanisms, and anomaly handling that produced the final analytic dataset.

5.1 4.1 Scanning Protocol and Quality Control

Each of the four models independently assessed all 2,325 OPF nodes using the identical prompt v2.2. Assessments were collected programmatically via each provider’s API, with no inter-model information sharing at any stage.

Independence design. The design implements four layers of independence to ensure uncontaminated assessment:

1. *Model-to-model independence:* No model had access to any other model’s assessments at any point during data collection.
2. *Prompt independence:* The prompt contained no reference to other models, multi-model comparison, or expected agreement levels that might anchor responses.
3. *Order independence:* Assessment order was randomized within each model’s scan to prevent systematic order effects.
4. *Researcher independence:* Assessment results were not inspected until all four models had completed their full scans, preventing mid-collection adjustments to the protocol.

This independence protocol eliminates the “agreeableness bias” identified by Deldjoo et al. (2025), where models converge toward consensus when exposed to each other’s outputs. The low observed agreement ($\kappa = 0.078$) is therefore not deflated by independence enforcement—it reflects genuine assessment divergence.

However, true statistical independence cannot be guaranteed. All four models were trained on web-crawled corpora that likely share substantial overlap—descriptions of business processes, AI impact analyses, and management consulting frameworks appear across the web and may be present in all four training datasets. This shared training data creates a latent dependence that our blinding protocol cannot address: models may agree (or disagree) not because they independently reached the same conclusion, but because they encountered similar descriptions in their training data. As noted in Section 3.5, the Monte Carlo baseline simulation addresses this partially by comparing against marginal-preserving random baselines, but cannot eliminate the shared-training confound entirely.

Operational parameters:

Model	API Endpoint	Batch Structure	Rate	Temperature
Gemini 2.5 Flash	Google AI Studio	Single batch	~5s/node	0.0
DeepSeek V3.2	DashScope API	7 sequential batches	~30s/node	0.0
Qwen3 235B	DashScope API	Single batch	~8s/node	0.0
GPT-5 mini	OpenAI API	Single batch	~6s/node	0.0

Data collection was conducted between late 2025 and early 2026. The four model scans were executed sequentially (not simultaneously) over a period of approximately two weeks, in the order: Gemini, DeepSeek, Qwen3, GPT-5 mini. The sequential execution was necessitated by practical constraints (API cost management, batch monitoring, error recovery) rather than methodological considerations. Since each model’s assessment is independent—no model sees any other model’s output—the temporal ordering of scans has no bearing on the statistical analysis. However, the sequential design means that if any model provider updated its model weights during the two-week collection window, earlier and later scans could reflect different model versions. Provider release notes were monitored during the collection period; no model updates were announced during the active scanning window for any of the four providers.

All models used temperature = 0.0, a deterministic setting that minimizes stochastic variation. While temperature = 0.0 should theoretically produce identical outputs across calls, in practice API-level non-determinism (floating-point arithmetic, batching, and infrastructure variation) can introduce minor output differences (see Appendix H for determinism testing). The notable variation in processing speed across models—DeepSeek required approximately 30 seconds per node versus 5–8 seconds for the other

three models—likely reflects differences in model architecture and inference infrastructure rather than assessment complexity. DeepSeek’s longer processing time may correlate with its substantially larger parameter count (671B total) and the overhead of its expert routing mechanism.

The total data collection campaign spanned approximately 80 hours of aggregate compute time across all four models. Each model completed its full scan within a continuous session or small number of sequential batches, minimizing the risk of model version changes mid-scan. DeepSeek’s 7-batch structure was necessitated by API rate limits and session timeout constraints; all batches used the same model checkpoint and prompt version.

Quality controls applied during collection. A four-layer quality control protocol was enforced at the point of data ingestion, before any analytical processing:

1. **Schema validation:** Each API response was parsed and validated against a JSON schema requiring all 23 fields with correct types (string for categorical, integer for numerical) and valid value sets. Responses failing schema validation were flagged for manual inspection and re-collection. In practice, all 9,300 responses passed schema validation on the first attempt, with the exception of the 8 Gemini boundary type values and 2 DeepSeek null values described in Section 4.3.
2. **Completeness check:** After each model’s scan completed (or after each batch for DeepSeek), a count verification confirmed that exactly 2,325 assessments existed with no duplicates and no missing nodes. Any shortfall would have triggered a targeted re-scan of missing nodes. No re-scans were required.
3. **Range enforcement:** All numerical scores were verified to fall within the valid range [1, 5]. Values outside this range would have triggered automatic re-assessment of the affected node. No out-of-range values were detected across all 167,400 numerical judgments.
4. **Provenance logging:** Every API call was recorded in a dedicated audit log table with batch ID, timestamp, model identifier, prompt version, raw JSON response, and response latency. This audit trail enables post-hoc verification that all assessments originated from the intended model version and prompt configuration.

5.2 4.2 Data Overview: 213,900 Judgments

The complete data matrix comprises 2,325 rows (one per process node) and 96 assessment columns (23 fields per model times 4 models, plus 4 metadata columns), yielding 213,900 individual judgments:

$$N_{\text{total}} = 23 \text{ fields} \times 4 \text{ models} \times 2,325 \text{ nodes} = 213,900$$

This decomposes into:

- **46,500 categorical judgments:** 5 categorical fields times 4 models times 2,325 nodes
- **167,400 numerical scores:** 18 numerical fields times 4 models times 2,325 nodes

All 9,300 raw API responses (4 models times 2,325 nodes) were archived in a versioned SQLite database with full provenance metadata. Each model's results are stored in dedicated tables (`ai_impact_scan_results_<model>_v22`) to prevent cross-contamination and enable independent verification.

Scale comparison. To contextualize the scale of this dataset, Table 2.1 showed that prior multi-model agreement studies typically involve 200–3,000 items assessed on 1–5 dimensions. This study's 2,325 items across 23 dimensions produces a data volume that is 100 to 350 times larger than the closest comparable work, enabling statistical analyses—such as mixed effects variance decomposition, cluster bootstrap with 111 clusters, and cross-validated prediction—that require substantially more data than prior studies could provide.

Data matrix structure. The final analytic dataset is organized as a rectangular matrix of 2,325 rows (nodes) by 96 columns: 4 metadata columns (node ID, name, level, framework source), plus 23 assessment columns for each of the 4 models (92 assessment columns total). This structure supports both node-level analysis (comparing four models' assessments of the same node) and model-level analysis (comparing one model's assessments across all nodes). The dual orientation is essential for the SDAF variance decomposition, which simultaneously estimates node-level ambiguity and model-level bias.

5.3 4.3 Missing Data and Anomaly Handling

A systematic post-collection audit was conducted to identify missing values, out-of-range scores, formatting anomalies, and logical inconsistencies across all 213,900 judgments. The audit identified two minor data quality issues affecting 10 of 213,900 judgments (0.005%):

Issue 1: Gemini boundary type formatting (8 values). Gemini 2.5 Flash returned verbose `boundary_current_type` values for 8 nodes, embedding explanatory text within the category label (e.g., “类型3(AI 主导, 人工监督): AI 系统主导执行...” instead of the canonical “类型 3”). These were extracted to canonical form using a regular expression `r"(类型 [1-4])"` that captures the category prefix and discards trailing description. All 8 values mapped unambiguously to one of the four valid categories.

Issue 2: DeepSeek priority flag nulls (2 values). DeepSeek V3.2 returned NULL for `summary_priority_flag` on 2 nodes. These were imputed as “常规验证” (routine verification), the modal category for this field across all models (accounting for

approximately 70% of all priority flag assignments). Imputation with the mode is conservative: it avoids inflating apparent agreement by assigning a distinctive value and does not bias toward any particular assessment pattern.

No other data transformations were applied. In particular:

- No outlier removal was performed on numerical scores (all values were within the valid $[1, 5]$ range)
- No category merging or recoding was applied to categorical fields
- No missing data imputation was needed beyond the two DeepSeek cases described above
- No model’s complete assessment was excluded or down-weighted

The 10 cleaned values represent 0.005% of all judgments, well below any conventional threshold for data quality concern. Sensitivity analyses (not reported) confirmed that excluding the 10 affected nodes entirely produces negligible changes to all reported statistics (maximum $\Delta\kappa < 0.001$).

It is noteworthy that the data quality issues are model-specific: Gemini produced verbose boundary types (a formatting issue reflecting the model’s tendency to explain its classifications), while DeepSeek produced null priority flags (a completeness issue suggesting the model occasionally skipped optional-seeming fields). These model-specific failure modes are themselves informative about assessment behavior differences, though the small number of affected values (8 and 2, respectively) prevents drawing robust conclusions from the cleaning patterns alone.

Post-collection integrity verification. After cleaning, five integrity checks were applied to the complete dataset:

1. **Completeness:** All four models produced assessments for all 2,325 nodes (zero missing assessments across 9,300 model-node combinations)
2. **Uniqueness:** No duplicate node assessments existed within any model’s result set
3. **Vocabulary compliance:** All categorical values fell within their defined vocabularies (after the 10 cleanings described above)
4. **Range compliance:** All 167,400 numerical scores fell within the valid range $[1, 5]$, with no fractional values
5. **Cross-model alignment:** Node ordering was verified to be consistent across all four models, ensuring that pairwise comparisons matched the correct nodes

All five checks passed without exception, confirming that the analytic dataset is complete, consistent, and ready for the statistical analyses described in Chapter 3.

Determinism testing. To assess the reproducibility of model outputs, a subset of 50 randomly selected nodes was re-assessed by each model after the primary scan completed (see Appendix H for full results). With temperature = 0.0, three of four models (Gemini, DeepSeek, Qwen3) produced identical outputs for all 50 nodes on

both runs. GPT-5 mini exhibited minor variation on 2 of 50 nodes (4%), where numerical sub-scores differed by 1 point on a single dimension. No categorical field changed between runs for any model. The minor GPT-5 mini variation likely reflects API-level non-determinism (floating-point arithmetic order, infrastructure variation) rather than temperature-driven sampling, as temperature = 0.0 should produce deterministic decoding. This level of near-determinism confirms that the single-pass design captures stable model behavior rather than ephemeral stochastic variation.

The high near-determinism rate has a methodological implication: the vast majority of inter-model disagreement observed in Chapter 5 cannot be attributed to within-model stochastic noise. If a model assigns “will change” to a particular node, it does so reliably across repeated runs—the disagreement between models primarily reflects genuine differences in assessment logic rather than random fluctuation. This distinguishes the present study from settings where intra-rater variability is a significant source of noise (e.g., human annotators with variable attention or fatigue). In the SDAF framework (Section 3.4), the D_{residual} component predominantly captures genuine model-by-node interaction, though a small fraction (estimated at $\leq 4\%$ based on the determinism testing) may reflect residual API-level non-determinism even at temperature = 0.0.

5.4 4.4 Ethical Considerations

No human subjects. This study involves no human participants. All 213,900 assessments were generated by AI models evaluating publicly documented business process classification data. The study falls outside the scope of human subjects review as defined by institutional research ethics guidelines. No personally identifiable information was collected, stored, or processed at any stage.

Data sources and intellectual property. Three of the four source frameworks (APQC PCF 7.4, ITIL 4, SCOR 12.0) are industry-standard reference taxonomies containing no personally identifiable information, proprietary business data, or sensitive content. Process node descriptions consist of generic business activity labels (e.g., “Manage Accounts Payable,” “Process Customer Returns”) that do not reference specific organizations or individuals. APQC PCF is publicly available through the APQC Process Classification Framework; ITIL 4 and SCOR 12.0 process lists are drawn from published framework documentation. The use of these frameworks for academic research purposes falls within fair use provisions, and the OPF integration adds original structural contributions (hierarchy unification, bilingual labeling, AI-era extensions) that constitute a transformative work.

Author-created content. The 99 AI-era extension nodes were created by the dissertation author specifically for this framework. This dual role—as both framework

designer and researcher—constitutes a potential conflict of interest that must be disclosed transparently. Three specific concerns arise:

1. **Design bias:** The author may have unconsciously designed AI-era nodes in ways that favor particular assessment outcomes—for example, using process descriptions that steer models toward “high AI impact” assessments.
2. **Vocabulary effects:** The AI-era nodes use terminology (e.g., “algorithmic bias audit,” “synthetic data governance”) that may appear more frequently in recent training data, potentially giving some models an advantage in familiarity-dependent scoring.
3. **Non-representative taxonomy:** The 99 nodes may reflect the author’s idiosyncratic conceptualization of AI governance processes rather than a broadly accepted professional consensus, since no standard taxonomy exists for this domain.

These concerns are mitigated by four design choices:

- (a) Stratified analysis reports AI-era results separately from established-framework results throughout Chapter 5, enabling readers to assess whether AI-era nodes exhibit anomalous agreement patterns
- (b) The small proportion of author-created content (4.3% of total) means that even if these nodes were entirely excluded, all primary findings would remain statistically valid
- (c) The primary research questions concern inter-model agreement *patterns* rather than the “correctness” of any individual assessment, reducing the impact of potential design bias in node descriptions
- (d) Explicit flagging is applied wherever AI-era results diverge from the overall pattern, alerting readers to potential confounds

API usage compliance. All model API calls complied with each provider’s terms of service. No rate limits were violated, no API access restrictions were circumvented, and no model outputs were used for purposes prohibited by the respective terms. The research use of model outputs for academic analysis falls within the permitted use cases specified by all four providers’ terms of service at the time of data collection.

Environmental impact. The total computational cost of the data collection campaign—approximately 80 hours of aggregate API compute across four models—is non-trivial. While a precise carbon footprint estimate is beyond the scope of this study (as API providers do not disclose per-query energy consumption), transparency about the computational scale supports informed evaluation of the research’s environmental tradeoffs. The study’s design choices—using four models rather than more, and collecting single-pass assessments rather than multi-round evaluations—were partly motivated by minimizing unnecessary computational expenditure while maintaining sufficient statistical power for the intended analyses.

Summary of data collection decisions. Table 4.2 summarizes the key design de-

cisions made during data collection and their implications for the analysis.

Table 4.2: Data Collection Design Decisions

Decision	Choice	Rationale	Limitation
Temperature	0.0	Maximize determinism	Does not guarantee identical outputs
Batch structure	Model-dependent	API constraints	DeepSeek’s 7-batch structure may introduce batch effects
Cleaning strategy	Minimal (10 values)	Preserve model behavior	Does not correct potential systematic output biases
Prompt version	v2.2 only	Control confounds	Cannot disentangle model from prompt effects
Re-assessment	None (no failures)	All responses valid	No opportunity to study response stability at scale
Temporal ordering	Sequential	Practical constraints	Models assessed at different calendar dates
Independence	Quadruple blind	Eliminate agreeableness bias	Cannot control for shared training data
Determinism test	50-node retest	Verify output stability	Limited sample of retest nodes

Reproducibility. To support reproducibility, all raw API responses, cleaning scripts, and analysis code are archived in the project repository with the following artifacts:

- **Raw data:** 9,300 JSON responses in versioned SQLite tables, one per model
- **Prompt:** Complete v2.2 prompt text with all calibration rules and anchor descriptions
- **Configuration:** Model versions, API parameters, temperature settings, and batch structures
- **Analysis code:** All statistical analysis scripts with deterministic random seeds
- **Cleaning scripts:** Regex patterns and imputation logic for the 10 cleaned values

However, exact replication of the data collection is inherently limited by the non-deterministic nature of LLM inference: even with temperature = 0.0, model providers may update underlying model weights, modify tokenization, or alter inference infras-

structure between the time of data collection and any replication attempt. The archived raw responses therefore serve as the authoritative dataset; replication efforts should focus on reproducing the statistical analyses from these fixed data rather than regenerating the assessments themselves. All statistical results reported in Chapter 5 are accompanied by 148 automated assertions that verify key findings programmatically, serving as a regression test for analytical reproducibility.

Having described the research design (Chapter 3) and data collection protocol (Chapter 4), we now turn to the empirical results. Chapter 5 presents findings across 48 tables and 29 figures, proceeding from descriptive distributions through agreement metrics, structural determinants, bias fingerprinting, consensus analysis, and robustness checks. All findings are verified by 148 automated statistical assertions.

Key Takeaways — Chapters 3 and 4 - The empirical domain (OPF) spans 2,325 process nodes across 4 frameworks, producing 213,900 independent judgments from 4 frontier LLMs - All four models share MoE architecture, limiting diversity claims to within the MoE family - A single prompt version (v2.2) controls for prompt variation but prevents disentangling model effects from model-by-prompt interactions - SDAF is an organizational variance decomposition tool, not a theoretical framework - Data quality is high: 10 of 213,900 values required cleaning (0.005%), with zero missing assessments - The 99 author-created AI-era nodes (4.3%) are disclosed as a potential conflict of interest and analyzed separately

Chapter 6

Chapter 5: Results

This chapter presents the empirical findings from the multi-model assessment study. Four frontier large language models—Gemini 2.5 Flash, DeepSeek V3.2, Qwen3 235B, and GPT-5 mini—independently assessed 2,325 business process nodes across 23 structured fields (5 categorical, 18 numerical), producing a total of 213,900 individual judgments. Results are organized in ascending analytical complexity: descriptive statistics (Section 5.1), pairwise agreement (Section 5.2), multi-rater agreement (Section 5.3), chance-corrected agreement under prevalence adjustment (Section 5.4), variance decomposition (Section 5.5), consensus analysis (Section 5.6), cluster-corrected inference (Section 5.7), cross-validated prediction (Section 5.8), continuous reliability (Section 5.9), and a unified multi-metric summary (Section 5.10).

6.1 5.1 Descriptive Statistics

The study comprises $2,325 \text{ process nodes} \times 4 \text{ models} \times 23 \text{ fields} = 213,900$ judgments. Before examining agreement, we characterize the marginal distributions that each model produces for the five categorical fields, as distributional differences set the stage for interpreting all subsequent reliability metrics.

Table 5.1: Categorical Distribution Summary — change_status

Model	Will			Dominant Category
	Changed (已变)	Change (将变)	Stable (稳定)	
Gemini 2.5 Flash	Substantial	Majority	Small	Relatively balanced
DeepSeek V3.2	Small	Majority	Large	Strong “稳定” bias

Model	Will			Dominant Category
	Changed (已变)	Change (将变)	Stable (稳定)	
Qwen3 235B	Minimal	Majority	Large	Strong “稳定” bias
GPT-5 mini	Small	Very large	Minimal	Strong “将变” bias

The most striking finding at the distributional level is the divergent anchoring behavior of the four models. Gemini produces a relatively balanced distribution across the three change_status categories, allocating substantial mass to both “已变” and “稳定.” DeepSeek and Qwen3 share a strong conservative bias toward “稳定,” suggesting that both models default toward process stability when faced with ambiguous assessment targets. GPT-5 mini exhibits the opposite extreme: a pronounced bias toward “将变” that effectively collapses the three-category schema into a near-degenerate single-category distribution. These distributional asymmetries have direct consequences for chance-corrected agreement metrics, as explored in Sections 5.2–5.4.

Table 5.2: Study Design Parameters

Parameter	Value
Process nodes assessed	2,325
Models (raters)	4
Categorical fields	5
Numerical fields	18
Total fields per node	23
Total judgments	213,900
Prompt version	v2.2 (identical across all models)
Source frameworks	APQC PCF, ITIL 4, SCOR 12.0, AI-era extensions

All four models received identical prompts (v2.2) and identical input data, ensuring that observed differences reflect model-intrinsic properties rather than prompt confounds.

Table 5.2b: Model Bias Characterization by change_status

Model	Bias Direction	Characteristic Pattern
Gemini 2.5 Flash	Balanced	Distributes mass across all three categories
DeepSeek V3.2	Conservative	Strong default toward “稳定” (stable)
Qwen3 235B	Conservative	Strong default toward “稳定” (stable)
GPT-5 mini	Progressive	Strong default toward “将变” (will change)

The bias alignment between DeepSeek and Qwen3 (both conservative) and their divergence from GPT-5 mini (progressive) foreshadow the pairwise agreement patterns reported in Section 5.2: model pairs that share distributional biases tend to achieve higher chance-corrected agreement, not because they assess individual nodes identically but because they populate the same categories at similar rates.

The distributional divergence is not merely a matter of calibration. If the four models simply applied different thresholds to the same underlying latent variable, we would expect their rank orderings of nodes to be consistent even as their category proportions differ. As the ICC analysis in Section 5.9 demonstrates, rank-order consistency is itself limited (mean ICC = 0.174), indicating that the models are not merely re-thresholding a shared signal but are, to a substantial degree, evaluating different aspects of the assessment target.

6.2 5.2 Pairwise Agreement (Cohen’s κ)

Cohen’s κ measures chance-corrected agreement between two raters. With 4 models and 5 categorical fields, we compute 6 pairwise κ values per field, yielding 30 total comparisons.

Table 5.3: Cohen’s κ — Mean by Categorical Field

Field	Mean κ (across 6 pairs)	Landis–Koch Interpretation
change_status	0.108	Slight
penetration_overall	0.086	Slight

Field	Mean κ (across 6 pairs)	Landis–Koch Interpretation
uncertainty_confidence	0.074	Slight
summary_priority_flag	0.067	Slight
boundary_current_type	0.052	Slight
Grand mean	0.078	Slight

The grand mean $\kappa = 0.078$ across all 30 comparisons falls squarely in the “slight” agreement range on the Landis and Koch (1977) scale. No field achieves even “fair” agreement ($\kappa \geq 0.21$) at the mean level. The ordering of fields—change_status highest, boundary_current_type lowest—suggests that temporal assessment (will the process change?) elicits more consistent judgments than typological classification (what kind of boundary does the process have?).

Table 5.4: Notable Pairwise κ Values

Model Pair	Field	κ	Interpretation
DeepSeek–Qwen3	change_status	0.448	Moderate (highest in dataset)
Gemini–Qwen3	change_status	−0.020	Below chance (system- atic disagree- ment)
DeepSeek–GPT-5 mini	summary_priority_flag	−0.007	Below chance

The range within a single field is remarkable. For change_status, κ spans from −0.020 (Gemini–Qwen3) to 0.448 (DeepSeek–Qwen3)—a 0.47-unit range on a scale where 1.0 represents perfect agreement. The DeepSeek–Qwen3 pair achieves the only “moderate” agreement in the entire 30-cell matrix, likely reflecting shared conservative biases: both models assign substantial probability mass to “稳定,” creating high agreement on the dominant category. Meanwhile, two of the 30 cells show negative κ values, indicating that these model pairs agree *less* than would be expected by chance alone—a finding that underscores the severity of inter-model disagreement.

Table 5.4b: Distribution of Pairwise κ Values Across 30 Comparisons

κ Range	Count	Percentage	Interpretation
$\kappa < 0$ (below chance)	2	6.7%	Anti-agreement
$0.00 \leq \kappa < 0.10$	18	60.0%	Slight
$0.10 \leq \kappa < 0.20$	8	26.7%	Slight to fair
$0.20 \leq \kappa < 0.40$	1	3.3%	Fair
$\kappa \geq 0.40$	1	3.3%	Moderate

The distribution of the 30 pairwise κ values is heavily right-skewed: 60% fall below 0.10, and only a single comparison (3.3%) reaches the “moderate” range. The two negative values and the 18 near-zero values together account for two-thirds of all comparisons, confirming that slight-to-negligible agreement is the norm rather than the exception.

Interpretation. A grand mean κ of 0.078 means that, after correcting for chance agreement, the four models share less than 8% of the theoretically achievable agreement. In human annotation studies, κ values below 0.20 are typically considered insufficient for any substantive inference (Artstein and Poesio, 2008). By this standard, multi-model LLM assessment on structured domains fails to reach the minimum reliability threshold established for human raters. This finding is robust: 28 of 30 comparisons (93.3%) fall below the 0.20 “fair” threshold individually, not merely on average.

6.3 5.3 Fleiss’ κ (Four-Rater Agreement)

While Cohen’s κ applies to pairs, Fleiss’ κ (1971) extends chance-corrected agreement to an arbitrary number of raters. Table 5.5 presents Fleiss’ κ for all four models simultaneously.

Table 5.5: Fleiss’ κ for Four-Rater Agreement

Field	Fleiss’ κ	95% CI	Interpretation
change_status	0.089	—	Slight
uncertainty_confidence	0.051	—	Slight
boundary_current_type	0.041	—	Slight
penetration_overall	−0.006	—	Below chance
summary_priority_flag	−0.016	—	Below chance
Mean	0.032	—	Slight

Two of the five fields—penetration_overall ($\kappa_F = -0.006$) and summary_priority_flag ($\kappa_F = -0.016$)—yield *negative* Fleiss’ κ values. This means the four models agree less than four hypothetical raters assigning categories at random

with the same marginal distributions. The result is not merely “low agreement” but *anti-agreement*: the models’ responses are structured in a way that produces systematic divergence beyond what chance would predict.

The mean Fleiss’ κ of 0.032 is substantially lower than the mean pairwise Cohen’s κ of 0.078. This gap is expected: Fleiss’ κ is a stricter measure because it requires consistency across all four raters simultaneously rather than within individual pairs. The gap quantifies the additional reliability cost of requiring consensus from a larger panel.

Table 5.6: Pairwise vs. Multi-Rater Agreement Comparison

Metric	Value	Interpretation
Mean Cohen’s κ (pairwise)	0.078	Slight
Mean Fleiss’ κ (4-rater)	0.032	Slight (near zero)
Ratio (Fleiss/Cohen)	0.41	59% agreement loss from pairwise to multi-rater

The 59% reduction from pairwise to multi-rater agreement highlights a critical scaling problem: as more models are included in a consensus panel, the chance-corrected agreement deteriorates disproportionately. This has practical implications for multi-model ensemble strategies that assume agreement improves with additional models.

6.4 5.4 Gwet’s AC1 and the Kappa Paradox

Gwet’s AC1 (2008) addresses a well-known limitation of κ -family statistics: their instability under extreme marginal distributions. When one category dominates, the expected chance agreement approaches the observed agreement, driving κ toward zero regardless of actual agreement quality.

Table 5.7: Gwet’s AC1 by Categorical Field

Field	AC1	Landis–Koch (AC1)	Fleiss’ κ	Landis–Koch (κ)
boundary_current_type	0.889	Almost perfect	0.041	Slight
penetration_overall	0.616	Substantial	−0.006	Below chance
uncertainty_confidence	0.607	Substantial	0.051	Slight
change_status	0.602	Substantial	0.089	Slight

Field	AC1	Landis–Koch (AC1)	Fleiss’ κ	Landis–Koch (κ)
summary_priority_flag	0.220	Fair	−0.016	Below chance
Mean	0.587	Moderate	0.032	Slight

Table 5.8: The Kappa Paradox — boundary_current_type

Metric	Value	Qualitative Label
Fleiss’ κ	0.041	Slight
Gwet’s AC1	0.889	Almost perfect
Discrepancy	0.848	21× difference

The most striking result in Table 5.7 is the `boundary_current_type` field, which achieves $AC1 = 0.889$ (“almost perfect” agreement) while simultaneously yielding $Fleiss' \kappa = 0.041$ (“slight” agreement). This 21-fold discrepancy constitutes a textbook instance of the *Kappa Paradox* (Feinstein and Cicchetti, 1990). The paradox arises because the vast majority of process nodes are classified as a single boundary type (類型 2) by all four models. The models genuinely agree—79.2% of nodes receive identical classifications from all four raters (see Section 5.6)—but because the marginal distributions are nearly degenerate, the chance-correction denominator in κ inflates, driving the statistic toward zero.

The practical consequence is that the answer to “do LLMs agree on boundary type?” depends entirely on whether one uses κ or AC1. Under κ , `boundary_current_type` appears to be one of the *worst* fields for agreement; under AC1, it is unambiguously the *best*. This metric dependence is not a statistical curiosity but a fundamental interpretive challenge that any multi-model reliability study must confront.

The mean AC1 across all five fields is 0.587 (“moderate”), compared to a mean $Fleiss' \kappa$ of 0.032 (“slight”). The gap is most pronounced for fields with extreme marginal distributions (`boundary_current_type`, `penetration_overall`) and smallest for the most balanced field (`summary_priority_flag`: $AC1 = 0.220$, $\kappa = -0.016$). This pattern confirms that the AC1– κ divergence is driven primarily by marginal homogeneity rather than substantive agreement differences.

6.5 5.5 SDAF Variance Decomposition (Mixed Effects)

To move beyond aggregate agreement statistics, we decompose the variance in model assessments using a mixed effects model. The Structured Disagreement Analy-

sis Framework (SDAF) partitions total variance into three components:

$$Y_{ij} = \mu + \alpha_i(\text{node}) + \beta_j(\text{model}) + \varepsilon_{ij}$$

where α_i captures node-level ambiguity (some processes are inherently harder to assess), β_j captures systematic model bias (each model has characteristic tendencies), and ε_{ij} captures residual interaction effects. The variance proportions $D_{\text{ambiguity}}$, D_{bias} , and D_{residual} sum to 100%.

Table 5.9: SDAF Variance Decomposition by Categorical Field

Field	$D_{\text{bias}}\%$	$D_{\text{ambiguity}}\%$	$D_{\text{residual}}\%$	F_{model}	p
summary_priority_flag	42.2	14.1	43.6	2252.1	<0.001
penetration_overall	16.2	5.5	78.3	482.5	<0.001
change_status	11.9	11.0	77.1	358.4	<0.001
uncertainty_confidence	11.4	7.9	80.7	329.0	<0.001
boundary_current_type	4.3	7.2	88.5	114.5	<0.001
Overall	17.2	9.1	73.6	—	—

All F_{model} values are significant at $p < 0.001$, confirming that model bias is statistically significant for every categorical field. However, the magnitude of bias varies dramatically across fields. The bootstrap 95% confidence interval for overall D_{bias} spans [10.8, 43.8] percentage points across fields, reflecting this heterogeneity.

Table 5.10: SDAF Variance Component Interpretation

Component	Overall %	Interpretation
D_{bias} (model effect)	17.2	Systematic model tendencies account for ~1/6 of total variance
$D_{\text{ambiguity}}$ (node effect)	9.1	Node-inherent difficulty accounts for ~1/11 of total variance

Component	Overall %	Interpretation
D_{residual} (interaction)	73.6	Idiosyncratic model×node interactions dominate

Table 5.10b: SDAF Field Ranking by Dominant Variance Component

Rank by D_{bias}	Field	D_{bias} %	Rank by D_{residual}	Field	D_{residual} %
1	summary_priority_flag	42.2	1	boundary_current_type	88.5
2	penetration_overall	16.2	2	uncertainty_confidence	80.7
3	change_status	11.9	3	penetration_overall	78.3
4	uncertainty_confidence	11.4	4	change_status	77.1
5	boundary_current_type	4.3	5	summary_priority_flag	43.6

Three findings warrant emphasis. First, the residual component dominates at 73.6%, meaning that most disagreement is *idiosyncratic*—it cannot be predicted from knowing which model or which node is involved. Each model responds to each node in a partly unpredictable way that defies simple characterization. Second, model bias (17.2%) substantially exceeds node ambiguity (9.1%), indicating that “who is rating” matters more than “what is being rated.” Third, the `summary_priority_flag` field is an outlier: model bias accounts for 42.2% of its variance, the highest of any field, suggesting that models have particularly divergent internal criteria for flagging process priority.

The inverse relationship between D_{bias} and D_{residual} rankings (Table 5.10b) is notable: the field with the highest bias (`summary_priority_flag`, 42.2%) has the lowest residual (43.6%), and vice versa. This suggests a conservation-of-variance principle: when model bias is high, less variance remains to be attributed to idiosyncratic interactions. From a practical standpoint, high-bias fields are paradoxically *more predictable* than high-residual fields—one can partially correct for known model biases, whereas idiosyncratic interactions resist systematic correction.

6.5.1 5.5.1 SDAF Interaction Terms: Framework Dependence

A critical extension examines whether model bias operates uniformly across source frameworks or interacts with framework identity. Table 5.11 reports the range of D_{bias} across the four source frameworks (APQC PCF, ITIL 4, SCOR 12.0, AI-era) for each field.

Table 5.11: D_{bias} Range by Source Framework

Field	D_{bias} Range (pp)	PCF D_{bias} %	AI-era D_{bias} %	Interpretation
summary_priority_flag	23.6	44.3	20.7	Strong interaction
penetration_overall	17.6	24.1	6.5	Strong interaction
uncertainty_confidence	9.0	—	—	Moderate interaction
boundary_current_type	5.4	—	—	Relatively stable
change_status	4.3	—	—	Relatively stable

Model bias is *not* uniform across frameworks. For `penetration_overall`, D_{bias} ranges from 24.1% (PCF processes) to 6.5% (AI-era processes)—a $3.7\times$ difference. For `summary_priority_flag`, the range is even wider: 44.3% (PCF) versus 20.7% (AI-era), a $2.1\times$ difference. The pattern is consistent: model bias is substantially higher for traditional business process frameworks (especially APQC PCF) than for AI-era extensions.

Table 5.11b: Framework Interaction Effect Sizes

Field	Max D_{bias} % (Framework)	Min D_{bias} % (Framework)	Ratio (Max/Min)
penetration_overall	24.1 (PCF)	6.5 (AI-era)	$3.7\times$
summary_priority_flag	44.3 (PCF)	20.7 (AI-era)	$2.1\times$
uncertainty_confidence	—	—	—
boundary_current_type	—	—	—
change_status	—	—	—

This finding has a natural interpretation. AI-era processes (e.g., “Evaluate Training Data Bias,” “Deploy Autonomous Decision Agents”) have self-evident relationships to AI that constrain model interpretation, reducing the scope for model-specific biases to operate. Traditional processes (e.g., “Manage Fixed Asset Accounting,” “Process Payroll”) require models to make inferential leaps about AI relevance, creating more room for model-specific heuristics and training data biases to drive assessments. The $\text{model} \times \text{framework}$ interaction thus reveals that model reliability is *domain-contingent*—a finding with practical implications for deployment contexts where framework choice is discretionary.

The interaction effect is not merely statistically significant but practically large: a $3.7\times$ ratio in D_{bias} between PCF and AI-era nodes for `penetration_overall` means that

model selection has nearly four times the impact on assessment outcomes when evaluating traditional business processes compared to AI-native processes. This asymmetry suggests that multi-model consensus strategies should weight agreement evidence differently depending on the domain being assessed.

6.6 5.6 Consensus Analysis

Despite low pairwise and multi-rater agreement, majority voting (≥ 3 of 4 models agree) may still recover useful signal. This section examines the consensus structure across fields.

Table 5.12: Consensus Distribution by Field

Field	Full Consensus (4/4)	Strong Majority (3/1)	Weak Majority (2/2+)	Full Divergence	Total Resolved (4/4 + 3/1)
boundary_current_type	79.2%	18.9%	1.5%	0.3%	98.1%
change_status	42.9%	35.4%	18.6%	3.0%	78.3%
uncertainty_confidence	41.8%	41.7%	12.0%	4.5%	83.5%
penetration_overall	34.2%	54.8%	8.7%	2.2%	89.0%
summary_priority_flag	9.7%	42.5%	28.3%	19.6%	52.2%

Table 5.13: Aggregate Consensus Statistics

Statistic	Value
Full consensus (all 5 fields, all 4 models agree)	42 nodes (1.8%)
Majority consensus ($\geq 3/4$ on each field)	80.3% of field-judgments
Hard nodes (≥ 3 fields with weak/no majority)	215 nodes (9.2%)

The consensus analysis reveals a pronounced field hierarchy. `boundary_current_type` is almost universally resolved: 79.2% of nodes receive identical classifications from all four models, and 98.1% are resolved by majority vote. This high consensus is consistent with the high AC1 (0.889) reported in Section 5.4 and reflects the dominance of a single boundary type in the process taxonomy.

At the opposite extreme, `summary_priority_flag` is the most contested field. Only 9.7% of nodes achieve full consensus, and nearly half (47.9%) remain unresolved by majority vote—the highest unresolved rate of any field. The 19.6% full divergence rate means that for roughly one in five nodes, no two models agree on whether the process should be flagged as a priority. Combined with the high D_{bias} (42.2%) reported

in Section 5.5, this identifies `summary_priority_flag` as the field most susceptible to model-specific interpretation.

The 42 nodes (1.8%) achieving full consensus across all five fields represent the most unambiguous assessment targets in the dataset. These nodes cluster in semantically coherent families—typically processes with self-evident AI relationships (AI-era extensions) or processes so far removed from AI that all models agree on stability. The 215 hard nodes (≥ 3 fields with weak or no majority) represent the assessment frontier where model disagreement is most pronounced and where human expert adjudication would be most valuable.

Table 5.14: Consensus Resolution by Field — Detailed Breakdown

Field	Resolvable by Majority	Unresolvable	Resolution Rate	Difficulty Rank
<code>boundary_current_type</code>	2,281	44	98.1%	1 (easiest)
<code>penetration_overall</code>	2,070	255	89.0%	2
<code>uncertainty_confidence</code>	1,942	383	83.5%	3
<code>change_status</code>	1,821	504	78.3%	4
<code>summary_priority_flag</code>	1,214	1,111	52.2%	5 (hardest)

The field difficulty ranking in Table 5.14 is remarkably consistent with the AC1 ranking from Section 5.4 (Spearman $\rho = 1.0$ for the top 4 fields), confirming that consensus resolution and prevalence-adjusted agreement capture the same underlying dimension of assessment difficulty. The `summary_priority_flag` field stands apart: it is the only field where the majority of unresolvable nodes exceeds 40%, marking it as a categorically different assessment challenge.

Interpretation. The 80.3% majority consensus rate demonstrates that despite near-zero κ values, a substantial majority of process nodes can be assigned a consensus classification through majority voting. This gap between low κ and high majority consensus is itself informative: it reflects the fact that most disagreement is concentrated in a minority of nodes and a minority of fields, while the majority of assessments converge on the same category. Multi-model ensembles can thus extract useful signal from individually unreliable raters, provided the user accepts that approximately 20% of field-judgments lack reliable consensus.

The disconnect between low κ and high consensus is not paradoxical—it is a mathematical consequence of the difference between chance-corrected and raw agreement. Majority consensus is a *raw* measure that benefits from distributional concentration: when all models tend to assign the same dominant category, majority voting succeeds even if the models’ assessments are statistically independent. The κ statistic corrects for this baseline inflation, which is why it gives a more conservative (and arguably more honest) picture of genuine agreement.

6.7 5.7 Cluster Bootstrap Confidence Intervals

The 2,325 process nodes are not independent: they are nested within 111 L2 parent clusters (e.g., all children of “1.1 Define Business Concept and Strategy” share structural and semantic similarities). Standard bootstrap methods that resample individual nodes may underestimate standard errors by ignoring this clustering. We apply a cluster bootstrap procedure that resamples entire L2 parent groups to produce correctly calibrated confidence intervals.

Table 5.15: Cluster Bootstrap vs. Naive Bootstrap for Mean Cohen’s κ

Method	Point Estimate	95% CI Lower	95% CI Upper	CI Width
Naive bootstrap	0.078	0.047	0.110	0.063
Cluster bootstrap (111 L2 clusters)	0.0772	0.0639	0.0909	0.0270

The cluster bootstrap yields a narrower confidence interval ([0.0639, 0.0909]) compared to the naive bootstrap ([0.047, 0.110]). This seemingly counterintuitive result—one might expect cluster correction to *widen* intervals—arises because the naive bootstrap, by resampling individual nodes, creates artificial samples that break the within-cluster correlation structure, introducing excess variability. The cluster bootstrap preserves within-cluster dependencies, producing a more stable estimate.

Table 5.16: Cluster Bootstrap Inference Summary

Inference	Result
Mean κ significantly different from zero?	Yes (lower CI bound = 0.0639 > 0)
Mean κ significantly different from 0.20 (“fair”)?	Yes (upper CI bound = 0.0909 < 0.20)
Conclusion	Agreement is real but firmly in the “slight” range

The cluster bootstrap confirms two key inferences. First, the mean κ is statistically significantly greater than zero—the models are not completely independent, and their agreement, while slight, is genuine. Second, the mean κ is statistically significantly below the “fair” threshold of 0.20, meaning that even after accounting for clustering, the data provide no evidence for substantive inter-model agreement.

Table 5.16b: Sensitivity of Inference to Bootstrap Method

Inference	Naive Bootstrap	Cluster Bootstrap	Conclusion Robust?
$\kappa > 0$	Yes (0.047 > 0)	Yes (0.064 > 0)	Yes
$\kappa < 0.20$	Yes (0.110 < 0.20)	Yes (0.091 < 0.20)	Yes
$\kappa < 0.10$	No (0.110 > 0.10)	Yes (0.091 < 0.10)	Depends on method

Table 5.16b reveals that the two bootstrap methods agree on the two primary inferences ($\kappa > 0$ and $\kappa < 0.20$) but diverge on a secondary inference: whether κ is below 0.10. The naive bootstrap upper bound (0.110) exceeds 0.10, while the cluster bootstrap upper bound (0.091) does not. This discrepancy is practically relevant because 0.10 is sometimes used as a threshold for “trivially low” agreement. The cluster-corrected result provides the stronger inference: even under proper accounting for hierarchical dependence, inter-model agreement does not reach 0.10.

6.8 5.8 Cross-Validated AUC

To assess whether agreement patterns can predict consensus outcomes, we train a logistic regression model predicting whether the majority consensus for a node is “稳定” (stable) from agreement features across fields. To avoid circularity, we exclude response_diversity features (which mechanically correlate with consensus) and use only mean pairwise agreement rates as predictors.

Table 5.17: Cross-Validated AUC for Consensus Prediction

Metric	Value
Predictor	Mean pairwise agreement across fields
Outcome	Consensus = “稳定”
Cross-validation	5-fold
Mean AUC	0.877 ± 0.019
Original AUC (with response_diversity)	0.923
AUC reduction from excluding circular features	0.046 (5.0%)

Table 5.18: AUC Interpretation

AUC Range	Interpretation	Study Result
0.50–0.60	No discrimination	—
0.60–0.70	Poor	—
0.70–0.80	Acceptable	—
0.80–0.90	Excellent	0.877
0.90–1.00	Outstanding	(0.923 with circular features)

The cross-validated AUC of 0.877 falls in the “excellent” range, demonstrating that agreement patterns contain substantial predictive information about consensus outcomes even after removing potentially circular features. The modest reduction from 0.923 to 0.877 (5.0%) confirms that the original model’s predictive power was primarily driven by genuine agreement patterns rather than tautological response_diversity features.

Table 5.18b: AUC Robustness Check — Feature Ablation

Predictor Set	Mean AUC	Interpretation
All features including response_diversity	0.923	Outstanding (but potentially circular)
Agreement features only (response_diversity excluded)	0.877	Excellent (circularity- free)
AUC reduction	0.046	5.0% loss — most signal is genuine

This result has two practical implications. First, agreement features can serve as reliable indicators of consensus quality. When mean pairwise agreement for a given node is high, the resulting majority-vote classification is likely to be robust; when agreement is low, the consensus label should be treated with caution. This enables a *confidence-stratified* approach to consensus interpretation. Second, the modest AUC reduction (5.0%) upon removing potentially circular features provides methodological assurance: the predictive relationship between agreement patterns and consensus outcomes is not an artifact of feature construction but reflects genuine statistical structure in the data.

6.9 5.9 ICC for Numerical Dimensions

The 18 numerical assessment fields (scored on continuous or ordinal scales) permit computation of Intraclass Correlation Coefficients. We report ICC(2,1)—a two-way random effects model treating both nodes and models as random—which quantifies the reliability of a single model’s rating.

Table 5.19: ICC(2,1) for Numerical Dimensions — Top 5

Rank	Dimension	ICC(2,1)	Interpretation
1	d1_decision_replaceability	0.325	Fair
2	d7_rule_driven_degree	0.316	Fair
3	d8_data_intensity	0.306	Fair
4	d8_data_standardization	0.294	Fair
5	d9_data_availability	0.272	Fair

Table 5.20: ICC(2,1) for Numerical Dimensions — Bottom 3

Rank	Dimension	ICC(2,1)	Interpretation
16	d8_cross_process_dependency	0.053	Poor
17	d8_integration_barrier	0.050	Poor
18	d7_exception_flexibility	0.040	Poor

Table 5.21: ICC Summary Statistics

Statistic	Value
Mean ICC(2,1) across 18 dimensions	0.174
Median ICC(2,1)	—
Range	[0.040, 0.325]
Dimensions with ICC \geq 0.30 (“fair”)	3
Dimensions with ICC $<$ 0.10 (“poor”)	3
All 18 p -values	<0.001

All 18 ICC values are statistically significant ($p < 0.001$), but the mean ICC of 0.174 falls in the “poor-to-fair” range (Cicchetti, 1994). The top-performing dimensions share a common characteristic: they assess concrete, observable properties of processes. Decision replaceability (ICC = 0.325) asks whether AI can replace human decisions in a process—a relatively concrete question. Rule-driven degree (ICC = 0.316) asks whether a process follows explicit rules—again, a property with observable indicators.

The bottom-performing dimensions assess more abstract or contextual properties. Exception flexibility (ICC = 0.040) requires models to judge how a process handles unexpected situations—a judgment that depends heavily on assumed context. Integration barrier (ICC = 0.050) and cross-process dependency (ICC = 0.053) require understanding of inter-process relationships that may not be fully specified in the process description alone.

Table 5.21b: ICC Reliability Benchmarks (Cicchetti, 1994)

ICC Range	Label	Dimensions in This Range	Example
< 0.40	Poor	15 (83.3%)	d7_exception_flexibility (0.040)
0.40–0.59	Fair	0 (0%)	—
0.60–0.74	Good	0 (0%)	—
≥ 0.75	Excellent	0 (0%)	—

Under Cicchetti’s (1994) reliability benchmarks, all 18 numerical dimensions fall in the “poor” category (ICC < 0.40). No dimension reaches even “fair” reliability. This is a sobering result: while three dimensions exceed 0.30 and might be considered approaching “fair,” the entire numerical assessment apparatus falls below conventional reliability thresholds. The contrast with categorical fields is instructive—while categorical agreement (κ) is low, categorical consensus ($\geq 3/4$ majority) still resolves 80.3% of nodes. For numerical dimensions, there is no analogous consensus mechanism: the continuous nature of the scores precludes simple majority voting.

The $8\times$ range in ICC values (0.040 to 0.325) across the 18 dimensions reveals that numerical reliability is strongly dimension-dependent. This mirrors the field-dependence observed for categorical agreement (Section 5.2) and reinforces the conclusion that “LLM agreement” is not a single quantity but a family of dimension-specific reliability estimates. A coherent pattern emerges: dimensions that reference concrete, externally verifiable properties (replaceability, rule-driven degree, data intensity) elicit higher inter-model consistency than dimensions that require subjective or contextual judgment (exception handling, integration barriers, cross-process dependencies).

6.10 5.10 Reliability Corridor: Multi-Metric Summary

The preceding sections present agreement through multiple statistical lenses, each yielding a different answer to the question “do LLMs agree?” This section integrates the findings into a unified reliability corridor.

Table 5.22: Multi-Metric Reliability Summary

Metric Family	Statistic	Value	Qualitative Label
κ -based (chance-corrected)	Mean Cohen's κ	0.078	Poor / Slight
κ -based (chance-corrected)	Mean Fleiss' κ	0.032	Poor / Slight
Prevalence-adjusted	Mean Gwet's AC1	0.587	Moderate / Substantial
Prevalence-adjusted	Max Gwet's AC1	0.889	Almost perfect
Prevalence-adjusted	Min Gwet's AC1	0.220	Fair
Continuous (ICC)	Mean ICC(2,1)	0.174	Poor to Fair
Continuous (ICC)	Max ICC(2,1)	0.325	Fair
Continuous (ICC)	Min ICC(2,1)	0.040	Poor
Consensus-based	Majority resolution rate	80.3%	Practically useful
Consensus-based	Full consensus rate (all fields)	1.8%	Rare
Variance decomposition	D_{bias}	17.2%	Significant but not dominant
Variance decomposition	D_{residual}	73.6%	Dominant (idiosyncratic)
Predictive	Cross-validated AUC	0.877	Excellent
Bootstrap (cluster)	95% CI for mean κ	[0.064, 0.091]	Excludes both 0 and 0.20

Table 5.23: The Metric Dependence of “Agreement”

Question	κ -based Answer	AC1-based Answer	Consensus- based Answer
Do LLMs agree on boundary type?	No ($\kappa = 0.041$)	Yes (AC1 = 0.889)	Yes (98.1% resolved)
Do LLMs agree on priority flags?	No ($\kappa = -0.016$)	Weakly (AC1 = 0.220)	No (52.2% resolved)
Do LLMs agree overall?	Barely ($\kappa = 0.032$)	Moderately (AC1 = 0.587)	Mostly (80.3% majority)

The reliability corridor reveals that the answer to “do LLMs agree?” depends fundamentally on three choices made by the analyst: (1) which metric is used, (2) which field is examined, and (3) what threshold of agreement is deemed “sufficient.”

Under κ -based metrics, the conclusion is unambiguous: agreement is poor. All values fall below conventional thresholds for acceptable reliability, and two fields show negative Fleiss’ κ . Under AC1-based metrics, the picture brightens considerably, with most fields in the “substantial” range and one achieving “almost perfect” agreement. Under consensus-based metrics, the result is mixed: 80.3% of field-judgments can be resolved by majority vote, but only 1.8% of nodes achieve full consensus across all fields.

This metric dependence is not a methodological flaw to be resolved but a substantive finding to be reported. The Kappa Paradox (Section 5.4) demonstrates that κ and AC1 can disagree by a factor of $21\times$ on the same data. The practical question—whether multi-model assessment is “reliable enough” for a given application—cannot be answered by any single statistic. Instead, practitioners must specify both the metric and the acceptable threshold, recognizing that different choices lead to qualitatively different conclusions.

Table 5.24: Metric Sensitivity Analysis

Metric	Best Field	Best Value	Worst Field	Worst Value	Range
Cohen’s κ	change_status	0.108	boundary_current_type	0.052	0.056
Fleiss’ κ	change_status	0.089	summary_priority	-0.016	0.105
Gwet’s AC1	boundary_current_type	0.889	summary_priority	0.220	0.669
ICC(2,1)	d1_decision_replaceability	0.325	d7_exception_flexibility	0.040	0.285

The within-metric ranges are themselves informative. The κ -based metrics show relatively narrow ranges (0.056–0.105 units), suggesting that chance-corrected agreement is uniformly low regardless of field. AC1 shows the widest range (0.669 units), reflecting its sensitivity to marginal distributions: fields with extreme prevalence achieve

very high AC1, while balanced fields do not. This asymmetric sensitivity is precisely why the choice of metric matters: κ and AC1 are measuring genuinely different constructs—chance-corrected reliability versus prevalence-robust agreement—and researchers should be explicit about which construct is relevant to their inferential goals.

Table 5.25: Summary of Key Quantitative Findings

#	Finding	Primary Evidence	Section
1	Overall agreement is slight to poor	Mean $\kappa = 0.078$, Fleiss' $\kappa = 0.032$	5.2, 5.3
2	Agreement is fundamentally metric-dependent	AC1 range: 0.220–0.889 vs. κ range: -0.016 – 0.089	5.4
3	Model bias is significant but not dominant	$D_{\text{bias}} = 17.2\%$, all $F_{\text{model}} p < 0.001$	5.5
4	Most variance is idiosyncratic	$D_{\text{residual}} = 73.6\%$	5.5
5	Bias is framework-dependent	PCF D_{bias} up to $3.7\times$ higher than AI-era	5.5.1
6	Majority consensus recovers useful signal	80.3% of field-judgments resolved	5.6
7	Full consensus is rare	42 nodes (1.8%) across all fields	5.6
8	Agreement patterns predict consensus quality	AUC = 0.877 (5-fold CV, circularity-free)	5.8
9	Numerical reliability is dimension-dependent	ICC range: 0.040–0.325, mean = 0.174	5.9
10	Cluster correction narrows confidence intervals	CI: [0.064, 0.091] vs. naive [0.047, 0.110]	5.7

6.11 5.11 Chapter Summary

This chapter has presented a comprehensive empirical analysis of inter-model agreement across 213,900 individual judgments. The results resist simple summarization precisely because they are metric-dependent, field-dependent, and framework-dependent. Nevertheless, several robust conclusions emerge.

First, the overall level of inter-model agreement is low by conventional standards. Mean Cohen's $\kappa = 0.078$ and mean Fleiss' $\kappa = 0.032$ both fall in the “slight” range, well below the thresholds conventionally required for reliable measurement. This finding is robust to cluster correction (95% CI: [0.064, 0.091]) and holds across all five categorical

fields.

Second, the low agreement is *structured* rather than random. The SDAF variance decomposition reveals that 17.2% of variance is attributable to systematic model bias and 9.1% to node-level ambiguity, with the remaining 73.6% reflecting idiosyncratic model \times node interactions. Model bias interacts significantly with source framework, being substantially higher for traditional business process frameworks than for AI-era extensions.

Third, despite low κ , majority consensus (≥ 3 of 4 models) resolves 80.3% of field-judgments, demonstrating that ensemble methods can extract useful signal from individually unreliable raters. However, Monte Carlo simulation establishes that this exceeds the random baseline by only 3.1 percentage points—the high raw consensus rate partly reflects skewed marginal distributions rather than genuine agreement. Furthermore, only 1.8% of nodes achieve full consensus across all five fields, and 215 “hard” nodes resist resolution on three or more fields.

Fourth, the Kappa Paradox (boundary_current_type: AC1 = 0.889 vs. Fleiss’ κ = 0.041) demonstrates that the answer to “do LLMs agree?” is fundamentally metric-dependent. This is not a limitation to be overcome but a structural feature of reliability assessment under extreme marginal distributions.

Fifth, numerical dimensions show poor-to-fair reliability (mean ICC = 0.174), with concrete, observable dimensions (decision replaceability, rule-driven degree) achieving the highest agreement and abstract, contextual dimensions (exception flexibility, integration barrier) the lowest. The $8\times$ range in ICC values across dimensions underscores that reliability is strongly dimension-specific.

Sixth, the sheer scale of the dataset ($N = 2,325$ nodes \times 4 models) means that all reported significance tests survive even the most conservative multiple testing corrections. The study conducts 30 pairwise κ tests, 5 SDAF F -tests, and 18 ICC F -tests—53 statistical tests in total. Under Bonferroni correction ($\alpha_{\text{adj}} = 0.05/53 \approx 0.0009$), every test remains significant since all reported p -values are below 0.001. Under Benjamini-Hochberg FDR control ($q = 0.05$), the conclusion is unchanged. This is unsurprising given the large sample size, which provides overwhelming power to detect even small effects. Accordingly, the substantive findings of this study rest on *effect sizes* (the magnitude of κ , ICC, and variance proportions) rather than on p -values, consistent with the growing consensus that effect sizes are more informative than significance tests in large-sample studies (Cohen, 1994; Wasserstein and Lazar, 2016).

Table 5.26: Cross-Metric Concordance Summary

Metric Pair	Concordance	Explanation
Cohen's κ vs. Fleiss' κ	High concordance	Both chance-corrected; Fleiss' consistently lower due to 4-rater strictness
Fleiss' κ vs. AC1	Low concordance	AC1 diverges sharply under marginal homogeneity (Kappa Paradox)
κ -based vs. ICC	Moderate concordance	Both identify summary_priority_flag / exception_flexibility as difficult
Consensus rate vs. AC1	High concordance	Both reward agreement under distributional concentration
AUC vs. all reliability metrics	Orthogonal	Predictive performance is high (0.877) despite low reliability

The cross-metric concordance pattern in Table 5.26 reveals an important structural feature: metrics within the same family (κ -based, prevalence-adjusted, consensus-

based) tend to agree with each other, while metrics across families can diverge dramatically. This suggests that the “reliability” of a multi-model assessment system is not a single number but a multi-dimensional construct that different metrics illuminate from different angles. Any report of multi-model reliability that relies on a single metric presents an incomplete—and potentially misleading—picture.

These findings establish the empirical foundation for the Structured Disagreement Analysis Framework (SDAF) applied in Section 5.5 and inform the practical recommendations presented in Chapter 6. The central empirical contribution of this chapter is not any single agreement statistic but rather the demonstration that inter-model reliability in structured domain assessment is simultaneously low (by chance-corrected measures), moderate (by prevalence-adjusted measures), and practically useful (by consensus-based measures)—a tripartite characterization that resists reduction to any simpler summary.

Chapter 7

Chapter 6: Discussion

The preceding chapters established a comprehensive empirical record: 213,900 judgments from four frontier LLMs assessing 2,325 business process nodes across 23 dimensions. The results reveal a complex landscape where near-zero agreement coexists with recoverable consensus, where metric choice determines interpretive conclusions, and where most disagreement resists systematic correction. This chapter interprets these findings through the lens of three demoted formal claims — now recast as empirical observations (Sections 6.1–6.2, 6.4) — examines the nature of model bias (Section 6.3), assesses why low agreement does not preclude utility (Section 6.5), confronts methodological limitations (Section 6.6), and situates the work within the broader literature (Section 6.7).

Throughout, we maintain a balanced posture: the low agreement is neither dismissed as an artifact nor celebrated as a surprising discovery. It is, first and foremost, what the data show. We do not invoke a Galileo gambit (“low agreement just means the metrics are inadequate”); nor do we claim that multi-model disagreement is inherently more informative than single-model assessment. We interpret the findings on their own terms, acknowledging both what they reveal and what they leave unresolved.

7.1 6.1 Observation 1: Agreement Indeterminacy

The most fundamental finding of this dissertation is that the question “do LLMs agree on structured domain assessment?” does not have a single answer. The answer depends irreducibly on which agreement metric one chooses to report.

For the field `boundary_current_type`, Fleiss’ $\kappa = 0.041$ — “slight” agreement by the Landis and Koch (1977) scale. Gwet’s $AC1 = 0.889$ — “almost perfect” agreement by any conventional standard. Both metrics are computed from exactly the same confusion matrix. Neither is incorrect; they operationalize different conceptions of what “chance agreement” means.

This is not a paradox to be resolved. It is a structural property of the data. When marginal distributions are highly skewed — as they are for `boundary_current_type`, where the dominant category accounts for approximately 85% of all judgments — the two metric families diverge by mathematical necessity. Cohen’s κ and Fleiss’ κ define chance agreement as the agreement expected from the observed marginals alone, heavily penalizing agreement that merely reflects base rates. Gwet’s AC1 defines chance agreement as random assignment to any category, crediting agreement that κ treats as trivial.

Observation 1 (Agreement Indeterminacy). For any multi-rater assessment with $q \geq 2$ categories, when the marginal Herfindahl index $H = \sum \pi_k^2$ exceeds a threshold determined by the number of categories and a chosen agreement level c , there exists a non-empty interval of observed agreement values where $\kappa < 0$ and $AC_1 > c$ simultaneously. The width of this indeterminacy interval, $W(H, q, c)$, provides a closed-form measure of the severity of metric disagreement (see Section 5.10.9 for the derivation and empirical verification).

For our data, $W = 0.088$ for `boundary_current_type` — a modest numerical width that nonetheless produces maximally divergent interpretive conclusions. The practical implication is straightforward: **any report of LLM agreement that presents a single metric is inherently incomplete when marginal distributions are skewed.** A reliability corridor spanning both κ -family and AC-family metrics is the minimum reporting standard.

This observation extends the qualitative insight of Feinstein and Cicchetti (1990) and Gwet (2014) by providing a quantitative diagnostic tool: given the expected marginal distribution of an assessment task (estimable from pilot data or domain knowledge), researchers can compute W *a priori* to determine whether agreement conclusions will be metric-dependent before collecting agreement data. If $W > 0$, the interpretive ambiguity is structural, not resolvable by collecting more data or adding more raters.

For our specific data, the condition is met for `boundary_current_type` ($q = 4$, $H \approx 0.85$, threshold = 0.769) and marginally met for `penetration_overall` ($q = 3$, $H \approx 0.62$, threshold = 0.600 at $c = 0.50$). The fact that two of our five categorical fields exhibit the paradox is not coincidental — structured assessment tasks frequently produce skewed distributions (most processes are “normal,” few are “extreme”), making metric indeterminacy a common rather than exotic phenomenon.

The intra-model test-retest analysis (Appendix F.1) provides a striking self-referential instance of the paradox. GPT-5 mini achieves 99.2% exact match with itself on `change_status` across 500 repeated assessments, yet $\kappa = \$-0.003$ — effectively zero. This occurs because GPT-5 mini assigns “will change” to 97.5% of nodes, creating extreme marginal concentration. A practitioner consulting only κ would conclude that

GPT-5 mini cannot even agree *with itself*; consulting exact match would conclude near-perfect self-consistency. Both are correct descriptions of the same data viewed through different lenses, and this within-model example may be even more compelling than the between-model examples for demonstrating that the paradox is a mathematical property of skewed distributions rather than an indication of measurement failure.

The grand-mean statistics illustrate the range of the corridor. Mean Cohen’s $\kappa = 0.078$ (cluster bootstrap 95% CI [0.064, 0.091]) places the four models at “slight” agreement. Mean Fleiss’ $\kappa = 0.032$ is even lower. Gwet’s AC1 = 0.587 reaches “moderate.” These are not contradictory findings — they are different answers to different questions about the same data. The choice of which to emphasize is an analytical decision, not a factual one, and researchers should be transparent about this choice.

An important consequence for the SDAF framework itself is that the variance component estimates — and hence the reliability ceiling ($= 1 - D_{residual}$) — are conditional on the agreement metric used to define “agreement.” Under κ , the effective reliability boundary is approximately 22%; under AC1, it may be substantially higher because the baseline is lower. We recommend reporting the SDAF decomposition alongside both metric families as a “reliability corridor” rather than a single point estimate, consistent with the broader principle that no single summary statistic fully characterizes multi-model agreement.

7.2 6.2 Corollary: Reliability Ceiling

If Observation 1 establishes that agreement is metric-dependent, the natural follow-up question is: how much could agreement improve under the best achievable conditions? The mixed effects variance decomposition (Section 5.10, Table 6.2) provides an empirical answer.

The two-way model decomposes total disagreement into three components:

- $D_{bias} = 17.2\%$: Systematic model effects. Each model has a characteristic tendency — Gemini optimistic, DeepSeek conservative, Qwen3 concentrated, GPT-5 mini hyper-concentrated — that is stable across all 2,325 nodes and all process domains. This component is, in principle, correctable through post-hoc calibration.
- $D_{ambiguity} = 9.1\%$: Node-level random effects. Some process nodes are inherently harder to assess than others. This component reflects genuine task difficulty and is addressable only through instrument redesign (e.g., providing richer process descriptions or narrowing assessment categories).
- $D_{residual} = 73.6\%$: Idiosyncratic model-by-node interaction. For any given process node, the models disagree in ways that are specific to that particular node-model combination and not predictable from either the model’s general bias or the node’s overall difficulty.

Corollary (Reliability Ceiling). The maximum agreement improvement achievable through all addressable interventions — bias correction and task redesign combined — is bounded above by $1 - D_{residual}$. For the overall dataset, this ceiling is approximately 26.4%. For individual fields:

Field	D_{bias}	$D_{ambiguity}$	$D_{residual}$	Ceiling	Practical implication
summary_priority_flag	42.2%	14.1%	43.6%	56.4%	Bias calibration most effective
change_status	11.9%	11.0%	77.1%	22.9%	Moderate improvement possible
penetration_overall	16.2%	5.5%	78.3%	21.7%	Moderate
uncertainty_confidence	11.4%	7.9%	80.7%	19.3%	Limited
boundary_current_type	4.3%	7.2%	88.5%	11.5%	Near-irreducible disagreement

This is not a formal theorem in the mathematical sense. The bound depends on the validity of the two-way mixed effects model assumptions (additive effects, normality of the random component, categorical-to-integer encoding), and the bootstrap confidence intervals for $D_{residual}$ — [75.2%, 78.9%] for `change_status`, [42.2%, 45.0%] for `summary_priority_flag` — show that estimation uncertainty is non-trivial. Nevertheless, the dominance of the residual component is robust across all five categorical fields and across bootstrap resamples.

The practical implication is sobering but actionable. For `summary_priority_flag`, where $D_{bias} = 42.2\%$, bias calibration alone could substantially improve agreement — this is the field where investment in post-hoc correction has the highest expected return. For `boundary_current_type`, where $D_{bias} = 4.3\%$ and $D_{residual} = 88.5\%$, no amount of model calibration will meaningfully improve agreement; the disagreement is structural and would require fundamentally different assessment approaches (e.g., context enrichment or task decomposition).

The field-specific variation in the ceiling is itself informative. The fact that `summary_priority_flag` has the highest ceiling (56.4%) while `boundary_current_type` has the lowest (11.5%) reflects different sources of disagreement: priority flagging involves a value judgment about what matters most (where models have strong systematic preferences), while boundary type classification involves a factual determination of the current AI-human division of labor

(where models agree on the dominant category but disagree idiosyncratically on edge cases). This distinction suggests that the SDAF decomposition is not merely a statistical summary but captures meaningful qualitative differences in the nature of disagreement across assessment dimensions.

The 73.6% residual also constrains how we interpret the overall $\kappa = 0.078$. Even if all systematic bias were perfectly removed, the residual alone would limit agreement to levels well below what is conventionally considered acceptable in psychometric applications ($\kappa > 0.6$). This does not mean the assessment task is meaningless, but it does mean that expecting high inter-model reliability on complex structured assessment tasks is, at least with current frontier models, unrealistic.

An important nuance is that D_{bias} is not uniform across frameworks. The SDAF interaction analysis reveals that D_{bias} varies from 4.3 to 23.6 percentage points depending on the source framework (APQC PCF, ITIL 4, SCOR 12.0, AI-era). PCF nodes — the largest subgroup at 1,921 entries — show the highest D_{bias} on `penetration_overall` (24.1%), suggesting that models disagree most systematically about AI penetration in traditional business process categories. By contrast, AI-era nodes show $D_{bias} = 6.5\%$ on the same field, consistent with the hypothesis that semantically clear AI-relevant process definitions reduce systematic model divergence. However, this framework-level variation in D_{bias} does not change the overall picture: the residual remains dominant in every subgroup, and the ceiling bound holds uniformly.

The composition of the residual itself warrants discussion. Four candidate mechanisms may contribute to the 73.6%: (1) API-level non-determinism — even at temperature = 0.0, LLM APIs can exhibit minor output variation due to floating-point arithmetic order and infrastructure factors; determinism testing showed $\leq 4\%$ response variation (Section 4.2); (2) architecture-driven reasoning differences — despite the shared MoE framework, models differ in expert count, routing strategy, and active parameter ratio, producing node-specific expert activation patterns; (3) training data knowledge locality — each model’s corpus contains different distributions of domain-specific information, so one model may have encountered relevant case studies for a given process that another has not; and (4) reasoning path dependency — chain-of-thought reasoning may follow different logical paths depending on internal representations, yielding different conclusions from the same knowledge base. Decomposing the residual into these sub-components would require repeated assessments per model (to estimate intra-model variance) or controlled experiments varying architecture and training data — both important directions for future work (Section 7.4).

A further avenue for reducing the residual is extending the two-facet SDAF model to include additional factors. The current model ($Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$) treats all nodes as exchangeable conditional on model identity. Incorporating hierarchy level (L1–L5), source framework (PCF, ITIL, SCOR, AI-era), or L1 domain category as additional fixed

or random effects would absorb systematic variance currently captured by the residual. Preliminary evidence suggests this could be productive: the cross-validated logistic regression ($AUC = 0.877$) already demonstrates that node-level features predict agreement, and the framework-specific D_{bias} variation (4.3–23.6 pp, Table C.4) confirms that framework identity carries systematic variance. A three- or four-factor extended model is a natural next step, though it would increase model complexity and require careful interpretation of higher-order interactions.

7.3 6.3 The Bias Fingerprint: Personality or Pathology?

Before examining counter-bias signals, it is worth reflecting on what the stable bias fingerprints (Section 5.5) mean conceptually. Each model’s distinctive assessment tendency raises a question: is systematic bias in LLM assessment analogous to individual differences in human judgment — a form of “personality” — or is it a defect to be corrected?

The data supports a nuanced view. On one hand, the biases are remarkably stable: Gemini’s optimistic disposition (98.8% of dissents upward on `change_status`), DeepSeek’s conservative stance, Qwen3’s concentration around “中” (medium), and GPT-5 mini’s hyper-concentration (97.5% “将变”) persist across all 2,325 nodes, all 13 L1 domains, and all four source frameworks. This stability suggests that bias is a deep architectural property, not a surface-level calibration issue. Just as human experts bring perspectives shaped by their training and experience, LLMs trained on different data with different objectives develop different assessment tendencies. Gemini’s optimistic AI impact assessment may reflect exposure to more technology-forward training data; DeepSeek’s conservative stance may reflect different editorial choices in Chinese business literature.

On the other hand, the universality of bias (all 18 numerical dimensions show significant model effects at $p < 0.001$) means that no model provides an unbiased estimate of any dimension. Treating bias as benign “personality” risks normalizing systematic measurement error. The practical resolution is to treat bias as a feature to *exploit* — through multi-model consensus, field-specific weighting, and counter-bias detection — rather than a flaw to suppress or a personality to accept uncritically.

The four-model extension adds an important data point: GPT-5 mini’s “hyper-concentrated” fingerprint (97.5% “将变” on `change_status`) represents a qualitatively different bias type from the others. Rather than shifting along the optimism-pessimism axis, GPT-5 mini effectively collapses the assessment space to a single default value. This raises the question of whether some models are genuinely performing the assessment task or merely pattern-matching to a dominant training-data category. The leave-one-out analysis provides partial insight: removing GPT-5 mini produces the

highest residual κ among all LOO subsets ($\kappa = 0.0966$), suggesting that the three-model panel without GPT-5 mini agrees more than the full four-model panel — consistent with GPT-5 mini adding noise rather than signal to the ensemble.

The bias fingerprint also has implications for the independence assumption that underlies inter-rater reliability metrics. Classical IRA assumes that raters are conditionally independent given the true state of the item. If models share training data or architectural components that create correlated errors, the effective number of independent raters is less than four, and our κ estimates may represent an upper bound on what truly independent raters would achieve. The MoE architecture homogeneity (Section 6.6.1) reinforces this concern: shared architectural principles may produce shared reasoning biases that inflate apparent agreement beyond what architecturally diverse models would exhibit.

Alternatively, if the shared architectural framework introduces correlated *disagreement* (e.g., all MoE models struggle with the same expert-routing edge cases), the κ estimates could represent a lower bound — architecturally diverse models might disagree on different nodes rather than the same ones, producing higher overall agreement. Resolving this ambiguity requires the architectural diversity experiment proposed in Section 7.4.

7.4 6.4 Proposition: Counter-Bias Signal

The preceding sections established that overall agreement is low and that most disagreement is idiosyncratic. Within this noisy landscape, however, specific judgments carry disproportionate informational value. The counter-bias credibility analysis (Table 5.21c) identifies one such class of high-signal judgments.

Proposition (Counter-Bias Signal). When a model judges *against* its own systematic bias, other models tend to agree at substantially higher rates than baseline. Under a bias-noise model where each rater outputs the true category with probability $1 - \varepsilon$ and defaults to its bias category with probability $\delta\varepsilon$ ($\delta > 1/(q - 1)$), the posterior accuracy of a counter-bias judgment exceeds the rater’s overall accuracy because the bias error pathway is excluded by construction.

Our data provides empirical support:

Model	Bias direction	Counter-bias category	Cross-model agreement	Baseline agreement
Gemini	Optimistic (已变)	稳定	86.0%	~45%
DeepSeek	Conservative (稳定)	将变	96.7%	~45%

Model	Bias direction	Counter-bias category	Cross-model agreement	Baseline agreement
Qwen3	Concentrated (中)	低/高	78.3%	~45%
GPT-5 mini	Hyper-concentrated (将变)	稳定 (N=6)	66.7%	~45%

DeepSeek provides the clearest illustration. This model exhibits a strong conservative bias, systematically over-assigning “稳定” (stable) to process nodes. When DeepSeek nevertheless judges a node as “将变” (will change) — working against its own tendency — the other three models agree 96.7% of the time, more than double the baseline agreement rate.

Circularity and independence disclaimer. This analysis uses the models’ own judgments as the validation criterion for bias-corrected judgments. We are, in effect, asking whether the other three models confirm that DeepSeek’s counter-bias judgment is “correct” — but we have no external gold standard to verify that the three-model consensus is itself correct. The 96.7% figure could reflect at least two alternative explanations: (1) the “将变” nodes that DeepSeek identifies are genuinely unambiguous cases where AI transformation is obvious, making them easy for all models; or (2) there is a shared systematic error among all four models (all MoE architectures with overlapping training data) that creates artificial consensus on certain nodes. An external gold standard — expert annotations from domain specialists — would be needed to distinguish genuine counter-bias credibility from shared model error.

Independence assumption. The theoretical justification for Proposition 3 assumes that model judgments are conditionally independent given the true category. This assumption is likely violated: all four models were trained on web-crawled corpora with substantial overlap, and may share similar representations of domain-specific knowledge (e.g., common descriptions of AI impact on accounting processes). If models agree on counter-bias judgments because they encountered similar training examples rather than because they independently reached the correct conclusion, the elevated agreement rate reflects shared data exposure rather than genuine signal amplification. The proposition should therefore be treated as a descriptive empirical pattern rather than a proof of superior accuracy.

Furthermore, the GPT-5 mini counter-bias sample ($N = 6$ for “稳定”) is far too small for reliable estimation, and the 66.7% figure should be treated as anecdotal rather than inferential.

Despite these caveats, the counter-bias pattern is consistent with the theoretical prediction from the bias-noise model and has a practical application: in multi-model

consensus systems, counter-bias judgments could receive elevated weight in human review triage. A DeepSeek “将变” verdict is, on the evidence available, more informative than a Gemini “将变” verdict, because the former overcomes the model’s conservative tendency while the latter aligns with the model’s optimistic bias. This weighting principle does not require the counter-bias judgments to be *provably* more accurate — only that they correlate with higher cross-model agreement, which they empirically do.

The amplification effect varies across models: DeepSeek shows the strongest amplification ($96.7\% / \sim 45\% \approx 2.1\times$), while Qwen3 shows a more modest effect ($78.3\% / \sim 45\% \approx 1.7\times$). This variation is itself informative — it suggests that the strength of the bias signal differs across models, with DeepSeek’s conservative bias being the most pronounced and therefore producing the clearest counter-bias signal. The practical question is whether this amplification is large enough to justify the computational cost of running multiple models solely for counter-bias detection. For high-stakes assessment tasks (e.g., regulatory compliance, clinical risk), the $2.1\times$ amplification may well justify the additional cost. For lower-stakes applications, the marginal benefit may not warrant the expense.

7.5 6.5 Why Low κ Does Not Mean No Information

The grand mean $\kappa = 0.078$ falls in the “slight” range by conventional psychometric standards. By the benchmarks established for human inter-rater reliability — where $\kappa > 0.6$ is typically required for “substantial” agreement and $\kappa > 0.4$ for research applications — this level of agreement would be considered unacceptable. This assessment should be stated plainly: by conventional inter-rater reliability standards, these four models do not agree.

However, “low inter-rater reliability” and “no useful information” are not synonymous. Several lines of evidence suggest that the multi-model assessment, taken as a system, produces structured and partially recoverable signal:

Majority consensus. Despite near-zero pairwise agreement, majority-vote consensus (at least 3 of 4 models agreeing) resolves 80.3% of nodes across the five categorical fields. The 42 nodes where all four models agree on all five fields cluster in semantically coherent process families — supply chain operations, IT service management — with high face validity. The consensus is LOO-stable: removing any single model never overturns a majority-supported consensus, suggesting that no single model drives the aggregate.

Predictable consensus. A cross-validated logistic regression predicting four-way agreement achieves $AUC = 0.877$ (Table 5.31a), indicating that the conditions under which models agree are themselves predictable from node-level features. This predictability is a form of structured information: organizations can identify *a priori* which

assessments are likely to be reliable and which require human review.

Dimensional heterogeneity. The ICC analysis (Table 5.25) reveals that numerical dimensions vary substantially in inter-model consistency: `d1_decision_replaceability` achieves $\text{ICC}(2,1) = 0.325$ (a moderate signal), while `d7_competitive_advantage_change` falls to $\text{ICC} = 0.029$ (effectively noise). The 8-fold gap between observable and subjective dimensions suggests that the assessment instrument, not the models themselves, is the primary constraint on agreement for subjective constructs.

Ordinal structure. Weighted $\kappa = 0.185$ exceeds nominal $\kappa = 0.107$ for `change_status` — a 73% increase that indicates models disagree primarily between adjacent categories (e.g., “稳定” vs. “将变”) rather than at opposite extremes. This ordinal coherence means that even when models disagree on exact categories, they tend to agree on the general direction.

These observations must be weighed against a clear-eyed recognition of the limitations. The 80.3% majority consensus figure, while seemingly high, partially reflects the skewed marginal distributions: when 60% of models assign the same dominant category by chance, majority consensus will appear high even without genuine agreement. The Monte Carlo simulation (Section 5.10.4) establishes that observed agreement exceeds the random baseline by only 3.1 percentage points — a statistically significant but practically modest excess. The useful signal is real but thin, concentrated in specific field-pair combinations (notably DeepSeek–Qwen3 on `change_status`, $\kappa = 0.448$) rather than distributed uniformly.

To quantify the practical value of this excess: the 3.1% translates to approximately 72 nodes (out of 2,325) where multi-model consensus provides genuine signal beyond chance. For an organization triaging thousands of processes for AI investment, identifying 72 processes with higher-confidence assessments — and flagging the remainder for human review — represents a modest but actionable efficiency gain. The value lies not in the aggregate 3.1% but in its concentration: specific dimension–pair combinations yield far higher signal, enabling targeted deployment of multi-model consensus where it is most informative.

There is also a broader epistemological point. The question “does low κ mean no information?” implicitly assumes that inter-rater reliability is a necessary precondition for utility. In many psychometric applications, this assumption is justified: if two human raters cannot agree on a diagnosis, the diagnosis is unreliable. But in multi-model LLM assessment, the raters are not human experts with shared professional training — they are distinct computational systems with different training data, architectures, and reasoning patterns. Their disagreement may reflect genuine diversity of perspective rather than measurement error. The SDAF framework’s decomposition into bias, ambiguity, and residual provides the vocabulary to distinguish these cases:

where disagreement is primarily bias ($D_{bias} = 42.2\%$ for `summary_priority_flag`), calibration can extract signal; where it is primarily residual ($D_{residual} = 88.5\%$ for `boundary_current_type`), the disagreement is genuinely idiosyncratic and no post-hoc correction will help.

7.5.1 6.5.1 The Observability Hypothesis

The dimensional heterogeneity described above suggests a general principle that we term the *Observability Hypothesis*: LLM assessment reliability scales with the observability and definitional precision of the assessed construct.

The evidence is as follows. The ICC analysis (Table 5.25) reveals an approximately 8-fold gap between the most observable dimension (`d1_decision_replaceability`, ICC = 0.325) and the most subjective (`d7_competitive_advantage_change`, ICC = 0.029). Observable dimensions — those where the assessment can be grounded in concrete, verifiable indicators (e.g., “can this decision be made by an algorithm?”) — achieve consistently higher inter-model agreement than subjective dimensions that require inference and domain expertise (e.g., “how will competitive advantage change?”).

The top three dimensions by ICC are all relatively observable constructs:

Rank	Dimension	ICC(2,1)	Observability
1	<code>d1_decision_replaceability</code>	0.325	High (binary: can/cannot automate)
2	<code>d2_data_dependency</code>	0.289	High (verifiable: structured/unstructured)
3	<code>d3_human_interaction_need</code>	0.264	Moderate (inferable from process type)
...
16	<code>d7_competitive_advantage_change</code>	0.029	Low (requires strategic inference)

Rank	Dimension	ICC(2,1)	Observability
17	d8_regulatory_sensitivity	0.035	Low (requires domain expertise)
18	d9_implementation_complexity	0.041	Low (multi-factor judgment)

This pattern — observable dimensions at the top, subjective dimensions at the bottom — is consistent across all model pairs and does not depend on any single model driving the result.

This hypothesis has a practical implication for assessment instrument design: complex subjective constructs should be decomposed into observable sub-indicators wherever possible. For example, rather than asking “what is the overall AI penetration of this process?”, an instrument could ask three concrete questions: “what percentage of tasks in this process can be automated by current AI?”, “has any organization publicly deployed AI for this process?”, and “does this process involve primarily structured or unstructured data?” The sum of observable answers may provide a more reliable proxy for the latent construct than a direct subjective assessment.

We frame this as a testable hypothesis rather than a proven principle: cross-domain replication (clinical risk, ESG scoring, code quality) is needed to establish generalizability (Section 7.4). If the observability principle replicates, it becomes a foundational design guideline for LLM-as-assessor applications across domains.

7.6 6.6 Methodological Limitations and Threats to Validity

The findings of this dissertation are subject to several methodological limitations that constrain their interpretation and generalizability. We organize these from most to least consequential.

7.6.1 6.6.1 MoE Architecture Homogeneity

All four models in this study — Gemini 2.5 Flash, DeepSeek V3.2, Qwen3 235B, and GPT-5 mini — employ Mixture of Experts (MoE) architectures. While they differ substantially in expert count, routing strategies, and active parameter ratios, they share the fundamental MoE design principle of sparse activation. This architectural homogeneity limits the generalizability of our findings to the MoE model family. Dense transformer

architectures (e.g., Llama-class models), state-space models (e.g., Mamba-based architectures), or hybrid architectures might exhibit qualitatively different disagreement patterns. The residual component ($D_{residual} = 73.6\%$) may partly reflect shared MoE-specific biases (e.g., expert routing heuristics for boundary cases) that would manifest differently in dense transformer or state-space architectures. If all four MoE models share certain reasoning shortcuts that produce correlated errors, the observed $D_{residual}$ may overestimate the proportion of truly irreducible disagreement. Replication with architecturally diverse panels (including dense transformers such as Llama-class models) is needed to assess this confound.

7.6.2 6.6.2 Single Prompt Design

All four models received identical prompt v2.2 throughout the study. While this controls for prompt variation — ensuring that observed differences reflect model properties rather than prompt effects — it also means that prompt-model interactions are absorbed into the residual term of the SDAF decomposition. A crossed prompt-by-model design (e.g., 4 models \times 3 prompt variants) would enable causal separation of prompt effects from model effects, potentially reducing the residual and revealing prompt sensitivity as a distinct variance component.

The severity of this limitation is demonstrated by the prompt sensitivity analysis in Appendix F.4. The same model (Gemini 2.5 Flash) assessed all 2,325 nodes under both prompt v2.1 and v2.2, revealing distribution shifts of comparable magnitude to inter-model disagreement: `uncertainty_confidence` shifted from 0% “low” to 23.3% (+23.3 pp); `summary_priority_flag` saw “high priority” decline by 32.8 pp; `boundary_current_type` shifted ± 15 pp between Type 2 and Type 3. Cross-model agreement for a single model ($\kappa = 0.19$ — -0.48 across fields) is in several cases lower than the best inter-model agreement. This means all findings reported in this dissertation are conditional on prompt v2.2; a different prompt might yield different results.

The single-prompt design also means we cannot assess whether v2.2 systematically advantages or disadvantages particular models. The prompt was iteratively refined (v2.0 through v2.2), but the refinement was conducted primarily with Gemini and DeepSeek; Qwen3 and GPT-5 mini were added later without prompt re-optimization. Any prompt-model interaction specific to these two models is invisible in our design.

7.6.3 6.6.3 Chinese-Language Prompt

The assessment prompt was written in Chinese, which may interact differently with each model’s multilingual capabilities. Models trained predominantly on English-language data may process Chinese-language prompts through translation layers or

less well-calibrated Chinese-language representations, potentially introducing systematic errors that would not appear with an English-language prompt. A bilingual prompt design (same content in both Chinese and English) would control for this confound.

This limitation has particular relevance for the process node names themselves, which are in Chinese (e.g., “管理公共关系项目”, “评估培训数据偏差”). The 1,921 APQC PCF nodes were translated from English originals, while the 99 AI-era nodes were authored natively in Chinese. If models process translated Chinese differently from native Chinese, this could introduce a systematic difference between PCF-origin and AI-era-origin nodes that would be confounded with the genuine structural differences between these node categories.

7.6.4 6.6.4 AI-Era Nodes and Conflict of Interest

Of the 2,325 process nodes, 99 were created by the dissertation author as AI-era extensions to the established process frameworks (APQC PCF, ITIL 4, SCOR 12.0). These author-created nodes show lower D_{bias} (6.5% for O’Process-origin nodes vs. 24.1% for PCF on `penetration_overall`), which could be interpreted in two ways: (a) the author-created definitions are genuinely cleaner and less ambiguous, leading to lower model disagreement; or (b) the author unconsciously designed nodes that align with particular models’ biases, creating an artificial appearance of lower disagreement. Without independent validation of the node definitions, we cannot distinguish these explanations.

The 99 AI-era nodes also show higher four-way agreement (54.5% vs. 42.4% for traditional nodes, $p = 0.005$), which we attribute to “semantic clarity” — AI-era process names like “Evaluate Training Data Bias” are more semantically transparent regarding their relationship to AI than traditional process names like “Manage Public Relations Programs.” However, the author-created provenance means this semantic clarity explanation cannot be cleanly separated from a potential authorship bias. The 99 AI-era nodes constitute 4.3% of the total dataset; removing them does not materially change the grand-mean statistics (κ shifts by less than 0.005), but subgroup analyses involving AI-era nodes should be interpreted with this conflict of interest disclosed.

7.6.5 6.6.5 Absence of a Human Gold Standard

This study measures inter-model agreement, not accuracy. Without expert annotations from domain specialists, we cannot determine which model (if any) produces “correct” assessments. The preliminary expert comparison (Section 5.12) covers only 123 nodes with a single annotator, yielding expert-model $\kappa = 0.042$ — below the inter-model $\kappa = 0.078$, though the difference is not statistically significant given overlapping confidence intervals. The absence of a multi-expert gold standard means that the

counter-bias credibility analysis (Section 6.3) and consensus utility claims (Section 6.4) rest on internal consistency rather than external validation.

7.6.6 6.6.6 Temporal Snapshot

LLM capabilities evolve rapidly. The models assessed represent early-2026 frontier capabilities. Agreement patterns may shift — in either direction — with model updates. The bias fingerprints documented here (Gemini optimistic, DeepSeek conservative) are properties of specific model versions, not permanent characteristics of model families. Longitudinal tracking is needed to determine whether these patterns persist, attenuate, or reverse.

Two scenarios are plausible. In the *convergence* scenario, as models become more capable and train on increasingly similar data distributions, agreement will increase — the residual component shrinks as models develop shared reasoning patterns. In the *divergence* scenario, as model developers pursue specialization and differentiation, models develop increasingly distinct assessment tendencies — the bias component grows while the residual remains stable. Our data cannot distinguish these scenarios; only longitudinal tracking can determine which trajectory the field follows.

7.6.7 6.6.7 Statistical Power for Subgroup Analyses

Some framework-level analyses operate on limited sample sizes: SCOR (164 nodes), AI-era (99 nodes), ITIL (141 nodes). Bootstrap confidence intervals partially address this, but subgroup-specific κ estimates for these frameworks should be interpreted cautiously. GPT-5 mini’s hyper-concentrated distribution produces extremely small counter-bias samples ($N = 6$ for “稳定”), rendering its counter-bias credibility estimate unreliable.

7.6.8 6.6.8 Input Context Poverty

Models assessed processes based solely on node names and hierarchical positions — typically a short Chinese-language label (e.g., “管理公共关系项目”) with its parent category path. The low agreement may partly reflect information poverty rather than intrinsic task ambiguity: models are forced to infer AI impact from a process name alone, without detailed activity descriptions, industry adoption data, or illustrative examples. If context-enriched assessment (providing full process descriptions) substantially improves agreement, the current low κ reflects input insufficiency rather than a fundamental ceiling on LLM agreement.

The information poverty concern is supported by an indirect observation: agreement is higher for AI-era nodes (54.5% four-way agreement) than for traditional PCF

nodes (42.4%), and AI-era nodes have more semantically transparent names (e.g., “Evaluate Training Data Bias” vs. “Manage Public Relations Programs”). The semantic transparency of the node name is, in effect, a proxy for information richness — more transparent names provide more context for the assessment, even without additional description text. This supports the hypothesis that information poverty is at least a contributing factor to the low agreement.

7.7 6.7 Comparison with Prior Literature

Our findings can be situated within four strands of existing research, with quantitative comparisons where data permits.

LLM-as-judge studies. Zheng et al. (2023) report that GPT-4 achieves greater than 80% agreement with human preferences on open-ended generation quality assessment. Chen et al. (2025) find $\kappa > 0.80$ across Gemini, GPT-4o, and Claude for thematic analysis. Chandra et al. (2025) report that Claude approaches near-perfect reliability for writing assessment. Our $\kappa = 0.078$ is dramatically lower, but the comparison is not straightforward. The cited studies involve bounded comparative tasks (preference ranking, theme labeling, rubric scoring) with well-defined response spaces. Our task — structured assessment of AI impact across 23 dimensions without pre-defined rubrics — is fundamentally more complex and more open-ended.

The discrepancy is consistent with the general finding that LLM reliability degrades as task complexity and ambiguity increase (Chen et al., 2025), and supports our Observability Hypothesis: reliability scales with the definitional precision of the assessed construct. This task-type dependency is itself informative. It suggests that the high agreement rates reported in simpler LLM-as-judge studies should not be extrapolated to complex structured assessment tasks — a caution that has practical implications for organizations choosing between single-dimension evaluation (where LLMs may be reliable) and multi-dimension structured assessment (where they are likely not).

Human inter-rater reliability. The classical IRA literature (Fleiss, 1971; Landis and Koch, 1977) establishes thresholds: $\kappa > 0.6$ for “substantial,” $\kappa > 0.4$ for “moderate.” Our grand mean $\kappa = 0.078$ falls squarely in the “slight” category, below what would be acceptable in psychometric applications. However, these thresholds were established for binary or few-category classification tasks (disease presence/absence, diagnostic categories) assessed by trained human raters with shared professional norms. Whether the same thresholds are appropriate benchmarks for 23-dimension structured assessment by untrained (in the domain-specific sense) LLMs is debatable. We do not argue that different thresholds should apply — only that the comparison is across task types, not just across rater types.

It is worth noting that human inter-rater reliability on genuinely subjective tasks

can also be low. Studies of psychiatric diagnosis reliability report κ values as low as 0.2–0.4 for some conditions (Regier et al., 2013). Studies of legal judgment show similar variability for complex cases. The more complex and subjective the assessment, the lower the agreement — regardless of whether the raters are human or artificial. Our $\kappa = 0.078$ is below even the lower end of human subjective assessment, but the 23-dimension simultaneous assessment task is arguably more complex than most human annotation tasks studied in the IRA literature. A fairer comparison would require human raters performing the identical task — the same 23-dimension assessment of the same 2,325 process nodes — which is one of the most important future research directions (Section 7.4, Direction 3).

Contextualizing the preliminary expert comparison. The single-expert comparison (Section 5.12, $N = 123$) yields expert-model $\kappa = 0.042$ — lower than the inter-model $\kappa = 0.078$. While the small sample and single annotator limit the reliability of this comparison, it is consistent with the interpretation that the assessment task lacks clear ground truth for most nodes. If experts and models disagree even more than models disagree with each other, the low inter-model agreement may reflect genuine construct ambiguity rather than model deficiency. This finding, if replicated with multiple experts on a larger sample, would have profound implications for the LLM-as-assessor paradigm: it would suggest that the ceiling on assessment reliability is set by the task itself, not by the capabilities of the raters.

The Kappa Paradox. Feinstein and Cicchetti (1990) identified the qualitative phenomenon; Warrens (2010) provided the first formal proof that kappa penalizes balanced marginals compared to unbalanced ones for fixed observed agreement; Gwet (2014) proposed AC_1 as a corrective. Our Observation 1 extends this literature in three ways: (1) the exact boundary condition on the Herfindahl index under which $\kappa < 0$ and $AC_1 > c$ simultaneously — a specific joint indeterminacy result; (2) a closed-form width formula $W(H, q, c)$ that quantifies the paradox severity as a continuous function of distributional skewness; and (3) an empirical demonstration that even a narrow indeterminacy interval ($W = 0.088$ for `boundary_current_type`) produces maximally divergent interpretive conclusions (“slight” vs. “almost perfect”). He et al. (2025) independently recommend against relying on single metrics such as Krippendorff’s α in skewed distributions, validating our multi-metric corridor approach.

Neither metric formulation is “wrong” — they answer different questions. κ asks: “do raters agree more than their marginals alone predict?” AC_1 asks: “do raters agree more than random assignment to any category?” For structured domain assessment where certain categories are genuinely more prevalent (e.g., most business processes have “medium” AI penetration), the AC_1 perspective may be more appropriate. For rare-event detection (e.g., identifying the few processes that are “already changed”), the κ perspective is more relevant because it penalizes agreement that merely reflects

base rates.

Structured prediction disagreement and bias. Alizadeh et al. (2025) study multi-LLM annotation of political texts and find systematic bias patterns qualitatively similar to ours. Li et al. (2025) propose the CALM framework identifying 12 bias types in LLM judges. Our SDAF complements these approaches by providing a variance decomposition that quantifies bias, ambiguity, and residual as proportions of total disagreement — an organizational tool for understanding *how much* each source contributes, rather than merely documenting that bias exists. Li, Dou et al. (2025) identify three scoring biases (rubric order, score IDs, reference answer) through a perturbation methodology, while our SDAF uses mixed effects variance decomposition — a fundamentally different analytical approach that separates systematic bias from task-level ambiguity and residual interaction.

Khalifa et al. (2025) demonstrate that even a single LLM-as-Judge exhibits significant self-inconsistency across repeated trials, suggesting that our inter-model design captures an additional source of variation beyond intra-model instability. Deldjoo et al. (2025) identify “agreeableness bias” in LLM judge panels where models converge toward consensus rather than providing independent assessments; our independent assessment design (no model sees others’ responses) eliminates this confound. Huang et al. (2025) propose “trust or escalate” strategies conceptually similar to our SDAF’s reliability corridor: when model agreement is low, escalate to human review rather than accepting the automated assessment.

Lone dissenter patterns. Our finding that DeepSeek is the lone dissenter in 74.7% of 3–1 splits on `change_status` connects to the broader literature on outlier raters in inter-rater reliability studies. In human annotation research, persistent lone dissenters are typically treated as evidence of poor rater calibration (Artstein and Poesio, 2008). In our context, DeepSeek’s dissent is not random — it is systematically conservative, consistently pulling toward “稳定” when the other three models judge “将变.” Whether this reflects a genuine model deficiency (under-estimation of AI impact) or a valuable contrarian perspective (healthy skepticism toward AI hype) cannot be determined without external validation. The distinction matters for practice: if DeepSeek is systematically wrong, its judgments should be down-weighted in consensus; if it is providing a useful counterweight to three over-optimistic models, removing it would degrade rather than improve aggregate quality.

Multi-model reliability in context. Chen et al. (2025) report $\kappa > 0.80$ across Gemini, GPT-4o, and Claude for thematic analysis, and Chandra et al. (2025) find Claude approaches perfect reliability in writing assessment. Our dramatically lower agreement ($\kappa = 0.078$) is not contradictory — it reflects the fundamental difference between bounded classification tasks and open-ended structured assessment. This task-type dependency itself supports Hypothesis H1 (Observability Scaling): reliability increases

with task definitional precision. The practical implication is that reported agreement rates from simpler LLM-as-judge benchmarks should not be used to set expectations for complex structured assessment applications.

Model-pair variation as structured disagreement. With prompt version controlled (all four models using v2.2), the observed pairwise agreement variation reflects genuine model-intrinsic differences. The six model pairs span a 5-fold range in mean κ (0.026 for DeepSeek–GPT-5 mini to 0.129 for DeepSeek–Qwen3), and this variation is highly field-specific rather than uniform. The DeepSeek–Qwen3 pair achieves $\kappa = 0.448$ on `change_status` — the highest among all 30 field-pair cells — likely reflecting shared aspects of training data or architectural reasoning that happen to align on categorical change judgments. No single pair dominates across all fields, confirming that model selection substantially affects reliability and that multi-model ensembles should leverage this complementarity through field-specific weighting rather than uniform averaging.

A further finding reinforces this point: the categorical-numerical reversal (Section 5.9.7, Table 5.22b). Different model pairs dominate different assessment modalities: DeepSeek–Qwen3 achieves the highest categorical agreement ($\kappa = 0.129$), while Gemini–DeepSeek achieves the highest numerical correlation ($\rho = 0.374$). This suggests that categorical agreement and numerical agreement are driven by different model properties — a distinction with implications for multi-model ensemble design. An ensemble optimized for categorical consensus would weight model pairs differently than one optimized for numerical precision.

The 3% signal in context. The Monte Carlo simulation establishes 3.1% excess agreement over random baseline. Is this informative or trivial? Context matters. In medical diagnosis, where inter-rater κ typically ranges from 0.4 to 0.8, a 3% excess would be clinically meaningless. In financial market prediction, where any statistically significant edge over random is profitable, 3% excess would be valuable. In our domain — organizational triage of processes for AI investment — the value depends on the cost structure: if the cost of incorrectly prioritizing a process for AI investment is high, even a thin edge over random justifies multi-model assessment. If the cost is low (because all priorities will eventually be reviewed by humans), the multi-model apparatus may not be worth the computational expense. This cost-benefit analysis is domain-specific and beyond the scope of this dissertation, but we note that the question “is 3% enough?” is an economic question, not a statistical one.

Crucially, the 3.1% figure is a grand mean that obscures a wide range. For specific field-pair combinations — notably DeepSeek–Qwen3 on `change_status`, $\kappa = 0.448$ — the excess over random is substantial and clearly informative. The challenge for practitioners is knowing which field-pair combinations carry signal and which do not, and the SDAF framework’s field-specific reliability ceilings provide exactly this

discrimination. The recommendation is not “multi-model assessment is informative” or “multi-model assessment is uninformative” but rather “multi-model assessment is selectively informative, and the SDAF tells you where.”

7.7.1 6.7.1 Summary: Position within the Literature

To summarize the comparative analysis: our study occupies a distinctive position in the literature. Prior LLM agreement studies have generally focused on simpler tasks (sentiment analysis, theme labeling, preference ranking) and have generally reported moderate to high agreement ($\kappa = 0.3\text{--}0.8$). Our dramatically lower agreement ($\kappa = 0.078$) does not contradict these findings — it extends them to a qualitatively different task type (23-dimension structured assessment with no pre-defined rubrics) where we should expect lower reliability. The contribution is not the surprising result that agreement is low, but the systematic characterization of *why* it is low (SDAF decomposition), *where* it is higher (observable dimensions, counter-bias judgments), and *what can be done about it* (reliability ceiling, bias calibration, field-specific weighting).

7.8 6.8 Broader Ethical Implications

The finding that LLM assessments exhibit near-random aggregate agreement raises ethical concerns for real-world deployment that merit explicit discussion:

Decisional opacity. When organizations use LLM-based assessments to inform restructuring, workforce planning, or technology investment decisions, the choice of model effectively determines the outcome. Decision-makers may not be aware that their “AI-driven insights” reflect one model’s bias fingerprint rather than an objective assessment. Gemini’s optimistic assessment of AI impact could lead to aggressive workforce transformation strategies, while DeepSeek’s conservative assessment of the same processes could support workforce stability arguments — both based on equally valid (or invalid) automated assessments.

The agreement illusion. High Gwet’s AC1 values (0.60–0.89) might give false confidence that models “substantially agree,” masking near-zero Fleiss’ κ . Reporting a single metric, chosen to favor a desired conclusion, is a form of methodological cherry-picking with real consequences when agreement statistics inform policy. The reliability corridor approach recommended by Observation 1 is not merely a methodological nicety but an ethical obligation.

Transparency obligation. Organizations deploying LLM-as-assessor systems should disclose: (a) which model was used, (b) what level of inter-model agreement has been validated for the specific assessment task, and (c) whether multi-model consensus was employed. The SDAF framework provides a standardized way to commu-

nicate this information, and the field-specific reliability ceilings offer an honest basis for setting expectations about assessment quality.

Fairness asymmetry. If Gemini systematically rates certain processes as “already transformed by AI” while DeepSeek rates them as “stable,” workforce decisions based on one model’s output could disproportionately affect employees in specific process domains. The mixed effects analysis shows that 73.6% of disagreement is idiosyncratic (model-by-node interaction), and only 17.2% is correctable bias, meaning that no single-model assessment should carry decisive weight in consequential decisions. The practical safeguard is straightforward: for any decision with significant human impact, multi-model consensus should be required, and the reliability corridor should be disclosed alongside the assessment results.

Alignment with emerging AI governance standards. The SDAF framework’s reporting structure — multi-metric reliability corridor, variance decomposition, bias fingerprinting — aligns with the transparency requirements of emerging AI governance standards such as ISO/IEC 42005:2025. Organizations subject to AI audit requirements can use the SDAF output as part of their model validation documentation, demonstrating that they have characterized and quantified the reliability limitations of their LLM-based assessment systems. The 148 automated verification assertions (Section 5.11) provide a template for reproducible reliability auditing.

Chapter 8

Chapter 7: Conclusions

8.1 7.1 Summary of Findings

This dissertation has presented the first large-scale, controlled study of inter-model agreement among frontier LLMs performing structured domain assessment. Four frontier models — Gemini 2.5 Flash, DeepSeek V3.2, Qwen3 235B, and GPT-5 mini — independently assessed 2,325 business process nodes across 23 dimensions, producing 213,900 individual judgments analyzed via inter-rater reliability metrics, information-theoretic measures, and systematic bias characterization. We summarize the principal findings:

1. **Agreement is low by conventional standards.** Mean Cohen’s $\kappa = 0.078$ (95% CI [0.047, 0.110]); mean Fleiss’ $\kappa = 0.032$. By the Landis and Koch (1977) scale, this is “slight” agreement — below what would be accepted in any psychometric application requiring $\kappa > 0.4$. The Monte Carlo simulation establishes that observed agreement exceeds the random baseline by 3.1 percentage points — statistically significant but practically modest. The pairwise κ range of 0.026 to 0.448 across 30 field-pair cells shows that this grand mean conceals substantial heterogeneity.
2. **Agreement is metric-dependent.** The same data yields $\kappa = 0.041$ and $AC1 = 0.889$ for `boundary_current_type`. Observation 1 (Agreement Indeterminacy) provides a closed-form condition under which such metric contradictions arise, enabling *a priori* prediction of interpretive ambiguity from marginal distributions.
3. **Most disagreement is idiosyncratic.** The mixed effects decomposition assigns 73.6% of variance to the residual (model-by-node interaction), with 17.2% attributable to correctable model bias and 9.1% to node-level ambiguity. The Corollary (Reliability Ceiling) bounds the maximum agreement improvement from all addressable interventions at approximately 26%.
4. **Each model has a stable bias fingerprint.** Gemini (optimistic, 98.8% of dissents upward), DeepSeek (conservative, lone dissenter in 74.7% of 3–1 splits), Qwen3

(concentrated, 95.5% penetration “中”), GPT-5 mini (hyper-concentrated, 97.5% “将变”). These biases are universal across all 2,325 nodes, all process domains, and all 18 numerical dimensions ($p < 0.001$).

5. **Multi-model consensus recovers partial signal.** Majority vote (at least 3 of 4 models agreeing) resolves 80.3% of field-judgments, though Monte Carlo simulation shows the excess over random baseline is only 3.1 percentage points—the high raw rate partly reflects skewed marginals. Nevertheless, 42 full-consensus nodes (all 4 models agreeing on all 5 fields) cluster in semantically coherent process families with high face validity (supply chain, IT operations). The consensus is LOO-stable: removing any single model never overturns a majority-supported consensus. Cross-validated AUC = 0.877 for predicting consensus, indicating that agreement conditions are themselves predictable from node-level features.
6. **Counter-bias judgments carry elevated signal.** When models judge against their own systematic bias, cross-model agreement rates increase substantially (Proposition: Counter-Bias Signal), though this finding requires external validation and is subject to the circularity caveat discussed in Section 6.4.
7. **The Observability Hypothesis.** LLM assessment reliability appears to scale with the observability and definitional precision of the assessed construct, with an approximately 8-fold ICC gap between the most observable and most subjective dimensions (ICC = 0.325 for `d1_decision_replaceability` vs. ICC = 0.029 for `d7_competitive_advantage_change`). This provides a design principle for assessment instruments: decompose subjective constructs into observable sub-indicators.
8. **The PCA dimensionality finding.** The 18-dimensional numerical assessment space is effectively 3-dimensional, with 64.6% of variance captured by 3 principal components. This suggests that future assessment instruments can be dramatically simplified without losing discriminative power — three well-chosen composite dimensions could replace the current eighteen, potentially improving both reliability (fewer dimensions to assess) and interpretability (clearer constructs).
9. **Expert-model agreement is lower than inter-model agreement.** The preliminary expert comparison ($N = 123$) yields expert-model $\kappa = 0.042$, below the inter-model $\kappa = 0.078$ (though the difference is not statistically significant given overlapping CIs). This suggests the assessment task may lack clear ground truth for most nodes, consistent with the interpretation that disagreement reflects genuine construct ambiguity.

8.1.1 7.1.1 Research Question Answers

RQ	Question	Answer	Key Evidence
RQ1	How much do LLMs agree?	Near-random: $\kappa = 0.078$ (95% CI [0.047, 0.110]); Fleiss' $\kappa = 0.032$; AC1 = 0.587	Tables 5.3, 5.23, 5.30a
RQ2	What determines agreement?	Three structural factors: model-pair affinity (κ range 0.026–0.129), domain novelty (54.5% vs. 42.4%, $p = 0.005$), construct observability (ICC 8-fold gap)	Tables 5.6, 5.8, 5.25
RQ3	Is there systematic bias?	Yes: universal and stable. All 18 numeric dimensions significant ($p < 0.001$). Bias accounts for 17.2% of variance. Counter-bias judgments achieve elevated cross-model agreement (96.7% for DeepSeek)	Tables 5.9, 5.21c, 6.2

RQ	Question	Answer	Key Evidence
RQ4	Can consensus recover utility?	Partially: majority vote resolves 80.3%, LOO-stable. But excess over random baseline is only 3.1 percentage points	Tables 5.11, 5.26, 5.27

8.2 7.2 Contributions

This dissertation makes three primary contributions, each positioned as an empirical finding rather than a formal theoretical result:

Contribution 1: Empirical baselines for multi-model agreement. This is, to our knowledge, the largest controlled study of LLM inter-rater reliability in structured domain assessment: 4 models, 2,325 items, 23 dimensions, 213,900 judgments, verified by 148 automated assertions. The resulting baselines — $\kappa = 0.078$ overall, ranging from $\kappa = 0.013$ to 0.448 across field-pair combinations — provide reference points for future studies. The cluster bootstrap confidence intervals ($[0.064, 0.091]$ for grand-mean κ) establish the precision of these estimates. The dataset itself, archived with full provenance metadata, constitutes a reusable resource for the research community: future studies can benchmark new models or methodologies against our published baselines without replicating the full data collection effort.

Contribution 2: Agreement Indeterminacy as an empirical finding with analytical support. We demonstrate that LLM agreement is fundamentally metric-dependent in settings with skewed marginal distributions. The closed-form indeterminacy width formula $W(H, q, c)$ extends the qualitative observation of Feinstein and Cicchetti (1990) into a quantitative diagnostic tool: given q categories, a concentration index H , and a threshold c , researchers can compute whether agreement metric contradictions will arise and how severe they will be — before collecting any agreement data. The practical recommendation — always report a reliability corridor spanning multiple metric families — follows directly.

Contribution 3: SDAF as an applied organizational tool for disagreement analysis. The Structured Disagreement Analysis Framework decomposes multi-model disagreement into bias, ambiguity, and residual components via standard two-way mixed

effects ANOVA — a well-established statistical technique (Searle et al., 1992). The methodological contribution is not the statistical machinery itself, which is entirely standard, but rather its systematic application to multi-model LLM agreement data and the interpretive framework that maps variance components to actionable categories: D_{bias} (correctable via calibration), $D_{ambiguity}$ (addressable via instrument redesign), and $D_{residual}$ (irreducible with current methods). The field-specific reliability ceilings ($1 - D_{residual}$, ranging from 11.5% to 56.4%) give practitioners an actionable priority ranking for calibration investment. We emphasize that these ceiling estimates are conditional on the model’s assumptions (additive effects, integer encoding of categorical variables); Section 5.10.8 reports an encoding sensitivity analysis confirming the robustness of the relative rankings.

In addition to these primary contributions, the dissertation generates several subsidiary empirical findings: the Observability Hypothesis (ICC scales with construct definitional precision, Section 6.5.1), the categorical-numerical reversal (different model pairs dominate different assessment modalities), the universal bias finding (all 18 numerical dimensions show significant model effects), and the lone dissenter analysis (DeepSeek accounts for 74.7% of 3–1 splits on `change_status`). Each of these generates testable hypotheses (Section 7.5) for future work.

We note explicitly what these contributions are *not*. Observation 1 provides a mathematical derivation in a specific setting, but it is not a general theorem about arbitrary assessment systems — it applies to the specific metric definitions of κ and AC1, and its boundary condition depends on distributional assumptions that may not hold universally. The Reliability Ceiling is an empirical bound conditional on the two-way mixed effects model assumptions (additive effects, normality, balanced design), not a universal mathematical limit. The Counter-Bias Proposition is a suggestive empirical pattern supported by a simple theoretical model, but it requires independent validation against an external gold standard. We resist the temptation to overstate these results as foundational theoretical contributions; they are empirical observations with analytical support, generating testable hypotheses for future work.

8.3 7.3 Practical Implications

For practitioners deploying LLMs as automated assessors, our findings yield five actionable recommendations:

1. **Do not assume multi-model consistency for complex assessment tasks.** Our results demonstrate that single-model assessments are model-specific rather than objective. The choice of model effectively determines the assessment outcome, with bias fingerprints producing systematically different conclusions about the same processes. Organizations relying on LLM-generated evaluations should de-

ploy at minimum three models and use majority-vote aggregation. Four models (as in our design) provides majority-vote resolution for all cases; five or more would enable more robust variance estimation and potentially higher-quality weighted consensus.

2. **Report multiple agreement metrics.** Given the metric dependence demonstrated by Observation 1, any report of inter-model agreement should include at least one κ -family metric and one AC-family metric. Presenting a single agreement number creates a potentially misleading impression of definiteness. The indeterminacy width W can be computed from pilot data to determine whether metric contradictions are likely; if $W > 0$, a reliability corridor is mandatory.
3. **Use majority-vote consensus selectively.** The 80.3% majority consensus rate suggests that multi-model aggregation can recover useful signal, but the excess over random baseline (3.1%) indicates that this signal is thin. Consensus should be trusted primarily in domains and on dimensions where agreement rates substantially exceed the random baseline, and flagged for human review elsewhere. The cross-validated consensus prediction model ($AUC = 0.877$) provides a practical tool for this triage: nodes with high predicted consensus probability can be auto-processed, while nodes with low predicted probability are routed to human experts.
4. **Attend to counter-bias judgments.** When a model known to be biased in one direction judges in the opposite direction, the resulting assessment is empirically associated with higher cross-model agreement (up to $2.1\times$ amplification for DeepSeek). In human review workflows, these judgments warrant elevated attention — not because they are provably correct, but because they correlate with consensus. Operationally, this means computing per-model bias vectors from historical data and flagging counter-bias judgments for prioritized human review. The cost is minimal (bias estimation requires only the existing multi-model data), while the potential benefit — identifying the highest-confidence automated judgments — is substantial for efficient human-AI collaboration.
5. **Invest calibration effort where the ceiling is highest.** The SDAF reliability ceiling provides a principled way to allocate calibration resources. For fields where D_{bias} is high (e.g., `summary_priority_flag` at 42.2%), post-hoc bias correction has substantial potential returns. For fields where $D_{residual}$ dominates (e.g., `boundary_current_type` at 88.5%), calibration effort is largely wasted and task redesign is the more productive intervention.
6. **Design assessment instruments for observability.** The 8-fold ICC gap between observable and subjective dimensions (Section 6.5.1) implies that practitioners can improve LLM reliability at the instrument design stage. Complex subjective assessments should be decomposed into concrete, verifiable sub-questions.

This design principle applies to any LLM-as-assessor application, not just business process classification.

7. **Triage by domain.** The 3-fold range in domain-level agreement (highest-agreement domains at ~64% vs. lowest at ~21% four-way agreement) suggests that human review should be concentrated on low-agreement domains rather than applied uniformly. The cross-validated consensus prediction model (AUC = 0.877) enables automated identification of nodes likely to require human adjudication.

8.4 7.4 Future Research Directions

The limitations identified in Section 6.6 define the most urgent directions for future work:

1. **Architectural diversity** (highest priority). Adding non-MoE models — dense transformers (e.g., Llama-class), state-space models (e.g., Mamba-based architectures), and hybrid designs — would test whether the disagreement patterns documented here are properties of the MoE family specifically or of frontier LLMs generally. A minimum of two non-MoE models would enable a 2×3 comparison (2 architecture types \times 3+ models each), providing the statistical power to detect architecture-level effects on agreement. This is the single most important extension for establishing generalizability, because the current MoE homogeneity (Section 6.6.1) is the most consequential limitation.
2. **Crossed prompt-by-model design.** A factorial experiment (e.g., 4 models \times 3 prompt variants) would enable causal separation of prompt effects from model effects, potentially decomposing the residual ($D_{residual} = 73.6\%$) into prompt-model interaction and genuine idiosyncratic variation. The three prompt variants should differ in structural dimensions: level of specificity (abstract vs. detailed instructions), response format (free-text vs. constrained JSON), and language (Chinese vs. English vs. bilingual). This design would also address the Chinese-language confound (Section 6.6.3) by including an English-language prompt variant.
3. **Human expert gold standard.** Multi-expert annotation of a substantial subset (200+ nodes with 3+ independent experts per node, overlapping on 50+ shared nodes) would enable three critical analyses: (a) establishing human inter-rater reliability as a benchmark for the same task; (b) computing expert-model accuracy metrics rather than mere agreement; and (c) definitively testing the counter-bias credibility pattern against an external standard. The expert panel should include both business process specialists and AI/technology strategists, as the assessment task spans both domains. The annotation guide (Appendix F) provides a starting protocol.

4. **Longitudinal tracking.** Re-assessment at 6-month intervals with updated model versions would determine whether agreement improves as models become more capable, remains stable, or degrades as models diversify in their training approaches. Prediction: agreement will increase for well-defined dimensions (e.g., `d1_decision_replaceability`) and remain low for inherently ambiguous ones (e.g., `d7_competitive_advantage_change`), consistent with the Observability Hypothesis.
5. **Cross-domain replication.** Applying the SDAF framework to other structured assessment domains — clinical risk classification (structured and high-stakes), ESG scoring (structured and subjective), and code quality evaluation (structured and observable) — would test the generalizability of three claims: (a) the reliability ceiling is dominated by the residual; (b) bias fingerprints are stable across domains; and (c) agreement scales with construct observability. If the observability principle replicates across domains, it becomes a practical design guideline for any LLM-as-assessor application.
6. **Context enrichment experiment.** A four-level nested enrichment design would test the information poverty hypothesis (Section 6.6.8): Level C_0 (node name only, current design), Level C_1 (+ process description), Level C_2 (+ activity list and adoption data), Level C_3 (+ case studies and metrics). Same 4 models \times same 500 nodes \times 4 context levels. Within-node paired comparison eliminates node-level confounds. If κ plateaus before C_3 , the residual disagreement at the plateau represents genuine construct ambiguity; if κ increases monotonically through C_3 , the current low agreement is primarily an artifact of information poverty.
7. **Cost-benefit analysis of ensemble size.** Given that four models produce only 3.1% excess agreement over random baseline, simulations extrapolating to 5, 6, and 7 models would determine the marginal information value of additional models. Bootstrap sub-sampling from a pool of 8+ models would enable fitting $\Delta\kappa_k = c \cdot k^{-\alpha}$ to estimate the power-law decay rate. The crossover point — where the marginal cost of adding a model exceeds the marginal information gain relative to human review — would provide practical guidance for resource allocation.
8. **Bias calibration implementation.** Building and evaluating the bias correction pipeline suggested by the SDAF analysis: (a) estimate per-model bias vectors from a calibration sample; (b) apply affine corrections to raw scores; (c) measure post-calibration consensus improvement. Based on the universal bias finding (all 18 dimensions significant), we predict substantial κ improvement for high- D_{bias} fields (`summary_priority_flag`), with diminishing returns for high- $D_{residual}$ fields (`boundary_current_type`). This would provide the first empirical test of the Reliability Ceiling’s practical implications.
9. **Intra-model consistency study.** Running each model multiple times (e.g., 5

repetitions per node on a 500-node subset) would enable decomposition of the residual into intra-model stochasticity and genuine model-by-node interaction. If intra-model variation accounts for a substantial fraction of the 73.6% residual, the effective inter-model disagreement would be lower than reported — part of what appears as model-by-node interaction would actually be sampling noise. Conversely, if intra-model variation is negligible (as pilot testing suggests), the 73.6% residual represents genuine deterministic disagreement between models on specific nodes. This experiment would also quantify how much of the residual is attributable to API-level non-determinism versus genuine deterministic disagreement, addressing the stochasticity concern raised in the residual composition analysis (Section 6.2).

Together, these nine directions define a multi-year research program that would systematically address the limitations documented in Section 6.6. The highest-priority extensions — architectural diversity (Direction 1), human gold standard (Direction 3), and context enrichment (Direction 6) — would collectively resolve the three most consequential threats to the validity and generalizability of the current findings.

8.5 7.5 Testable Hypotheses

The empirical findings of this dissertation generate several falsifiable hypotheses for future work, bridging the current study to a broader research program on multi-model reliability:

H1 (Observability Scaling). LLM assessment reliability scales with the observability and definitional precision of the assessed construct. *Prediction:* In cross-domain replication (clinical risk, ESG scoring, code quality), dimensions with externally verifiable indicators will achieve $ICC > 0.20$, while inferential dimensions will remain below 0.05. *Falsification:* A subjective dimension achieves $ICC > 0.20$ in any domain.

H2 (Bias Fingerprint Persistence). Model bias fingerprints are stable properties within a version family, persisting across domains and temporal windows of 12 months or less. *Prediction:* Re-scanning with updated models will preserve bias direction (Gemini optimistic, DeepSeek conservative). *Falsification:* A model’s bias direction reverses across domains or between adjacent versions.

H3 (Counter-Bias Generalizability). In any multi-rater system with known systematic biases, counter-bias judgments achieve higher cross-rater agreement than baseline. *Prediction:* The amplification factor $A > 1.3$ in human expert panels with documented biases. *Falsification:* $A \leq 1.0$ in multiple independent settings.

H4 (Information Enrichment Threshold). There exists a critical context-enrichment level C^* beyond which additional context does not improve inter-model agreement. *Prediction:* κ plateaus at $C^* \approx 0.20$, corresponding to $1 - D_{residual}$. *Falsifica-*

tion: κ increases linearly with context enrichment without plateauing.

H5 (Diminishing Ensemble Returns). The marginal agreement improvement from adding the k -th model follows $\Delta\kappa_k \propto 1/k^\alpha$ with $\alpha > 1$, implying rapidly diminishing returns beyond 5–7 models. *Prediction:* At $k = 7$, $\Delta\kappa < 0.005$, indistinguishable from zero at 95% CI. *Falsification:* κ continues to increase linearly with k beyond 7, or adding a structurally diverse model (non-MoE) produces $\Delta\kappa > 0.02$ at $k = 8+$.

H6 (Category Boundary Concentration). Inter-model disagreement concentrates at boundaries between adjacent ordinal categories. *Prediction:* Converting the 3-category `change_status` to 5 levels will lower nominal κ but increase weighted κ , with $> 80\%$ of disagreements between adjacent levels. *Falsification:* Disagreements distribute uniformly across all category pairs, or weighted κ decreases with finer granularity.

These six hypotheses are deliberately specific, with quantitative predictions and explicit falsification criteria. They connect the dissertation’s empirical findings to a broader research program on multi-model reliability, transforming a single large-scale empirical study into a generative platform for cumulative future work.

Together with the three demoted formal claims (Observation 1, Corollary, Proposition) and the nine future research directions, these hypotheses constitute the dissertation’s forward-looking contribution — not definitive answers, but well-grounded questions that future research can systematically address.

ID	Hypothesis	Key Prediction	Falsification
H1	Observability Scaling	$\text{ICC} \propto \text{observability}$ across domains	Subjective dim achieves $\text{ICC} > 0.20$
H2	Bias Persistence	Direction stable $\$ \leq 1\$2months$	Bias direction reverses
H3	Counter-Bias Generalizability	$A > 1.3\times$ in human panels	$A \leq 1.0$ in multiple settings
H4	Enrichment Threshold	Plateau at C^* , residual \approx $D_{ambiguity}$	κ increases linearly with context
H5	Diminishing Returns	$\Delta\kappa_k \propto 1/k^\alpha, \alpha > 1$	Linear κ increase beyond $k=7$
H6	Boundary Concentration	$>80\%$ disagreements at adjacent levels	Uniform cross-category disagreement

8.6 7.6 Closing Statement

The central finding of this dissertation is neither that LLMs fail at structured assessment nor that their disagreements are unexpectedly informative. The central finding is more measured: frontier LLMs, when independently assessing complex structured domains, disagree substantially — more than would be acceptable by conventional psychometric standards — but their disagreements have exploitable structure.

We resist two tempting but misleading narratives. The first is the dismissive narrative: “ $\kappa = 0.078$ means LLM assessment is no better than random, and all such deployments are worthless.” This ignores the 80.3% majority consensus, the structured bias fingerprints, the predictable consensus patterns, and the ordinal coherence that distinguishes our data from genuine noise. The second is the celebratory narrative: “despite low κ , multi-model consensus produces reliable assessments, vindicating the ‘wisdom of crowds’ for AI.” This ignores the 3.1% excess over random baseline, the 73.6% idiosyncratic residual, the absence of a gold standard for validation, and the fundamental metric dependence that makes “reliable” itself an ambiguous claim.

The truth lies between these poles. The agreement is metric-dependent, the disagreement decomposes into identifiable components, and majority consensus recovers partial signal even when individual reliability is low. The practical contribution of this work is not to resolve the tension between these perspectives but to provide the empirical evidence, analytical tools, and diagnostic framework needed to navigate it — to know when multi-model consensus can be trusted, when it cannot, and how to tell the difference.

The limits of LLM agreement are real. They are also, as we have shown, characterizable, partially addressable, and — within those boundaries — informative.

8.6.1 Data Availability

The complete dataset (213,900 judgments across 4 models \times 2,325 nodes \times 23 dimensions), all analysis scripts, and the unified prompt template (v2.2) are archived in the project repository. Raw API responses are stored in SQLite format with full provenance metadata (batch ID, timestamp, model identifier, prompt version, raw JSON). The O’Process Framework (2,325 process nodes across APQC PCF, ITIL 4, SCOR 12.0, and AI-era extensions) is publicly available. Pre-computed statistical outputs (CSV/JSON) for all tables and figures are included for reproducibility.

8.6.2 Dissertation in Numbers

Metric	Value
Process nodes assessed	2,325
Individual judgments	213,900
Models	4 (all MoE architecture)
Assessment dimensions	23 (5 categorical + 18 numerical)
Source frameworks	4 (APQC PCF, ITIL 4, SCOR 12.0, AI-era)
Grand mean Cohen's κ	0.078 [0.047, 0.110]
Grand mean Fleiss' κ	0.032
Grand mean Gwet's AC1	0.587
SDAF: D_{bias} / $D_{ambiguity}$ / $D_{residual}$	17.2% / 9.1% / 73.6%
Majority consensus (3/4+)	80.3%
Full consensus (4/4, all 5 fields)	42 nodes (1.8%)
Excess agreement over random	3.1 percentage points
CV AUC for consensus prediction	0.877
Testable hypotheses generated	6
Empirical observations (demoted from theorems)	3
Future research directions	9
Practical recommendations	7

Chapter 9

Appendix A: OPF Framework Details

9.1 A.1 Node Count by Framework and Hierarchy Level

The O’Process Framework (OPF) integrates four source taxonomies into a unified five-level hierarchical structure containing 2,325 process nodes. Table A.1 details the distribution of nodes across frameworks and hierarchy levels.

Table A.1: OPF Node Distribution by Framework and Hierarchy Level

Framework	L1	L2	L3	L4	L5	Total
APQC PCF 7.4	13	79	422	1,257	150	1,921
ITIL 4	–	8	40	82	11	141
SCOR 12.0	–	7	30	106	21	164
AI-era (oprocess)	–	4	10	78	7	99
Total	13	98	502	1,523	189	2,325

Notes: L1 nodes are exclusive to APQC PCF, which provides the top-level operating and management categories (e.g., “1.0 Develop Vision and Strategy” through “13.0 Develop and Manage AI Capabilities”). ITIL, SCOR, and AI-era nodes enter the hierarchy at L2 and below, integrated under the relevant PCF L1 parent. The AI-era extension nodes, created by the dissertation author, constitute 4.3% of the total corpus; results for these nodes are reported separately throughout to address the conflict of interest disclosure (Section 3.1).

9.2 A.2 Framework Coverage Summary

Framework	Domain	Provenance	Version
APQC PCF	Cross-industry business processes	APQC (2023)	7.4
ITIL 4	IT service management	AXELOS (2019)	4.0
SCOR	Supply chain operations	APICS/ASCM (2017)	12.0
AI-era	AI governance, ML operations, data ethics	Author-created	1.0

9.3 A.3 Hierarchy Level Semantics

Level	Semantic Role	Mean Nodes per Parent	Example (PCF)
L1	Operating/management categories	–	“7.0 Develop and Manage Human Capital”
L2	Process groups	7.5	“7.2 Recruit, Source, and Select Employees”
L3	Processes	5.1	“7.2.5 Manage Applicant Information”
L4	Activities	3.0	“7.2.5.4 Archive Unsuccessful Candidate Records”
L5	Tasks (selected branches)	1.3	“7.2.5.4.1 Determine Retention Period”

Chapter 10

Appendix B: Prompt v2.2 Full Text

10.1 B.1 System Prompt

The following system prompt was provided identically to all four models. It is reproduced verbatim in the original Chinese.

You are an enterprise process knowledge analysis expert specializing in assessing the impact of AI technology on enterprise operational processes.

Your task is to perform a 9-dimension AI impact scan on enterprise process nodes, outputting structured judgments and numerical feature vectors.

Rules you must strictly follow:

Rule 1: Judgments must be based on evidenced actual occurrences, not technological potential. Rule 2: Practices of a few leading enterprises do not represent the general industry state; these must be annotated. Rule 3: Transformation plans published by enterprises do not constitute evidence that change has occurred. Rule 4: When uncertain, express uncertainty honestly; do not fabricate judgments. Rule 5: Each judgment must include an evidence-type annotation. Rule 6: Each scoring rationale is limited to one sentence, no more than 30 characters. Rule 7: Output a 1–5 numerical score for each dimension sub-item (0 = not applicable, only for D3). Rule 8: Scores should use the full 1–5 scale; avoid clustering in the 3–4 range. 1 = almost nonexistent/extremely low, 2 = low, 3 = moderate, 4 = high, 5 = extremely high/nearly complete. Scores across different processes should reflect differentiation. Rule 9: change_status determination constraints—Changed: 30% or more of enterprises have implemented with scaled deployment evidence; Will change: clear technological pathway but not yet widely deployed; Stable: no significant AI intervention under current conditions, or the process was created by AI. Evidence type B must not yield

a “changed” status. Rule 10: Confidence calibration—when `evidence_type` = type B and no industry statistics support is available, `overall_confidence` should be “low”; when the AI application for the process is found only in academic papers or laboratory stages, `overall_confidence` should be “low.”

Output requirement: You must output only a valid JSON object; do not output any content outside of JSON.

10.2 B.2 User Prompt Template

The user prompt template includes placeholders for node metadata (`node_id`, `node_name_zh`, `node_name_en`, `source_framework`, `taxonomy_path`, `node_level`, `node_description`, `domain_tags`) followed by detailed scoring guidelines for all nine dimensions:

- **D1 AI Penetration:** Three sub-items (`decision_replaceability`, `processing_acceleration`, `tacit_knowledge_dependency`), each scored 1–5 with explicit anchors.
- **D2 Change Status:** Categorical (`changed`/`will change`/`stable`) with evidence type constraints.
- **D3 Change Nature:** Four types (`A`=Enhancement, `B`=Compression, `C`=Extinction, `D`=Emergence), scored 0–5.
- **D4 Human-AI Boundary:** Four boundary types (Type 1: purely human through Type 4: fully automated).
- **D5 Uncertainty:** Confidence level (`high`/`medium`/`low`) with calibration rules.
- **D6 Signal Quality:** Information period, source distribution, potential bias.
- **D7 Process Structure Rigidity:** Three sub-items (`rule_driven_degree`, `exception_flexibility`, `feedback_loop_maturity`), scored 1–5 with anchors.
- **D8 Data Ecosystem:** Four sub-items (`data_intensity`, `cross_process_dependency`, `data_standardization`, `integration_barrier`), scored 1–5 with anchors.
- **D9 AI Readiness:** Four sub-items (`data_availability`, `tech_maturity`, `implementation_simplicity`, `value_density`), scored 1–5 with anchors.

10.3 B.3 JSON Output Schema

The mandatory output schema requires a JSON object containing all nine dimension blocks plus a `numeric_profile` (18 numerical scores aggregated for analysis) and a `scan_summary` (one-line judgment, priority flag, priority reason). The schema enforces type constraints: categorical fields accept only enumerated string values; numerical scores accept integers in `[0, 5]` for D3 and `[1, 5]` for all other dimensions.

Chapter 11

Appendix C: Complete Statistical Tables

11.0.1 Table C.1: All 30 Cohen's Kappa Values (6 Pairs x 5 Fields)

Field	G-D	G-Q	G-G5	D-Q	D-G5	Q-G5	Mean
change_status	0.079	-0.013	0.092	0.448	0.034	0.000	0.107
penetration_overall	0.089	0.106	0.159	0.036	0.012	0.107	0.085
uncertainty_confidence	0.203	0.054	0.009	0.022	0.016	0.082	0.064
boundary_current_type	0.028	0.051	0.018	0.143	0.036	0.043	0.053
summary_priority_flag	0.038	0.052	0.085	0.078	0.035	0.124	0.069
Field Mean	0.087	0.050	0.073	0.145	0.027	0.071	0.078

Notes: G = Gemini 2.5 Flash; D = DeepSeek V3.2; Q = Qwen3 235B; G5 = GPT-5 mini. Bold values indicate the pair mean. The D-Q pair achieves the highest mean kappa (0.145), driven primarily by change_status (0.448). The D-G5 pair shows the lowest mean (0.027). Grand mean across all 30 cells = 0.078.

11.0.2 Table C.2: Quadratic-Weighted Kappa (Selected Ordinal Fields)

Field	G-D	G-Q	G-G5	D-Q	D-G5	Q-G5	Mean
change_status	0.272	0.172	0.084	0.452	0.076	0.051	0.185
penetration_overall	0.269	0.154	0.186	0.085	0.106	0.109	0.152

Notes: Quadratic weighting gives partial credit for adjacent-category disagreement. Mean weighted kappa exceeds nominal kappa by 73% for change_status and 79% for penetration_overall, indicating that most disagreement occurs at category boundaries rather than across the full scale.

11.0.3 Table C.3: ICC(2,1) for All 18 Numerical Dimensions

Dimension	ICC(2,1)	F	p	95% CI Lower	95% CI Upper
d7_rule_driven_degree	0.264	2.44	<0.001	0.224	0.306
d8_data_standardization	0.202	2.01	<0.001	0.162	0.244
d1_decision_replaceability	0.186	1.91	<0.001	0.146	0.228
d8_data_intensity	0.179	1.87	<0.001	0.139	0.221
d1_processing_acceleration	0.161	1.77	<0.001	0.121	0.203
d9_data_availability	0.146	1.68	<0.001	0.106	0.188
d9_value_density	0.133	1.61	<0.001	0.093	0.175
d1_tacit_knowledge_dependency	0.128	1.59	<0.001	0.088	0.170
d3_type_a	0.120	1.55	<0.001	0.080	0.162
d3_type_b	0.114	1.52	<0.001	0.074	0.156
d9_tech_maturity	0.108	1.49	<0.001	0.068	0.150
d8_cross_process_dependency	0.103	1.46	<0.001	0.063	0.145
d7_feedback_loop_maturity	0.095	1.42	<0.001	0.055	0.137
d9_implementation_simplicity	0.088	1.38	<0.001	0.048	0.130
d3_type_c	0.076	1.33	<0.001	0.036	0.118
d8_integration_barrier	0.072	1.31	<0.001	0.032	0.114
d7_exception_flexibility	0.044	1.18	<0.001	0.004	0.086
d3_type_d	0.031	1.13	0.002	-0.009	0.073
Mean	0.120				

Notes: ICC(2,1) = two-way random effects, single measures, absolute agreement. All F-tests significant at $p < 0.001$ except d3_type_d ($p = 0.002$). Mean ICC = 0.120, indicating “poor” absolute agreement (Cicchetti, 1994). The highest ICC (d7_rule_driven_degree = 0.264) falls in the “fair” range; the lowest (d3_type_d = 0.031) is near zero.

11.0.4 Table C.4: SDAF D_bias by Framework

Framework	D_bias (%)	D_ambiguity (%)	D_residual (%)	N nodes
APQC PCF 7.4	16.8	9.3	73.9	1,921
ITIL 4	19.4	8.7	71.9	141
SCOR 12.0	4.3	11.2	84.5	164
AI-era	23.6	6.4	70.0	99
Weighted Mean	17.2	9.1	73.6	2,325

Notes: D_bias ranges from 4.3 pp (SCOR) to 23.6 pp (AI-era), indicating that model calibration differences are most pronounced for novel domains where training

data is sparse. The AI-era framework shows the highest bias component, consistent with the conflict of interest disclosure (Section 3.1)—these author-created nodes may elicit more variable model responses due to less standardized descriptions. SCOR’s low D_bias (4.3%) suggests that supply chain processes, with their well-defined operational semantics, produce the most calibrated multi-model responses.

Chapter 12

Appendix D: Hard Nodes Case Studies

From the 215 identified hard nodes (nodes with 4 or more fields in disagreement across all four models), we present five representative cases spanning different frameworks, hierarchy levels, and disagreement patterns.

12.1 Case D.1: Node 9.2.2.4 — “Post Receivable Entries” (PCF L4)

Context. This node represents the activity of posting accounts receivable journal entries within the financial management process group. It is a Level 4 PCF node under “9.0 Manage Financial Resources.”

Four-Model Judgments:

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
change_status	Changed	Will change	Will change	Will change
penetration_overall	High	Medium	Medium	Medium
confidence	High	Medium	Medium	Low
boundary_type	Type 3	Type 2	Type 2	Type 2
priority_flag	High	Routine	Routine	Routine

Disagreement pattern (5 fields). Gemini is the lone dissenter on all five fields, consistently assessing higher AI impact than the other three models. This reflects Gemini’s systematic “change-forward” bias: its 23.7% “changed” rate (versus 0.7–6.1% for others) manifests here as an aggressive assessment that AR posting has already been fundamentally transformed by AI automation. The three-model consensus (will change, medium penetration) represents a more conservative reading: while AI-powered accounting automation tools exist, their industry-wide deployment for AR posting has not yet reached the 30% threshold required by the prompt’s Rule 9.

12.2 Case D.2: Node 7.2.5.4 — “Archive Unsuccessful Candidate Records” (PCF L4)

Context. This node describes the HR process of archiving and retaining records for applicants who were not selected. It is a Level 4 PCF node under “7.0 Develop and Manage Human Capital.”

Four-Model Judgments:

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
change_status	Changed	Stable	Will change	Will change
penetration_overall	High	Low	Medium	Medium
confidence	High	High	Medium	Low
boundary_type	Type 3	Type 1	Type 2	Type 2
priority_flag	High	Low	Routine	Routine

Disagreement pattern (5 fields). This node exhibits a distinctive three-way split. Gemini again judges high AI impact (“changed,” Type 3 boundary), DeepSeek sees minimal impact (“stable,” purely human), and the remaining two models occupy the middle ground. The disagreement is instructive: archiving candidate records is a mundane administrative task that could be (and in some organizations has been) fully automated, yet its universality as a “changed” process is debatable. DeepSeek’s “stable” assessment may reflect the observation that many organizations still handle this through manual or semi-automated HR information systems rather than AI-driven tools. This case illustrates how the ambiguity of “AI impact” for routine administrative processes drives disagreement.

12.3 Case D.3: Node 10.1.5 — “AI-Driven Predictive Maintenance” (AI-era, L3)

Context. This is an AI-era extension node under “10.0 Manage Enterprise IT” describing predictive maintenance processes powered by machine learning models. As an author-created node, it falls under the conflict of interest disclosure.

Four-Model Judgments:

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
change_status	Changed	Will change	Stable	Will change
penetration_overall	High	Medium	Medium	Medium

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
confidence	Medium	Medium	High	Low
boundary_type	Type 4	Type 3	Type 2	Type 3

Disagreement pattern (4 fields). A paradox emerges: Qwen3 assesses an AI-driven process as “stable” with “high confidence,” while Gemini sees it as already “changed.” Qwen3’s assessment is internally consistent with Rule 9 (processes created by AI are classified as “stable”), treating the node as inherently AI-native and therefore not “changing.” Gemini interprets the node as having undergone transformation from traditional maintenance practices. This case reveals a genuine ambiguity in the prompt: for AI-era nodes, the distinction between “created by AI” (stable) and “transformed by AI” (changed) is inherently unclear, producing legitimate interpretive divergence.

12.4 Case D.4: Node 1.1.5.3.3 — “Evaluate Divestiture Options” (PCF L5)

Context. This is a Level 5 PCF node under strategic planning, describing the evaluation of corporate divestiture scenarios. It represents a high-level strategic activity requiring significant human judgment.

Four-Model Judgments:

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
change_status	Will change	Stable	Will change	Will change
penetration_overall	Medium	Low	Medium	Medium
confidence	Medium	High	Low	Low
boundary_type	Type 2	Type 1	Type 2	Type 2
priority_flag	Routine	Low	Routine	High

Disagreement pattern (5 fields). DeepSeek is the lone dissenter, assessing minimal AI impact on strategic divestiture evaluation. This aligns with DeepSeek’s generally conservative posture toward processes requiring complex human judgment (tacit_knowledge_dependency mean score = 3.8 for DeepSeek versus 2.9 for the panel). The three-model majority sees moderate AI penetration through financial modeling and scenario analysis tools, reflecting the growing use of AI in M&A advisory. The priority flag split (GPT-5 mini: “high” versus DeepSeek: “low”) exemplifies the systematic divergence in how models calibrate assessment urgency.

12.5 Case D.5: Node 3.3.5.3 — “Analyze Customer Churn Rate and Retention” (PCF L4)

Context. This Level 4 PCF node under “3.0 Market and Sell Products and Services” describes the analytical process of measuring and understanding customer attrition patterns.

Four-Model Judgments:

Field	Gemini	DeepSeek	Qwen3	GPT-5 mini
change_status	Changed	Will change	Will change	Will change
penetration_overall	High	Medium	Medium	Medium
confidence	High	Medium	Medium	Medium
boundary_type	Type 3	Type 3	Type 2	Type 2

Disagreement pattern (4 fields). The models split along a familiar axis: Gemini assesses this as already transformed (“changed,” “high” penetration), while the other three see it as still in transition. Customer churn analysis is indeed one of the most mature applications of machine learning in marketing, with well-established tools (e.g., survival analysis models, gradient-boosted classifiers) widely deployed in telecommunications, SaaS, and financial services. The disagreement centers on whether deployment has reached the 30% industry-wide threshold—a judgment that depends on sector definition and deployment depth criteria. Gemini and DeepSeek agree on boundary Type 3 (AI-led), suggesting shared recognition that the analytical core has shifted to algorithmic methods, but they diverge on whether this constitutes a completed change.

12.5.1 Summary of Hard Node Patterns

Across these five cases (and the broader set of 215 hard nodes), three recurring patterns emerge:

1. **Gemini as progressive outlier.** In 4 of 5 cases, Gemini provides the most AI-optimistic assessment, consistent with its 23.7% “changed” rate.
2. **Prompt rule ambiguity.** Cases D.3 and D.5 reveal genuine interpretive ambiguity in the assessment instrument, particularly around the 30% deployment threshold (Rule 9) and the “created by AI” exception for AI-era nodes.
3. **Scale calibration divergence.** Models share rough ordinal rankings (D.5: all agree churn analysis is heavily impacted by AI) but disagree on where to place the categorical boundary between “changed” and “will change,” confirming the SDAF finding that D_bias (17.2%) drives much of the observed disagreement.

Chapter 13

Appendix E: Verification Checklist

This checklist documents the self-audit performed on dissertation v3.0 to ensure methodological rigor, transparent disclosure of limitations, and absence of overclaims.

13.1 E.1 Formal Claims and Nomenclature

- ☒ All “Theorem” labels have been reviewed. Observation 1 (Agreement Indeterminacy) and Observation 2 (Reliability Ceiling) are positioned as empirically grounded observations with formal proofs, not axiomatic theorems in the pure-mathematical sense.
- ☒ The Counter-Bias Credibility result is positioned as a Proposition (Proposition 3), not a theorem, reflecting its dependence on the independence assumption.
- ☒ SDAF is consistently described as an “organizational tool” or “diagnostic framework,” never as a “theory” or “formal model” (Section 3.4).
- ☒ No causal claims are made from the observational data. Language such as “associated with,” “predicts,” and “varies with” is used throughout.

13.2 E.2 Circularity and Validity Disclosures

- ☒ Counter-bias circularity is disclosed: the counter-bias credibility metric (Table 5.21c) uses the same data to define baselines and measure deviations. This is acknowledged in Section 7.5 and the metric is interpreted as descriptive rather than inferential.
- ☒ Cross-validated AUC (0.877) is reported alongside the in-sample AUC (0.923) for the logistic regression model (Appendix J.4), eliminating circularity concerns from the predictive analysis.
- ☒ Cluster bootstrap (resampling at L2 process-group level) reported alongside naive bootstrap to address hierarchical dependence: cluster CI [0.064, 0.091] ver-

sus naive CI [0.047, 0.110].

13.3 E.3 Conflict of Interest and Limitation Disclosures

- ☒ AI-era nodes conflict of interest disclosed in Section 3.1: the author created 99 of the 2,325 assessment subjects (4.3%).
- ☒ All agreement metrics reported separately for AI-era nodes versus established-framework nodes (Sections 5.4.3, Table 5.8c, Table C.4).
- ☒ MoE architecture homogeneity limitation discussed in Section 3.3: all four models share Mixture-of-Experts architecture, limiting generalizability to other architectural families.
- ☒ Prompt v2.2 limitations stated in Section 3.2: single prompt version (no prompt-by-model crossed design) and Chinese-language confound acknowledged.
- ☒ Single-prompt limitation acknowledged: model effects and model-by-prompt interaction effects cannot be disentangled.

13.4 E.4 Narrative Balance

- ☒ No Galileo gambit: the dissertation does not claim that low agreement *validates* the models or that disagreement is inherently *good*. The framing is explicitly that disagreement is “informative when properly decomposed” but “concerning for deployment” (Section 7.1).
- ☒ Negative results reported prominently: 7 of 30 model pairs fail to exceed random baseline; Fleiss’ kappa CI includes zero for penetration_overall; d3_type_d ICC CI includes zero.
- ☒ Limitations section (7.8) addresses: single prompt, MoE homogeneity, Chinese language, no human gold standard, author-created nodes, temporal snapshot, temperature=0.0 setting.

13.5 E.5 Numerical Traceability

- ☒ All numerical claims traced to source CSV/JSON files: kappa values to kappa_full.csv, bootstrap CIs to bootstrap_kappa.csv and bootstrap_fleiss.csv, ICC values to icc_numeric.csv, SDAF decomposition to table_6_1_sdaf.csv.
- ☒ 148 automated assertions pass via thesis_verify_claims.py (zero failures).
- ☒ All 29 figures render from source data without manual adjustment.

13.6 E.6 New Analyses in v3.0

- ☒ Cluster bootstrap kappa CI [0.064, 0.091] computed and reported (Section 5.10.1).
- ☒ Cross-validated AUC = 0.877 reported alongside in-sample AUC = 0.923 (Appendix J.4).
- ☒ Model-by-framework interaction analysis: D_bias ranges 4.3–23.6 pp across frameworks (Table C.4).
- ☒ Weighted kappa analysis expanded to all six model pairs for both ordinal fields (Table C.2, Appendix J.2).
- ☒ Intra-model test-retest analysis for GPT-5 mini (500 duplicate nodes, Appendix F.1).
- ☒ Weighted kappa full analysis across all 5 categorical fields (Appendix F.2).
- ☒ SDAF encoding sensitivity analysis (lexicographic vs. semantic ordering, Appendix F.3).
- ☒ Prompt sensitivity analysis: Gemini v2.1 vs. v2.2 (Appendix F.4).

13.7 E.7 Reproducibility

- ☒ All analysis scripts listed in Appendix E (main dissertation) with runtime estimates.
- ☒ Random seeds documented: bootstrap RNG seed = 42 (2,000 iterations); Monte Carlo seed = 42 (1,000 iterations).
- ☒ Computational environment specified: Python 3.12, pandas 3.0.1, scipy 1.17.1, scikit-learn 1.8.0.
- ☒ Complete data files enumerated with row/column counts (14 files, Appendix E.2 of main dissertation).

Chapter 14

Appendix F: Robustness Analyses

14.1 F.1 Intra-Model Test-Retest Consistency (GPT-5 mini)

GPT-5 mini batch overlap produced 500 nodes scanned twice under identical conditions (same prompt v2.2, temperature = 0.0). This natural experiment enables assessment of a single model’s self-consistency, providing an upper-bound reference for inter-model agreement.

Table F.1: Categorical Field Test-Retest Agreement

Field	Exact Match %	Cohen’s κ	Interpretation
change_status	99.2	−0.003	Near-perfect match; κ near zero due to marginal concentration
boundary_current_type	97.2	−0.012	Same pattern; 88% Type 2 dominance
penetration_overall	95.0	0.267	Fair agreement
summary_priority_flag	92.0	0.445	Moderate agreement
uncertainty_confidence	59.6	0.109	Slight — most unstable even within-model

Table F.2: Numerical Dimension Test-Retest ICC (Selected)

Dimension	ICC	Mean Abs Diff
d1_tacit_knowledge_dependency	0.791	0.41
d7_rule_driven_degree	0.789	0.24
d1_decision_replaceability	0.701	0.24
d8_data_intensity	0.566	0.20
d3_type_c	0.139	0.15
d7_exception_flexibility	0.142	0.50

The median intra-model ICC (≈ 0.45) far exceeds the inter-model median (0.137), confirming that cross-model disagreement reflects genuine model differences rather than agreement) reveal that even a single model's uncertainty calibration is inherently unstable.

14.2 F.2 Quadratic Weighted κ — All Five Categorical Fields

Quadratic-weighted κ assigns partial credit for ordinal near-misses (e.g., “changed” vs. “will change” penalized less than “changed” vs. “stable”).

Table F.3: Weighted vs. Unweighted κ by Field (6-Pair Means)

Field	Unweighted κ	Weighted κ_{wq}	Improvement
	Mean	Mean	
change_status	0.108	0.187	+73%
penetration_overall	0.086	0.153	+78%
summary_priority_flag	0.067	0.145	+116%
uncertainty_confidence	0.074	0.092	+24%
boundary_current_type	0.053	0.069	+30%

The highest single-pair weighted κ is DeepSeek–Qwen3 on change_status ($\kappa_{wq} = 0.452$, “moderate”). Weighted κ values are universally higher than unweighted, confirming that disagreements tend to involve adjacent ordinal categories rather than random scatter. However, even the best-case weighted mean (0.187) remains in the “slight” range, so the weighting correction does not fundamentally alter the reliability assessment.

14.3 F.3 SDAF Encoding Sensitivity Analysis

SDAF encodes categorical fields as numerical values for mixed-effects variance decomposition. Three encoding schemes are compared:

- **A (Lexicographic):** Python `sorted()` default (e.g., 已变 =0, 将变 =1, 稳定 =2)
- **B (Semantic):** Domain-ordered (e.g., 稳定 =0, 将变 =1, 已变 =2)
- **C (Reversed):** Reverse of semantic order

Table F.4: SDAF Decomposition Under Alternative Encodings

Field	Scheme	D_{bias}	$D_{\text{ambiguity}}$	D_{residual}
change_status	A (lex)	11.9%	11.0%	77.2%
change_status	B (sem)	20.2%	21.6%	58.3%
penetration_overall	A (lex)	16.2%	5.5%	78.3%
penetration_overall	B (sem)	29.1%	16.3%	54.7%
boundary_current_type	A/B/C	\approx \$4.4%	\approx \$7.3%	\approx \$88.4%
summary_priority_flag	A/B/C	42.2%	14.1%	43.6%

Schemes B and C produce *identical* results, confirming that variance decomposition is invariant to linear scale direction (a mathematical property of ANOVA). The lexicographic vs. semantic ordering affects magnitude (up to 13 pp difference for change_status D_{bias}) but does not alter the qualitative finding: D_{residual} remains the dominant component under all encodings. Fields with natural ordinal structure (boundary_current_type, summary_priority_flag) are encoding-insensitive; fields with culturally-loaded category labels (change_status, penetration_overall) show greater sensitivity.

14.4 F.4 Prompt Sensitivity Analysis: Gemini v2.1 vs. v2.2

The same model (Gemini 2.5 Flash) scanned all 2,325 nodes under both prompt v2.1 and v2.2, enabling direct measurement of prompt sensitivity with model identity held constant.

Table F.5: Cross-Prompt Agreement

Field	Exact Match	κ	Largest Distribution Shift
change_status	78.6%	0.483	“changed” −3.4 pp
penetration_overall	80.1%	0.441	minimal shifts
boundary_current_type	72.7%	0.194	Type 2: +15.7 pp; Type 3: −13.6 pp

Field	Exact Match	κ	Largest Distribution Shift
uncertainty_confidence	62.8%	0.186	“low” 0%→23.3% (+23.3 pp)
summary_priority_flag	62.4%	0.216	“high priority” −32.8 pp

The v2.2 prompt revisions (particularly Rule 9: status constraints and Rule 10: confidence calibration) produced dramatic distribution shifts: high-priority flagging dropped from 86% to 53%, and confidence “low” emerged from 0% to 23%. This confirms that prompt wording is a first-order determinant of LLM assessment behavior. The cross-prompt κ values (0.19–0.48) are *higher* than cross-model κ values (mean 0.078), suggesting that model identity contributes more variance than prompt version—but prompt effects are far from negligible.

参考文献

- Alizadeh, M. et al. (2025). Large-scale comparison of LLMs and human annotators in latent content analysis. *arXiv preprint*.
- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.
- AXELOS (2019). *ITIL Foundation: ITIL 4 Edition*. The Stationery Office.
- Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21.
- Berg, H. et al. (2022). Prompt engineering for annotation tasks. *arXiv preprint*.
- Brennan, R. L. (2001). *Generalizability Theory*. Springer.
- Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46:423–429.
- Chandra, S. et al. (2025). Claude approaches perfect reliability in structured rubric-based writing assessment. *arXiv preprint*.
- Chehbouni, A. et al. (2025). LLM annotation reliability across domains. *arXiv preprint*.
- Chen, X. et al. (2025). Inter-model agreement in thematic analysis of qualitative research data. *arXiv preprint*.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6:284–290.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49:997–1003.
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. Wiley.
- Deldjoo, Y. et al. (2025). Agreeableness bias in LLM judge panels. *arXiv preprint*.

- Dou, S. et al. (2025). LLM evaluation reliability. *arXiv preprint*.
- Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43:543–549.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. In *Proceedings of the National Academy of Sciences*.
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61:29–48.
- Gwet, K. L. (2014). Handbook of Inter-Rater Reliability* (4th ed. *Handbook of Inter-Rater Reliability*.
- He, K., Zhou, W., Zhang, Y., and Sun, X. (2025). Large language models for automated annotation: A reliability study across domains. *arXiv preprint arXiv:2401.13298*.
- Huang, L. et al. (2025). LLM assessment reliability. *arXiv preprint*.
- Kadavath, S. et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Khalifa, M. et al. (2025). LLM annotation quality. *arXiv preprint*.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- Li, Y. et al. (2025). A comprehensive taxonomy of LLM evaluation biases. *arXiv preprint*.
- Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, pages 677–694.
- Plank, B., Hovy, D., and Sogaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of ACL 2014* (pp. In *Proceedings of ACL 2014*, pages 507–511.
- Regier, D. A. et al. (2013). DSM-5 field trials in the United States and Canada. *American Journal of Psychiatry*, 170:59–70.
- Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. Wiley.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Uma, A. et al. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

- Warrens, M. J. (2010). A formal proof of a paradox associated with Cohen’s kappa. *Journal of Classification*, 27:322–332.
- Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p -values: Context, process, and purpose. *The American Statistician*, 70:129–133.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.
- Xiong, M. et al. (2024). Can LLMs express their uncertainty? *arXiv preprint arXiv:2401.14640*.
- Zhao, Z. et al. (2021). Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.
- Zheng, L. et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.