智能体检索中的关键词查询：稀疏与通用密集检索的比较分析

颜京平*，张恒然*，郭嘉丰，毕可平†

智能算法安全全国重点实验室；中国科学院计算技术研究所；中国科学院大学

研究动机

智能体检索会反复调用检索器处理大模型生成的中间查询。这类查询通常较短、以实体为中心，并呈现明显的关键词特征。

核心问题：**何时应使用轻量稀疏检索？何时通用密集模型的语义收益足以抵消语料向量化、GPU 显存与查询编码开销？**

研究问题

- 在关键词查询和描述类查询下，两类模型的相对优势是否一致？
- 这些优势主要体现在靠前排序质量，还是体现在候选集合召回能力？
- 当查询来自真实智能体轨迹时，轻量稀疏模型是否仍能作为高性价比的底层检索器？

实验方法

■ **模型选择 (Retrievers)：对比 3 类代表性模型**
- **稀疏检索 (Sparse)**：BM25, QLD, SDM
- **BERT 密集检索 (PLM-based)**：BGE-M3
- **LLM 密集检索 (LLM-based)**：BOOM-4B-v1, Qwen3-Embedding-4B/8B, F2LLM-v2-8B

■ **数据集与查询形态 (Datasets & Query Types)**：
- **真实智能体轨迹**：LRAT
- **多源评测集**：WebAP (噪声网页), PsgRobust/Robust04 (规整新闻), NQ (百科问答)
- **查询变体**：对比关键词查询 (Key) 与 描述型查询 (Desc)

■ **评测维度 (Evaluation Dimensions)**：
- **检索效果**：Recall@100, P@10, NDCG@10, MRR@10
- **工程开销**：离线向量化耗时、在线查询编码/检索延迟、最大 GPU 显存/内存占用

实验一：关键词查询下的候选召回能力（Recall@100）

检索模型	WebAP		PsgRobust		NQ		Robust04	
	Key	Desc	Key	Desc	Key	Desc	Key	Desc
BM25	25.1	22.1	79.5	49.7	83.8	75.1	41.4	37.5
QLD	23.2	22.0	80.6	57.4	80.9	73.4	41.1	39.4
SDM	24.8	21.6	**81.9**	51.9	84.5	76.4	42.6	38.9
BGE-M3	29.3	36.4	49.2	45.8	94.6	94.0	33.1	34.5
BOOM-4B-v1	38.6	44.1	64.5	56.0	96.9	97.1	43.0	45.0
Qwen3-Embedding-4B	39.4	43.6	67.0	61.6	97.7	97.4	43.5	47.0
Qwen3-Embedding-8B	39.6	43.4	68.2	**63.3**	97.5	97.5	**44.7**	**47.2**
F2LLM-v2-8B	**40.8**	**45.8**	63.2	60.0	**97.8**	**98.2**	41.4	44.3

实验一：查询形式如何改变前排排序质量（NDCG@10）

检索模型	WebAP		PsgRobust		NQ		Robust04	
	Key	Desc	Key	Desc	Key	Desc	Key	Desc
BM25	15.4	14.2	40.5	29.5	39.6	30.6	44.3	40.3
QLD	12.6	12.4	38.8	31.4	35.1	26.5	43.0	40.4
SDM	14.6	14.1	40.7	30.3	40.5	31.9	45.3	42.0
BGE-M3	19.0	28.3	27.7	27.7	57.4	60.6	41.9	45.6
BOOM-4B-v1	29.9	36.9	39.7	35.1	64.0	65.4	52.3	57.4
Qwen3-Embedding-4B	27.7	32.7	41.9	38.4	63.4	63.6	53.4	62.7
Qwen3-Embedding-8B	27.5	33.7	**45.2**	**40.1**	64.7	64.8	**55.2**	**62.9**
F2LLM-v2-8B	**33.1**	**40.4**	41.2	38.5	**67.7**	**69.3**	51.6	59.2

实验一：关键发现

- 在关键词查询和描述类查询下，**两类模型的相对优势不一致**：稀疏模型通常更偏好短查询和关键词查询；当查询变成长描述时，密集检索模型通常更能利用描述类查询中的语义结构。
- 在规整新闻语料（比如PsgRobust和Robust04 ）中，SDM 等稀疏模型在候选**召回能力上表现很强**，说明当查询是精确关键词且语料质量较高时，词项匹配仍能以较低成本提供可靠候选池。
- 密集模型的优势更多体现在需要**语义泛化或头部排序质量**的场景。在 WebAP 这类真实网页语料上，文本噪声、排版内容和非规范表达会导致稀疏模型召回仅包含关键词但语义无关的结果。

实验二：真实智能体轨迹上的效果验证

表 6 模型在 LRAT 数据集（关键词查询）上的性能对比（数值为百分比，保留一位小数）[1,2]

模型类型	检索模型	LRAT (关键词查询)			
		Recall@100	P@10	NDCG@10	MRR@10
稀疏检索模型	BM25	93.7[2]	8.1[2]	60.3[2]	53.9[2]
	QLD	93.0[2]	7.8[2]	57.3[1]	50.9[1]
	SDM	94.3[2]	8.2[2]	62.3[2]	56.1[2]
基于 BERT	BGE-M3	87.1	6.8	50.0	44.3
基于大语言模型（LLM）	BOOM-4B-v1	96.3	8.4	**65.0**	**59.0**
	Qwen3-Embedding-4B	96.0	**8.5**	64.6	58.1
	Qwen3-Embedding-8B	**96.6**	**8.5**	64.5	58.2
	F2LLM-v2-8B	90.5	7.4	56.3	50.9

■ 关键发现：
- 真实智能体关键词查询下的结果**并不是简单的 “密集模型全面优于稀疏模型**：SDM 的Recall@100 达到 94.3%，高于 BGE-M3 的 87.1%，甚至高于大规模参数没模型 F2LLM-v2-8B 的 90.5%。进一步说明在真实中间查询已经较为精确时，稀疏模型可以低成本提供强候选池，而中等规模密集检索模型并不一定是更好的折中选择。
- 大规模密集向量检索模型在靠前排序结果上保持最好性能，和其他数据集类似。

实验三：效果收益需要结合工程开销判断

表 7 检索模型离线与在线工程开销对比 [1,2,3,4]

模型类型	检索模型	离线耗时 / s	在线查询耗时 / s		硬件运行资源最大足迹 / GB	
		语料向量化/索引	查询语句编码	匹配/相似度检索	最大 GPU 显存	最大 CPU 内存
稀疏检索模型	BM25	9.31	0.00	3.95	0.00	0.67
基于 BERT	BGE-M3	104.68	0.35	0.23	6.95	15.73
基于 LLM	Qwen3-Embedding-4B	292.46	0.27	0.38	48.67	74.09
	Qwen3-Embedding-8B	249.70	0.28	0.22	67.27	108.62

■ 关键发现：
- BM25 离线建索引仅 9.31 s、无需 GPU；BGE-M3 离线处理 104.68 s，约为 BM25 的 11 倍。
- Qwen3-Embedding 需要 48.67–67.27 GB GPU 显存，部署门槛显著更高。
- 因此，**关键词查询或资源受限时优先稀疏检索；语义泛化或前排质量优先且资源充足时采用密集检索。**

结论

- 检索器选择应综合查询形态、语料噪声、指标目标和资源预算。
- 关键词型查询与规整语料下，稀疏检索能够低成本提供可靠候选池。
- 描述型查询、噪声语料或前排质量优先时，密集检索更有优势。
- 智能体检索应采用查询感知和资源感知的动态检索策略。