
Anatomy of a $2-3\sigma$ Hint

An Independent Stress Test of the DESI DR2 Evolving-Dark-Energy Preference

Bobby Morong (independent researcher), with AI-assisted implementation · July 2026

Canonical edition · July 2026

SCIENTIFIC STATUS

A calibrated, LRG2-sensitive $2-3\sigma$ hint within published compressed likelihoods. Not a discovery.

Includes selection-calibrated influence, direct Λ CDM-versus-CPL time variation, held-out LRG2 prediction, full-Boltzmann verification, and explicit frontier evidence gates.

Numerical results frozen at analysis commit 28894e7. Canonical source: [whitepaper.md](#).

Abstract

We independently reimplement the compressed likelihoods behind the DESI DR2 preference for evolving dark energy (w_0w_a CDM over Λ CDM) and subject the result to influence diagnostics, calibration stress tests, parameterization swaps, parametric bootstraps, a comparator-measurement swap test, full-Boltzmann (CAMB) verification, and posterior sampling under the collaboration's own priors. We reproduce the published $\Delta\chi^2_{\text{MAP}}$ ladder ($-4.7/-4.9/-8.0$ published vs $-4.7/-5.1/-8.5$ here), marginalized posteriors ($w_0 = -0.45 +0.24/-0.23$, $w_a = -1.66 +0.65/-0.71$ vs published -0.42 ± 0.21 , -1.75 ± 0.58), the w_0 – w_a posterior correlation (-0.978 vs -0.975), and ΔDIC (-4.96 vs -4.4). We then find the preference is sharply localized. Removing the single LRG2 ($z = 0.706$) BAO measurement collapses the compressed-CMB preference from 2.4σ to 1.6σ (full CAMB: $2.41\sigma \rightarrow 1.53\sigma$). Substituting the earlier eBOSS DR16 measurement at the same redshift — a lower-precision, partially sample- and footprint-correlated comparator, not an independent experiment — yields a preference identical to having no measurement there. The $z \approx 0.7$ anomaly changed observable between DESI's releases — transverse (D_M) in DR1, radial (D_H) in DR2. A selection-calibrated test repeats all seven tracer deletions inside each of 5,000 Λ CDM mocks. The observed maximum loss of preference is $M = 4.125$ in $\Delta\chi^2$, at LRG2; 78/5,000 null mocks have an equal-or-larger selected maximum, $p = 0.0156$ [0.0124, 0.0194]. However, among the 70 null mocks whose global w_0w_a preference is at least as strong as observed, 34 have an equal-or-larger maximum, $p_{\text{cond}} = 0.486$ [0.364, 0.608]. LRG2 is therefore descriptively load-bearing, but the concentration is not exceptional conditional on a global fluctuation of this strength. A held-out Λ CDM fit predicts LRG2 at joint $p = 0.0408$ (2.05σ); the discrepancy is mainly in the isotropic distance D_V/r_d ($p = 0.0204$), while the Alcock–Paczynski ratio is ordinary ($p = 0.374$). On the supernova side, a single free magnitude offset on the DES5Y external low- z sample fits at $+0.036$ mag and reduces the DESI+DES5Y preference from 2.2σ to 0.2σ (1.3σ with the CMB anchor); this is a degeneracy test, not a calibration diagnosis. Evolving families outperform constant w , but the direct question is calibrated separately: in DESI+cCMB, CPL w_0w_a CDM improves over w CDM by $T = 7.851$, with 30/5,000 constant- w mocks as large, $p = 0.0060$ [0.0041, 0.0086] (2.75σ). Removing LRG2 reduces this direct time-variation test to $T = 3.426$, $p = 0.0698$ [0.0629, 0.0772] (1.81σ). A separate global Λ CDM bootstrap gives $p = 0.0140$ [0.0109, 0.0177], consistent with Wilks. A pre-registered cross-probe consistency test returned an INTERMEDIATE verdict. Finally, a full-CMB variant including lensing (PR3-native likelihoods) gives $\Delta\chi^2 = -10.35$ (2.77σ) with LRG2 and -3.76 (1.43σ) without it. We conclude that the compressed likelihood contains a calibrated, CPL-specific hint of time variation, but its strength remains LRG2-sensitive and the selected localization is ordinary among equally strong Λ CDM fluctuations. A hint, precisely mapped; not a discovery.

1. Motivation and scope

The DESI DR2 BAO analysis (arXiv:2503.14738) reports a preference for a time-varying dark-energy equation of state reaching 3.1σ with full CMB data and 2.8 – 4.2σ with supernova compilations — the most significant challenge to Λ CDM in a generation, if it holds. Claims of this magnitude attract two failure modes: uncritical amplification and uncritical dismissal. This project attempts a third path: full independent reimplementation followed by systematic attempts to break the result, with every finding published regardless of direction.

Scope, stated precisely: this is an independent implementation of the *published compressed likelihoods* — DESI DR2 BAO summary values, Pantheon+ and DES-SN5YR distance moduli with full covariances, and the DR2 paper's own compressed early-CMB prior (its eqs. 35–36). It is not a raw-data reanalysis. In addition to the compressed branch, a full-CMB variant (Planck 2018 native low- ℓ TT/EE + plik-lite TTTEEE + lensing, via cobaya+CAMB) is included; DESI's own analysis uses PR4 CamSpec and PR4/ACT lensing, so this variant tests the lensing-inclusive structure of the result, not its exact published numbers. Four AI systems performed software, statistics, or manuscript checks during development; these are not independent scientific peer review. One re-executed the pipeline and matched all χ^2 values to four decimals. A clean-room human rerun remains the outstanding verification step.

2. Data and provenance

All inputs are public and primary-sourced under a fixed rule: no number enters the pipeline from memory. DESI DR2 BAO values (7 tracers, correlated D_M/D_H pairs plus BGS D_V) and the compressed CMB prior were machine-extracted from the arXiv HTML of the DR2 paper. Pantheon+ (1,701 light curves; 1,580 after standard cuts; full STAT+SYS covariance) from the official PantheonPlusSH0ES release. DES-SN5YR from the official repository's 2025 Dovekie-recalibrated release ($N = 1,820$; inverse covariance unpacked per the repository's own likelihood code; external low- z

sample identified by IDSURVEY). eBOSS DR16 LRG consensus values from the official SDSS files in the cobaya bao_data release. DESI DR1 values from arXiv:2404.03002. BBN and θ^* priors from DR2 eqs. 14 and 16; $w(z)$ priors ($w_0 \in U[-3,1]$, $w_a \in U[-3,2]$, $w_0 + w_a < 0$) verbatim from the paper.

3. Methods

Likelihoods are Gaussian in the published summary statistics (mirroring the official compressed treatment), with supernova magnitude offsets marginalized analytically and, where used, subsample calibration offsets marginalized jointly. Late-time distances use a validated fast integrator (relative error $< 3 \times 10^{-7}$, verified by external audit). Early-universe quantities (r_{drag} , θ^*) use, in the production pipeline, a calibrated surrogate — standard fitting formulas pinned to the paper's own r_d formula and CMB column, a labeling adopted after audit — and, decisively, full CAMB Boltzmann computations (RECFAST recombination, PPF dark energy) for verification. Two validation gates preceded all dark-energy fits and caught two real implementation bugs before any result existed. The Λ CDM-versus- $w_0 w_a$ statistic is calibrated with 5,000 Λ CDM mocks. The direct one-extra-parameter test of time variation compares w CDM to CPL $w_0 w_a$ CDM and is calibrated with 5,000 mocks at the observed best-fit constant- w null, both with and without LRG2. The official w_0 , w_a , and $w_0 + w_a$ prior domain is enforced in every fit.

4. Reproduction

Quantity	This work	Published
BAO-alone Λ CDM	$\Omega_m = 0.2973 \pm 0.0083$, $h \cdot r_d = 101.55 \pm 0.72$	0.2975 ± 0.0086 , 101.54 ± 0.73
Pantheon+-alone Λ CDM	$\Omega_m = 0.3326 \pm 0.0180$ (N=1580)	0.334 ± 0.018
$\Delta\chi^2$: DESI / +Pantheon+ / +cCMB	-4.7 / -5.1 / -8.5	-4.7 / -4.9 / -8.0
w_0 , w_a (DESI+cCMB, MAP)	-0.44, -1.68 ($H_0 = 63.6$)	-0.42, -1.75 (63.6)
w_0 , w_a (posterior 68%)	-0.45 +0.24/-0.23, -1.66 +0.65/-0.71	-0.42 $\pm$ 0.21, -1.75 $\pm$ 0.58
w_0 - w_a posterior correlation	-0.978	-0.975
Δ DIC (DESI+cCMB)	-4.96 (MAP plug-in; $p_D = 4.75/3.0$)	-4.4

An AI software audit re-ran this pipeline and confirmed all χ^2 values to four decimals, verified the minima against differential evolution, and confirmed covariance positive-definiteness. This is a software check, not independent scientific review.

5. Influence diagnostics: where the preference lives

Leave-one-tracer-out refits of DESI+cCMB show the preference is not distributed: dropping any of six tracers leaves 2.3 – 2.6σ (removing BGS or Ly α *strengthens* it), while dropping LRG2 ($z = 0.706$) collapses it to 1.6σ . The χ^2 decomposition attributes the $w_0 w_a$ improvement to LRG2 (-4.1) and the CMB geometric anchor (-3.9), with LRG1 contributing -1.2 , all others ≈ 0 , and Ly α opposing at $+1.0$. Within LRG2, the radial measurement carries more model leverage: retaining only D_H/r_d gives $\Delta\chi^2 = -6.5$, only D_M/r_d gives -4.9 , neither -4.3 . The selection-calibrated statistic $I_j = \Delta\chi^2_{-j} - \Delta\chi^2_{\text{full}}$ and $M = \max_j I_j$ gives $M_{\text{obs}} = 4.125$ at LRG2. Repeating the full fit and all seven deletion fits inside each of 5,000 Λ CDM mocks yields an unconditional selected tail of 78/5,000, $p = 0.0156$ [0.0124, 0.0194]. Conditioning on a global preference at least as strong as observed leaves 70 mocks, of which 34 meet or exceed M_{obs} : $p_{\text{cond}} = 0.486$ [0.364, 0.608]. Thus LRG2's leverage is a valid description of where the observed fit changes, but it is not an additional anomaly beyond the rarity of the global fit.

A targeted held-out check fits Λ CDM to DESI+cCMB without LRG2, propagates the fitted-parameter covariance, and then predicts the excluded (D_M/r_d , D_H/r_d) vector. The prediction is (17.530, 20.066), compared with (17.351, 19.455) observed. Repeating the full exclusion, fit, prediction, and score in 5,000 fitted- Λ CDM mocks gives a joint tail of 204/5,000, $p = 0.0408$ [0.0355, 0.0467] (2.05σ). Its decomposition changes the earlier interpretation: the isotropic distance D_V/r_d is low ($p = 0.0204$), but the Alcock–Paczynski ratio D_M/D_H is unexceptional ($p = 0.374$). LRG2's radial coordinate

carries more leverage for the w_0w_a model comparison, but the held-out residual is not an anomalous anisotropy. Because LRG2 was selected after inspection, these targeted p-values are diagnostic and do not replace the selected-null calibration above. Shifting LRG2 halfway toward Λ CDM lowers the global preference below 2σ ; inflating only its errors degrades the signal smoothly, indicating statistical leverage rather than numerical pathology. On the supernova side, the significance of DESI+Pantheon+ falls monotonically from 1.8σ to 1.3σ as low- z cuts rise from $z > 0.01$ to $z > 0.1$, with w_0 pinned near -0.89 throughout.

6. The $z \approx 0.7$ measurement across experiments

Three analyses have measured BAO at this redshift. Against the Λ CDM best-fit prediction: eBOSS DR16 ($z = 0.698$; partially correlated with DESI in sample and footprint) deviates jointly by 0.83σ , with D_M high if anything; DESI DR1 by 1.40σ , driven by low D_M ; DESI DR2 by 1.51σ , driven by low D_H . The anomaly changed observable between DESI's own releases. The swap test quantifies the consequence: with DR2's LRG2 the preference is 2.4σ ; with DR1's, 1.9σ ; with eBOSS's, 1.6σ — numerically identical to having no $z \approx 0.7$ measurement at all. Balanced reading: DR2's measurement is 1.7 – $1.8\times$ more precise than either comparator and may be resolving what noisier data could not; DESI's blinded internal-consistency checks passed; eBOSS could not strongly confirm a real signal at its precision, and the footprints partially overlap. But a discovery narrative must carry both facts: the signature is unstable across releases, and the earlier partially overlapping measurement at this redshift is Λ CDM-consistent (the swap is a comparator test, not an independent non-confirmation). Reproducing LRG2 from its official likelihood surface is the single most decisive open test.

7. Supernova calibration

On the Dovekie-recalibrated DES5Y release, DESI+DES5Y (no CMB) prefers w_0w_a at 2.2σ ($\Delta\chi^2 = -7.3$) — against -13.6 (3.3σ) published with the 2024 release. The updated Dovekie release changed cross-calibration, SALT training, scatter modeling, bias simulations, and sample treatment together; the smaller preference coincides with the new release, but attribution among those changes is not established (the same machinery reproduces Pantheon+ at -5.1 vs -4.9). As a sensitivity test (not a diagnosis — a free offset can absorb calibration, selection, population evolution, survey differences, or genuine cosmological curvature alike), freeing one magnitude offset on the 197 external low- z supernovae returns $D = +0.037$ mag under Λ CDM — matching Efstathiou's ~ 0.04 mag estimate from Pantheon+/DES5Y cross-matching — and the preference collapses to 0.2σ , with $w_0 = -0.99$. With the CMB anchor attached, the intercept still reduces 2.9σ to 1.3σ ($D = +0.036$). Counter-arguments are part of the record: the DES team reproduces the offset but attributes much of it to improved intrinsic-scatter and host-galaxy modeling (reverting those assumptions moves their result $3.9\sigma \rightarrow 3.3\sigma$, not to zero); the intercept is a post-hoc test motivated by a critic and may double-count covariance systematics; and per Cortês & Liddle, the overlapping supernova compilations must never be averaged.

8. Robustness of the statistical framework

The preference is not unique to CPL: JBP and logarithmic $w(z)$ families find similar improvements. That parameterization swap, however, is not itself a calibrated test that w changes with time. We therefore test the nested CPL question directly, comparing constant- w CDM to w_0w_a CDM with only w_a added. For DESI+cCMB, $T = \chi^2_{\min}(w\text{CDM}) - \chi^2_{\min}(w_0w_a\text{CDM}) = 7.8506$. In 5,000 mocks generated at the observed best-fit constant- w model, 30 are at least as large: $p = 0.0060$ [0.0041, 0.0086], or 2.75σ ; the one-degree-of-freedom Wilks value is 2.80σ . No mock fit failed or reached the prior boundary. After removing LRG2, T falls to 3.4263, with 349/5,000 exceedances: $p = 0.0698$ [0.0629, 0.0772], or 1.81σ . Thus the compressed data contain a direct, calibrated, CPL-specific preference for time variation, but most of its strength is not retained without LRG2. This does not establish arbitrary time dependence or a physical dark-energy field.

Separately, the 5,000-mock Λ CDM-versus- w_0w_a global calibration yields 70 exceedances of the unrounded observed $\Delta\chi^2 = -8.4575$: $p = 0.0140$ [0.0109, 0.0177], versus the Wilks expectation 0.0146. The selection-calibrated maximum-influence result is reported in §5; its conditional result shows that localization should not be counted as independent evidence against Λ CDM. Full CAMB computation (a separate execution environment, not scientific independence) confirms the baseline and LRG2-deletion fits to ≈ 0.2 in $\Delta\chi^2$. Posterior sampling under the official priors shows no pileup at the $w_0+w_a < 0$ boundary; the naive mean-plug-in DIC convention (-8.99) is documented as an artifact of the curved w_0-w_a ridge, with the standard MAP plug-in giving -4.96 .

8a. Pre-registered cross-probe consistency test (H-X)

With hypotheses and kill conditions registered before execution (HYPOTHESIS-cross-probe.md), per-probe w_0w_a posteriors were sampled for P1 = DESI BAO+cCMB, P2 = Pantheon+ +cCMB+BBN, P3 = DES5Y+cCMB+BBN. Pairwise 2D parameter-shift tensions: P1–P2 = 1.44σ , P1–P3 = 1.35σ , P2–P3 = 0.24σ . Phantom-crossing redshifts: P1 $z \times = 0.487$ [0.418–0.553]; P2 0.292 [0.187–0.502]; P3 0.296 [0.244–0.403]. Verdict against the registered conditions: INTERMEDIATE — neither the consistency ($<1\sigma$ all pairs) nor inconsistency ($>2\sigma$ any pair) threshold met; reported without promotion. The registered follow-up is complete: P1 without LRG2 gives $z \times = 0.478$ [0.384–0.564], so LRG2 controls much of the preference's strength but not the surviving BAO posterior's crossing location. Caveats: Gaussian shift metric on curved posteriors; surrogate likelihood (CAMB-validated); shared-anchor correlation.

8b. The full-CMB lensing variant

Using cobaya with CAMB and the klik-free native Planck 2018 likelihoods (low- ℓ TT, low- ℓ EE, plik-lite TTTEEE, lensing) plus the DESI BAO likelihood, MAP minimization (BOBYQA, dual converged runs) under DESI's official $w(z)$ priors gives, after removing the uniform-prior volume constant: with LRG2, $\Delta\chi^2 = -10.35$ (2.77σ); without LRG2, $\Delta\chi^2 = -3.76$ (1.43σ). The with-LRG2 value sits below the published PR4-based -12.5 (3.1σ) by an amount consistent with the PR3-vs-PR4 likelihood difference (pre-registered caveat). The central finding stands at the full-likelihood level: **CMB lensing does not substitute for LRG2** — the preference collapses from $\sim 2.8\sigma$ to $\sim 1.4\sigma$ on removal of one measurement. Caveats: PR3-native likelihoods; MAP not posterior comparison; single pipeline pending clean-room rerun; the selected-null calibration and its conditional limit are in §5.

9. What this establishes — and what it does not

Established: the published compressed likelihoods genuinely contain the reported preference; it survives independent reimplementations, optimizer attack, full Boltzmann physics, parameterization changes, and bootstrap calibration. Within CPL, a direct constant- w null test gives a calibrated 2.75σ preference for time variation, falling to 1.81σ without LRG2. The held-out LRG2 vector is a modest 2.05σ discrepancy, driven more by isotropic distance scale than by its shape ratio. Among Λ CDM mocks with a global preference this strong, a maximum deletion influence at least as large as observed occurs about half the time. Not established: that dark energy evolves; that LRG2 is wrong; that localization or the targeted held-out p-value is independent evidence; that the supernova calibration explanation is correct; or that the PR3-native full-CMB variant reproduces DESI's PR4/ACT stack. The honest status is **a reproducible, LRG2-sensitive 2–3 σ hint — not proof.**

10. Frontier-standard boundary and development roadmap

A frontier-roadmap review was compared against the frozen analysis rather than treated as new evidence. Several recommendations are adopted as the next evidentiary gates: native-likelihood reconstruction of LRG2; survey splits and nuisance templates for fiber assignment, reconstruction, redshift failures, observing conditions, covariance, and window functions; flexible or nonparametric expansion reconstructions calibrated inside null mocks; prior-sensitive Bayesian evidence or out-of-sample predictive scores; growth consistency through RSD, $f\sigma_8$, weak lensing, and CMB lensing; and a hierarchical, survey-aware supernova calibration model.

The same review also proposed interpretations that exceed this project's evidence. This compressed-likelihood analysis does not identify fiber assignment as the cause of LRG2, prove that the supernova seam is a calibration error, validate a Gaussian-process crossing, support Early Dark Energy, or favor interacting/decaying dark matter or modified gravity. Those ideas remain hypotheses until they improve predictions outside the data that generated the anomaly and survive appropriate null calibration and model-complexity penalties. The complete comparison is recorded in FRONTIER_REVIEW.md.

The frontier claim ladder is therefore explicit:

Level	Required evidence	Current status
Reproducible likelihood feature	independent implementation and stable optimization	Met for the published compressed inputs
Calibrated anomaly	null simulations and selection-aware statistics	Met for the global, maximum-influence, direct-time-variation, and held-out tests
Measurement diagnosis	native likelihood, survey splits, nuisance-template and covariance audit	Open
Cross-probe physical consistency	expansion, growth, and lensing predicted by one model	Open
Model preference	prior-robust Bayesian or predictive advantage over flexible alternatives	Open
Discovery	independent data and human replication at discovery-grade significance	Not met

11. Falsifiable expectations

If the signal is physics: DESI's final likelihood (2027) retains the $z \approx 0.7$ distance-scale departure at higher precision; Rubin-calibrated supernovae reproduce the low- z crossing with no magnitude offset required; growth and lensing co-vary as dynamical dark energy predicts. If it is systematics: the LRG2 distance residual regresses; the supernova preference tracks calibration choices across compilations; residuals decorrelate across probes. We pre-commit to updating this document against those outcomes.

Reproduction

Everything runs from the released directory. The selection calibration is `python3 scripts/max_influence_mocks.py --mocks 5000 --workers 8 --fresh`. The direct test is `python3 scripts/time_variation_mocks.py --mocks 5000 --workers 8 --fresh`; repeat with `--drop-index 2 --` output `results/time_variation_no_lrg2_mocks.json` for the no-LRG2 branch. The held-out test is `python3 scripts/lrg2_posterior_predictive.py --mocks 5000 --workers 8 --fresh`. Results ship as JSON in `results/`; canonical values, hashes, and completion states are in `RESULTS_MANIFEST.json`. The blind human-verifier protocol and combined runners are in `cleanroom/`.

Acknowledgments and provenance of assistance

Implementation, literature triage, and drafting were AI-assisted (multiple systems, adversarially cross-checked); four AI software/statistics/manuscript audits are incorporated with attribution in the README, but these are not independent scientific replication. A human clean-room rerun remains pending. Errors are the author's. Data credits: DESI, SDSS/eBOSS, Pantheon+/SH0ES, DES, Planck, and the cobaya project's data releases.

Primary sources

1. DESI Collaboration, “DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints,” arXiv:2503.14738.
2. DESI Collaboration, “DESI 2024 III: Baryon Acoustic Oscillations from Galaxies and Quasars,” arXiv:2404.03000, and “DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations,” arXiv:2404.03002.
3. Brout et al., “The Pantheon+ Analysis: Cosmological Constraints,” arXiv:2202.04077.
4. DES Collaboration, “The Dark Energy Survey: Cosmology Results With ~1500 New High-redshift Type Ia Supernovae Using the Full 5-year Dataset,” arXiv:2401.02929.

5. Efstathiou, “Evolving Dark Energy or Supernovae Systematics?”, arXiv:2408.07175.
6. Vincenzi et al., “Comparing the DES-SN5YR and Pantheon+ SN cosmology analyses,” arXiv:2501.06664.
7. Cortès and Liddle, “Interpreting DESI’s evidence for evolving dark energy,” arXiv:2504.15336.

Appendix A. Complete test ledger

#	Test	Result	Verdict
1	Λ CDM reproduction (BAO-alone, SN-alone)	Ω_m to ± 0.0002 / ± 0.0014 of published	PASS
2	$w_0 w_a \Delta\chi^2$ ladder reproduction	$-4.7/-5.1/-8.5$ vs published $-4.7/-4.9/-8.0$	PASS
3	Best-fit parameter reproduction	w_0, w_a, Ω_m, H_0 to ≤ 0.01	PASS
4	Posterior reproduction (emcee, official priors)	intervals match; correlation -0.978 vs -0.975 ; ΔDIC -4.96 vs -4.4	PASS
5	External re-execution (independent AI)	all χ^2 to 4 decimals; minima verified vs differential evolution	PASS (software check)
6	Leave-one-tracer-out (7 BAO tracers)	only LRG2 matters: $2.4\sigma \rightarrow 1.6\sigma$; others $2.3-2.6\sigma$	LOCALIZED
7	χ^2 pull decomposition	LRG2 -4.1 , CMB anchor -3.9 , LRG1 -1.2 , $\text{Ly}\alpha$ $+1.0$ (opposes)	LOCALIZED
8	LRG2 D_M/D_H split (external contribution, verified)	radial D_H carries the leverage (-6.5 vs -4.9)	LOCALIZED
9	eBOSS swap (partially correlated comparator)	eBOSS in the $z \approx 0.7$ slot $\equiv$ no measurement at all (1.6σ)	NOT CORROBORATED
10	DR1 vs DR2 $z \approx 0.7$ comparison	anomaly changed observable: transverse (DR1) $\rightarrow$ radial (DR2); eBOSS sees neither	UNSTABLE SIGNATURE
11	Efstathiou intercept test (DES5Y)	fitted offset $+0.036-0.037$ mag (matches his ~ 0.04); $2.2\sigma \rightarrow 0.2\sigma$ no-CMB, $2.9\sigma \rightarrow 1.3\sigma$ with cCMB	SN LEG ABSORBABLE (sensitivity test; DES rebuttal stands)
12	SN low- z cut sweeps (both compilations)	significance decays monotonically as low- z SNe removed; crossing is a low- z phenomenon	LOCALIZED (SN side)
13	Dovekie-release comparison	DES5Y contribution $-13.6 \rightarrow -7.3$ across releases (multi-element change; attribution open)	RELEASE-DEPENDENT
14	$w(z)$ parameterization swap (CPL/JBP/LOG/wCDM)	evolving families all $2.1-2.9\sigma$; constant- w $0.8-1.2\sigma$	EVOLVING FLEXIBILITY REWARDED; DIRECT TEST IN #23
15	Parametric bootstrap, 5,000 Λ CDM mocks (unrounded reconciled runner)	$P(\text{false pos.}) = 1.40\%$ [$1.09-1.77\%$] vs Wilks 1.46%	STATISTICS CALIBRATED
16	Anchor-uncertainty propagation	surrogate pins significance to $2.4\sigma \pm 0.5\sigma$	BOUNDED, then superseded by #17
17	Full CAMB Boltzmann verification	baseline 2.41σ vs surrogate 2.45σ ; no-LRG2 1.53σ vs 1.58σ	SURROGATE VINDICATED
18	Full-CMB + lensing variant (PR3-native, cobaya)	2.77σ with LRG2 (pub. PR4: 3.1σ); 1.43σ without LRG2	LENSING DOES NOT RESCUE
19	H-X pre-registered cross-probe consistency	BAO-SN shifts $1.44\sigma/1.35\sigma$; SN-SN 0.24σ ; $z\times$: BAO 0.49 ± 0.07 , SNe ≈ 0.30	INTERMEDIATE (per kill conditions)
20	H-X registered follow-up: BAO $z\times$ without LRG2	$z\times = 0.478$ [$0.384-0.564$] — unchanged	CROSSING LOCATION IS DISTRIBUTED, not an LRG2 artifact
21	H-Y registered: SN crossing vs intercept degeneracy	with intercept marginalized: $w_a = +0.50$ [$-0.33, +1.03$] (includes 0, flips sign); $w_0 \rightarrow -1.17$	KILL CONDITION MET: SN crossing degenerate with calibration seam
22	Selection-calibrated maximum influence, 5,000 Λ CDM mocks	$M_{\text{obs}} = 4.125$ at LRG2; unconditional 78/5,000, $p=0.0156$; conditional on equally strong global preference 34/70, $p=0.486$	LOCALIZED, BUT NOT EXCEPTIONALLY SO CONDITIONAL ON SIGNAL STRENGTH
23	Direct time-variation test: wCDM vs CPL, 5,000 constant- w mocks	$T=7.851$; 30/5,000; $p=0.0060$ [$0.0041-0.0086$], 2.75σ	CPL TIME VARIATION PREFERRED

#	Test	Result	Verdict
24	Direct time-variation test without LRG2	$T=3.426$; 349/5,000; $p=0.0698$ [0.0629–0.0772], 1.81σ	MOST STRENGTH IS LRG2-SENSITIVE
25	Held-out LRG2 Λ CDM prediction, 5,000 mocks	joint $p=0.0408$; D_V $p=0.0204$; D_M/D_H $p=0.374$	MODEST SCALE DISCREPANCY; SHAPE ORDINARY; TARGETED

Appendix B. Availability

Code, data-download instructions, results, chains, and the pre-registered hypothesis chain: <https://github.com/GrobeStreet/de-stress-lab> (v1.0, MIT). Archived with DOI via Zenodo.