

评估指标汇总

碳排放因子智能检索 MCP 服务器（Carbon Factor Matcher） • 零碳逻辑

一、大模型参与候选选择的消融实验

设计：4 种嵌入模型 × 2 种配置（纯向量检索，以及向量召回前 10 候选后由大模型语义重排，重排模型为 deepseek-chat，temperature = 0.1）；词汇失配基准 50 用例（查询用功能描述而非材料名）；95% bootstrap 置信区间（2,000 次重采样）。

表 1: 消融实验结果（50 用例）

模型	配置	Precision@1	MRR	NDCG@5	Recall@5
BERT (768d)	Embedding Only	24%	0.320	0.346	46%
BERT (768d)	LLM-Enhanced	40%	0.452	0.460	50%
Sentence-BERT (384d)	Embedding Only	32%	0.378	0.396	48%
Sentence-BERT (384d)	LLM-Enhanced	46%	0.476	0.484	52%
gte-large (1024d)	Embedding Only	38%	0.431	0.438	50%
gte-large (1024d)	LLM-Enhanced	50%	0.524	0.529	56%
text2vec-base (768d)	Embedding Only	10%	0.127	0.139	18%
text2vec-base (768d)	LLM-Enhanced	16%	0.167	0.169	18%

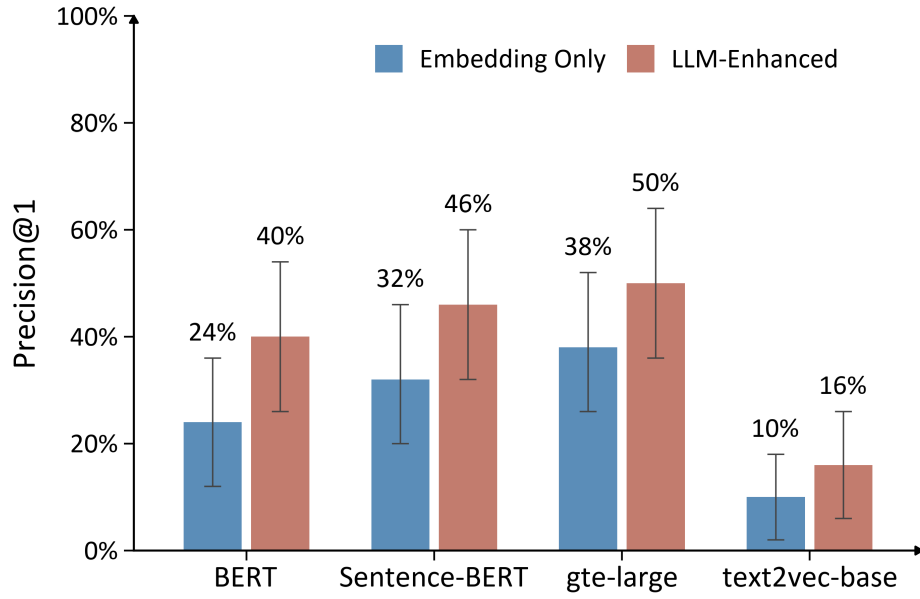


图 1: Precision@1 对比（误差棒为 95% bootstrap 置信区间）

结论：大模型语义复核对 4 个模型一致带来 +6 ~ +16 个百分点的 Precision@1 提升，且不以召回率下降为代价。据此，系统采用“服务端返回排序候选、由调用方智能体完成最终选择”的架构。

二、清单处理端到端评估（A/B 对照）

在本地库（ELCD 与 ecoinvent，共 7,742 条）上开展；基线为不做清单处理，处理组为去重加缺失替代。

表 2: 清单处理 A/B 实验结果

实验	设置	结果
去重	20 条物料、5 组真重复（复数/大小写/同义表述）	Precision = 1.000、Recall = 1.000（基线 Recall = 0）
缺失替代	8 种聚合物，无嵌入保守模式（类别回退）	MAE 改善 +8.4%；千单位排放偏差改善 +77.1%（19,143.5 → 4,381.7 kgCO ₂ e）；替代均标注“高”敏感性
端到端排放偏差	15 条真实清单（含 2 组重复、1 个缺失因子），对照 ELCD 真实因子值	11.7% → 0.0% （去重 2 条、替代 3 条；0.0% 表示与按真实因子直接计算完全一致，即处理环节未引入额外误差）

三、完整管线评估（混合检索 + 质量排序）

标准基准 17 例（含对抗性用例）实测：Recall@1 = 52.9%，Recall@5 = 64.7%，MRR = 0.665。

错配代价分析：错配案例平均排放影响达 595.6 kgCO₂e/千单位（正确匹配为 0），最严重个案（国别电网因子错配）达 1,521 kgCO₂e/千单位——印证了“返回带质量评级与出处的候选列表、由智能体复核选择”这一架构的必要性。

四、技术指标对照

表 3: 预期指标与实测对照

指标	目标	实测
词汇失配 Precision@1（纯向量）	≥ 35%	38%（gte-large）
词汇失配 Precision@1（加大模型复核）	≥ 45%	50% （gte-large）
大模型复核 P@1 提升（4模型）	一致为正	+6 ~ +16 pp
数据覆盖	≥ 5,000 条、多库	8,404 条 / 4 库
清单去重 Precision / Recall	≥ 0.95	1.000 / 1.000
清单端到端排放偏差	≤ 5%	11.7% → 0.0%
标准集 Recall@5 / MRR	——	64.7% / 0.665
单次检索响应	秒级	检索毫秒级；含大模型复核 1.4–2.2 s