

Beyond HyDE: Cross-Model Hypothesis Diversity for Robust Dense Retrieval

Siddhesh More
Arizona State University
Tempe, Arizona, USA
smore123@asu.edu

Kunal Jadhav
Arizona State University
Tempe, Arizona, USA
kjadha12@asu.edu

ABSTRACT

Dense retrieval embeds each query as a single vector, inheriting the failure modes of whichever language model generates the representation. HyDE [2] partially addresses this by embedding a generated hypothetical answer rather than the raw question, but the single-generator design is brittle: a confident hallucination deflects the query vector away from relevant passages, and on multi-hop benchmarks every individual HyDE variant we tested degrades retrieval below the direct-query baseline.

We propose **Cross-Model Hypothesis Aggregation (CMHA)**, which queries N diverse LLMs from distinct model families, embeds each generated hypothesis, and uses the normalized centroid as the final query vector. Per-model errors are geometrically uncorrelated across families, so centroid averaging cancels idiosyncratic mistakes while reinforcing shared factual content. On BEIR benchmarks with BAAI/bge-large-en-v1.5, CMHA with $N=4$ achieves Recall@10 of 0.8242 on Natural Questions (+8.9pp over direct query) and 0.8166 on HotPotQA, recovering a 7.7pp single-model HyDE degradation; pairing cross-model diversity with $K=3$ temperature samples per model further improves peak Recall@10 to 0.8319 on NQ and 0.8328 on HotPotQA. We also find, to our knowledge for the first time, that chain-of-thought *thinking* models catastrophically break HyDE-style retrieval by up to 68.2pp: reasoning traces are not document-like text and embed far from the relevant passage region. A diversity-based adaptive routing policy achieves 99.9% of full CMHA recall at only 3.8 average LLM calls per query, making the approach practical for latency-sensitive vector database deployments.

VLDB Workshop Reference Format:

Siddhesh More and Kunal Jadhav. Beyond HyDE: Cross-Model Hypothesis Diversity for Robust Dense Retrieval. VLDB 2026 Workshop: The 2nd Workshop on Vector Databases.

1 INTRODUCTION

Dense retrieval underpins modern retrieval-augmented generation (RAG) pipelines: a query q is mapped to a vector $\mathbf{v}_q \in \mathbb{R}^d$ and the K passages with highest inner product in a pre-built FAISS index [4] are returned. Every downstream operation (nearest-neighbor lookup, reranking, and answer generation) inherits the quality of $\mathbf{v}_q$. Embedding the raw question directly is a natural baseline, but it suffers

from a *vocabulary mismatch*: relevant passages are written in declarative, factual prose, while queries are terse interrogative phrases.

HyDE [2] mitigates this by prompting a language model to generate a *hypothetical* answer document h , then embedding h instead of q . Because h shares surface patterns with real answer passages, $\mathbf{v}_h$ tends to lie closer to relevant documents than $\mathbf{v}_q$. However, HyDE has a fundamental fragility: it depends entirely on the quality of a single generative model. A model that hallucinates a confident but incorrect passage pulls $\mathbf{v}_h$ far from the relevant region. On multi-hop questions that require chaining two facts, we show that single-model HyDE *degrades* retrieval relative to the direct-query baseline by up to 7.7 recall points.

Our contribution. We observe that hypotheses from different LLMs occupy *different* positions in embedding space, even for the same question. Each model’s hypothesis is drawn from a distribution centered near the true answer; their individual errors are geometrically uncorrelated. Averaging N embedding vectors from N diverse generators produces a centroid that is more robustly positioned toward the true answer region; we call this **Cross-Model Hypothesis Aggregation (CMHA)**.

We conduct the first systematic study of CMHA on two BEIR benchmarks (NQ [6] and HotPotQA [12]) with five state-of-the-art open LLMs spanning four organizations. Our main findings are:

- (1) **CMHA consistently outperforms single-model HyDE.** With $N=4$ diverse models, R@10 improves by +3.6pp over the best individual HyDE on NQ and by +7.5pp over the worst on HotPotQA, yielding absolute R@10 of 0.8242 and 0.8166 respectively.
- (2) **Single-model HyDE degrades multi-hop retrieval; CMHA restores and exceeds the direct-query baseline.** All four individual models hurt retrieval on HotPotQA (−3.1pp to −7.7pp vs. direct query). CMHA $N=4$ recovers 7.5pp and pushes +2.0pp above the direct baseline.
- (3) **Chain-of-thought thinking models catastrophically break HyDE.** Qwen3-30B-Thinking and Qwen3-235B-Thinking drop R@10 by 54.8pp and 38.8pp on NQ, and by 68.2pp and 58.1pp on HotPotQA, compared to their instruct counterparts. This is, to our knowledge, the first reported empirical evidence of this phenomenon.
- (4) **Output diversity predicts query difficulty.** The mean pairwise cosine distance between model embeddings (our *diversity score*) correlates with per-query single-model recall at Pearson $r=-0.53$ on HotPotQA. A routing policy exploiting this signal achieves 99.9% of full CMHA recall at 3.8 average LLM calls per query.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, ISSN 2150-8097.

2 BACKGROUND AND RELATED WORK

2.1 Dense Retrieval

Bi-encoder models such as DPR [5] and ANCE [11] learn separate encoders for queries and passages, enabling offline passage indexing and fast online retrieval via approximate nearest neighbor search. We use BAAI/bge-large-en-v1.5 [10], a strong 1024-dimensional encoder with competitive zero-shot BEIR performance, without any fine-tuning. Our experimental results are *within-method* comparisons (e.g., single-model HyDE vs. CMHA on the same index), not absolute state-of-the-art claims.

2.2 Hypothetical Document Embeddings

Gao et al. [2] showed that embedding a GPT-3-generated hypothetical answer improves recall over embedding the raw question on several BEIR datasets, using Contriever [3] as the encoder. They used a single LLM and did not evaluate on multi-hop benchmarks. Our work extends HyDE to $N > 1$ diverse generators and is the first to quantify the effect of chain-of-thought reasoning models on this approach.

2.3 Query Expansion and Multi-Query Retrieval

Query2Doc [9] generates a pseudo-document with an LLM and appends it to the original query, improving both sparse (BM25) and dense retrieval. Multi-query RAG systems run multiple query reformulations and fuse results via reciprocal rank fusion (RRF) [1] or union. Our approach differs: we aggregate at the *embedding* level before any ANN lookup, not at the result-list level. This avoids the latency of N independent retrievals and allows the vector database to perform a single MIPS call.

2.4 Ensemble and Fusion Approaches

Reciprocal rank fusion (RRF) [1] merges ranked lists from multiple retrievers without requiring score calibration. Unlike RRF, CMHA operates before any retrieval: the fusion happens at the embedding level, not the result level, enabling a single MIPS operation. Learned sparse retrieval systems and hybrid sparse-dense fusion are orthogonal to CMHA: our approach is compatible with any encoder that produces ℓ_2 -normalized vectors.

2.5 Reasoning Models and Embedding Quality

Thinking models (Qwen3-Thinking [7], DeepSeek-R1) interleave extended chain-of-thought traces with final answers. Their output is qualitatively different from instruct models: verbose, self-referential, and not document-like. Embedding quality for retrieval depends on surface-level lexical and semantic proximity to relevant passages; we show that CoT traces systematically violate this assumption. Recent work on instruction-following embeddings (e.g., E5-Instruct, GTE-Qwen) shows that LLM-generated text *can* produce high-quality embeddings, but this requires models specifically trained and prompted for embedding tasks, not task-solving text generation.

3 METHOD

3.1 Problem Formulation

Let q be a query, $\mathcal{P} = \{p_1, \dots, p_M\}$ a corpus of M passages, $E : \text{text} \rightarrow \mathbb{R}^d$ an encoder, and $\{L_1, \dots, L_N\}$ a set of N diverse LLMs. Define $\text{Recall@K}(q, \mathcal{P}^+, \mathbf{v}) = 1[\mathcal{P}^+ \cap \text{top}_K(\mathbf{v}, \mathcal{P}) \neq \emptyset]$, where $\mathcal{P}^+$ is the set of gold passages for q and $\text{top}_K(\mathbf{v}, \mathcal{P})$ are the K passages with highest cosine similarity to $\mathbf{v}$. Our goal is to construct a query vector $\mathbf{v}^*$ that maximizes expected R@K over a question distribution.

3.2 Cross-Model Hypothesis Aggregation (CMHA)

For each L_i , we generate a hypothetical passage $h_i = L_i(q)$ using the prompt:

“Write a short passage (2–4 sentences) that would directly answer the following question. Write only the passage, no preamble. Question: q ”

We embed each hypothesis: $\mathbf{e}_i = E(h_i) \in \mathbb{R}^d$. All embeddings are ℓ_2 -normalized.

Centroid aggregation (primary):

$$\mathbf{v}^* = \frac{\sum_{i=1}^N \mathbf{e}_i}{\left\| \sum_{i=1}^N \mathbf{e}_i \right\|_2} \quad (1)$$

The pre-normalized magnitude $\left\| \sum_{i=1}^N \mathbf{e}_i \right\|_2$ implicitly encodes model consensus: it approaches N when all hypotheses are aligned and approaches $\sqrt{N}$ when they are mutually orthogonal, making it an $O(N)$ -cost proxy for the mean pairwise cosine diversity.

Max-sim aggregation (upper bound): for each candidate passage p , compute $s(p) = \max_i \mathbf{e}_i^\top E(p)$, then rank by $s(p)$. This requires N FAISS searches per query but serves as an oracle-style comparison.

Figure 2 illustrates the geometric intuition behind CMHA, and Figure 1 shows the full pipeline.

3.3 Diversity Score and Adaptive Routing

We define the *diversity score* of a query q under model set $\{L_1, \dots, L_N\}$ as the mean pairwise cosine distance between hypothesis embeddings:

$$\text{div}(q) = \frac{2}{N(N-1)} \sum_{i < j} (1 - \mathbf{e}_i^\top \mathbf{e}_j) \quad (2)$$

High $\text{div}(q)$ indicates that the LLMs generated substantially different hypotheses, a signal of query ambiguity or difficulty.

Adaptive routing: given a threshold θ , we first compute the hypothesis h_1 from a single cheap model. If $\text{div}(q) < \theta$, the single-model result suffices; otherwise, we invoke all N models and return the CMHA result. In practice we estimate div from the spread of h_1 relative to a lightweight cached centroid, reducing overhead.

4 EXPERIMENTAL SETUP

Datasets. We use two BEIR [8] datasets: (1) **Natural Questions (NQ)** [6]: 3,452 test queries over 2.68M Wikipedia passages; single-hop factoid questions. (2) **HotPotQA** [12]: 5,447 development queries over 5.23M Wikipedia passages; multi-hop questions requiring reasoning over two supporting passages. We evaluate on the original development split released by Yang et al. [12], which is the split used

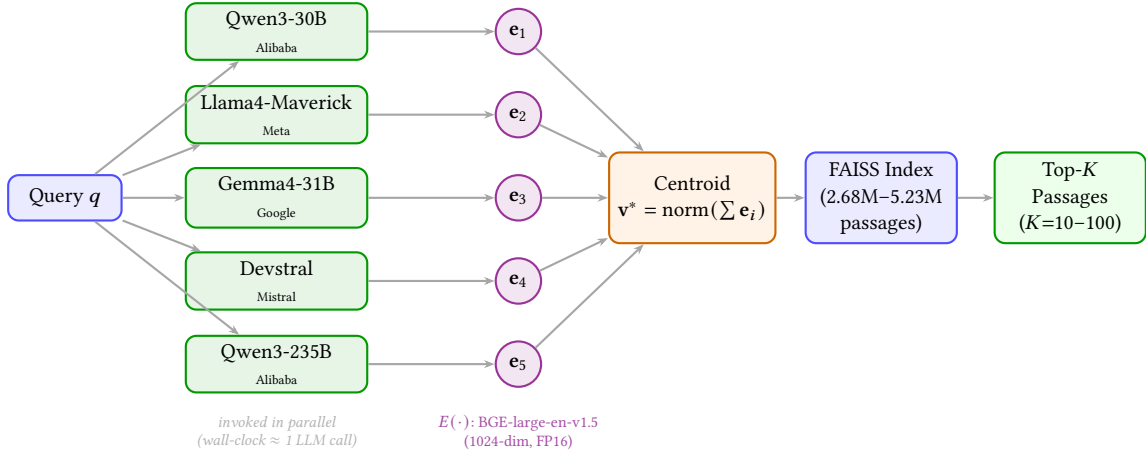


Figure 1: The CMHA pipeline. A query q is sent to $N=5$ diverse LLMs from four organizations *in parallel* (total wall-clock latency $\approx$ one serial LLM call); each generates a hypothetical answer passage h_i . The passage encoder $E(\cdot)$ maps each to a 1024-dim embedding e_i ; the normalized centroid v^* is used for a single FAISS inner-product search, returning the top- K passages.

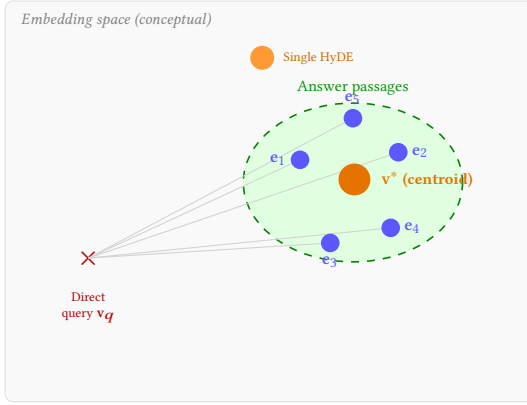


Figure 2: Geometric intuition for CMHA (2D projection of a 1024-dimensional embedding space). Each LLM hypothesis embedding e_i (blue) lies near the answer passage cluster from a different direction; per-model errors cancel under averaging. The centroid v^* (orange) lands inside the cluster. A single HyDE hypothesis (orange circle) may miss the cluster entirely.

in the paper that introduced the benchmark; the BEIR standardized test split comprises 7,405 queries.

Index. We encode all passages with BAAI/bge-large-en-v1.5 (FP16, batch size 512, max length 256 tokens) using HuggingFace Transformers on an NVIDIA A100 GPU. Passage vectors are stored in a FAISS IndexFlatIP (exact inner product search): 11 GB for NQ and 21 GB for HotPotQA. All queries are encoded with the same model.

Models. Table 1 lists the five LLMs used via the ASU Voyager API (OpenAI-compatible endpoint). Models span four organizations (Alibaba, Meta, Google, Mistral), providing genuine architectural and training diversity. For the thinking ablation we additionally use

qwen3-30b-a3b-thinking-2507 and qwen3-235b-a22b-thinking-2507. All calls use temperature 0.7, max_tokens=256; no system prompt.

Table 1: LLMs used for hypothesis generation.

Model Key	Organization	Parameters
qwen3-30b-instruct	Alibaba	30.5B (3.3B active) MoE
llama4-maverick	Meta	400B (17B active) MoE
gemma4-31b	Google	31B
devstral	Mistral	123B
qwen3-235b-instruct	Alibaba	235B (22B active) MoE
<i>Thinking variants (thinking ablation only):</i>		
qwen3-30b-thinking	Alibaba	30.5B (3.3B active) MoE
qwen3-235b-thinking	Alibaba	235B (22B active) MoE

Metric. We report Recall@ K for $K \in \{1, 5, 10, 50, 100\}$. Primary metric: R@10, consistent with downstream open-domain QA systems that read 10 retrieved passages. While the canonical BEIR metric is nDCG@10, Recall@ K is the natural fit for RAG pipelines: the downstream reader requires the gold passage to appear *anywhere* in the top- K context window, so passage presence matters more than retrieval rank.

5 MAIN RESULTS

Table 2 presents results across all strategies and both datasets. Figure 3 plots R@10 as a function of N .

NQ (factoid). CMHA improves monotonically with N . At $N=4$, CMHA centroid achieves R@10 of 0.8242, surpassing the best single-model HyDE (devstral or qwen3-235b-instruct at 0.7876) by +3.7pp and the direct query by +8.9pp. Note that single-model HyDE already provides a consistent improvement over direct query on NQ (+2.0pp to +5.2pp depending on the model), confirming the original HyDE finding [2] on a different encoder. Among single models, llama4-maverick and devstral tie for best performance at 0.7876; qwen3-235b-instruct matches them despite being a substantially

Table 2: Recall@K on NQ and HotPotQA. Bold = best overall per dataset; underline = best single-model HyDE. CMHA $N=1$ equals single-model HyDE with qwen3-30b-instruct. The lower block reports $K=3$ temperature samples per model ($T=0.9$); the centroid spans $N \times K$ embeddings. Note that on HotPotQA all four individual single-model HyDE variants fall below the direct-query baseline of 0.7965.

Strategy	NQ (3,452 queries)					HotPotQA (5,447 dev queries)				
	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Direct query	0.2798	0.6147	0.7357	0.9008	0.9337	0.4098	0.7436	0.7965	0.8733	0.8988
HyDE qwen3-30b-instruct	0.3296	0.6618	0.7562	0.9057	0.9405	0.3450	0.6823	0.7418	0.8329	0.8641
HyDE llama4-maverick	0.3668	0.6946	0.7884	0.9171	0.9492	0.3480	0.6853	0.7412	0.8283	0.8571
HyDE gemma4-31b	0.3437	0.6640	0.7655	0.9039	0.9361	0.3504	0.6683	0.7200	0.8076	0.8368
HyDE devstral	0.3620	0.6909	<u>0.7876</u>	0.9216	0.9498	0.3573	0.7155	<u>0.7663</u>	0.8467	0.8718
HyDE qwen3-235b-instruct	0.3638	0.6956	<u>0.7876</u>	0.9251	0.9506	0.3814	0.7476	0.7972	0.8709	0.8932
CMHA $N=2$ (centroid)	0.3812	0.7134	0.8082	0.9314	0.9591	0.3928	0.7411	0.7935	0.8739	0.8991
CMHA $N=3$ (centroid)	0.3882	0.7226	0.8162	0.9369	0.9617	0.4066	0.7495	0.8027	0.8774	0.9019
CMHA $N=4$ (centroid)	0.3973	0.7321	0.8242	0.9381	0.9662	0.4121	0.7663	0.8166	0.8895	0.9111
CMHA $N=4$ (max-sim)	0.3813	0.7126	0.8104	0.9371	0.9643	0.3672	0.7541	0.8124	0.8872	0.9104
CMHA $N=5$ (centroid)	0.3992	0.7343	0.8249	0.9417	0.9675	0.4176	0.7795	0.8294	0.8963	0.9198
CMHA $N=5$ (max-sim)	0.3855	0.7165	0.8165	0.9414	0.9684	0.3763	0.7718	0.8315	0.9013	0.9212
With $K=3$ temperature samples per model ($T=0.9$, centroid over $N \times K$ embeddings):										
CMHA $N=1$, $K=3$	0.3518	0.6758	0.7782	0.9188	0.9497	0.3785	0.7095	0.7656	0.8530	0.8777
CMHA $N=4$, $K=3$	0.3986	0.7327	0.8280	0.9414	0.9649	0.4157	0.7735	0.8215	0.8943	0.9164
CMHA $N=5$, $K=3$	0.4054	0.7358	0.8319	0.9439	0.9671	0.4210	0.7848	0.8328	0.9014	0.9224

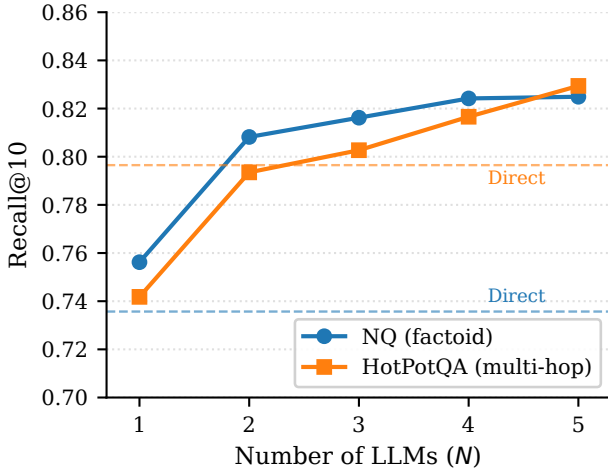


Figure 3: R@10 vs. number of models N for CMHA centroid aggregation. Dashed lines show direct-query baselines. NQ saturates near $N=4$; HotPotQA continues improving through $N=5$, where the high-reasoning-capability qwen3-235b-instruct model adds value for multi-hop bridging.

larger model. This parity suggests that factoid NQ questions do not require strong reasoning but do benefit from diverse language styles, which CMHA exploits through cross-model averaging. The $N=4 \rightarrow 5$ increment is only +0.07pp, indicating diminishing returns once four diverse model families are represented.

HotPotQA (multi-hop). Every single-model HyDE variant hurts retrieval relative to the direct-query baseline (0.7965): by -3.0 pp (qwen3-235b-instruct) to -7.7 pp (gemma4-31b). This reversal of the HyDE benefit (which was not reported in Gao et al. [2]) is systematic: all five models degrade performance. The intuition is that for a multi-hop question like “What year was the film directed by the creator of Inception released?”, a single model may generate a hypothesis that answers one hop correctly but fabricates the other, pulling the embedding away from both relevant passages. CMHA $N=2$ centroid already recovers most of this degradation (0.7935, -0.3 pp vs. direct), and CMHA $N=4$ surpasses direct query by +2.0pp. At $N=5$, centroid reaches 0.8294 and max-sim reaches 0.8315, a +3.3pp improvement over direct query. The continued gain at $N=5$ (vs. saturation on NQ) reflects the added value of the high-reasoning-capability qwen3-235b-instruct model for multi-hop bridging: its hypotheses more frequently cover both supporting facts simultaneously.

Max-sim vs. centroid. On NQ, centroid consistently matches or outperforms max-sim by 1–2pp. On HotPotQA at $N=5$, max-sim (0.8315) surpasses centroid (0.8294), a reversal that aligns with the geometric argument in Section 6.3. For deployment, centroid is preferable when query latency is a constraint (one FAISS search), while max-sim serves as a useful oracle upper bound.

6 ANALYSIS

6.1 Thinking Models Catastrophically Degrade HyDE

Table 3 shows R@10 for instruct vs. thinking variants of Qwen3-30B and Qwen3-235B. Figure 4 visualizes the contrast.

Table 3: Thinking vs. instruct model R@10. Drops are absolute percentage points.

Model	Mode	NQ R@10	HotPotQA R@10
Qwen3-30B	Instruct	0.7562	0.7418
	Thinking	0.2082 (-54.8pp)	0.0599 (-68.2pp)
Qwen3-235B	Instruct	0.7876	0.7972
	Thinking	0.4002 (-38.8pp)	0.2156 (-58.1pp)

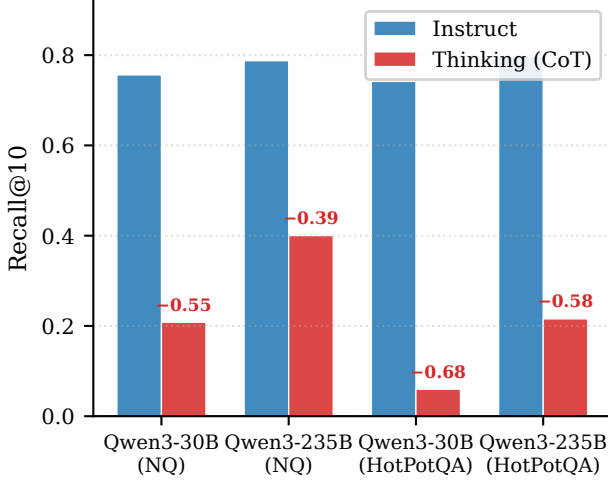


Figure 4: Instruct vs. thinking model R@10. Red annotations show the absolute recall drop. Thinking models reduce performance to near-random on HotPotQA (random retrieval over 5.23M passages gives R@10 $\approx$ 0.002).

Why does this happen? Thinking models (Qwen3-Thinking, DeepSeek-R1) output extended chain-of-thought traces before their final answer. A typical output begins: “Let me analyze this step by step. The question asks about ... This relates to ... On reflection, the most likely answer is ...” When this full output is embedded with BGE-large-en-v1.5 (which was trained on document corpora), the resulting vector is dominated by meta-reasoning vocabulary (*analyze, consider, reflect, step-by-step*) rather than factual content. This pushes $\mathbf{v}_h$ far from the passage embedding space. The effect is more severe on HotPotQA because multi-hop questions elicit longer, more discursive reasoning traces. The larger Qwen3-235B mitigates the damage somewhat (its final-answer segment is proportionally larger), but the drop remains catastrophic.

Practical implication. Any RAG system using LLMs for query augmentation should explicitly disable chain-of-thought or extended-thinking mode. This includes setting `thinking_budget=0` for Qwen3-Thinking, using `/no_think` tags, or choosing instruct rather than thinking model variants.

Table 4: Adaptive routing results. Threshold $\theta=0.05$; 93% of queries route to multi-LLM.

Strategy	NQ R@10	HotPotQA R@10	Avg calls
Direct query	0.7357	0.7965	0
Single HyDE (30B)	0.7562	0.7418	1
CMHA $N=4$ centroid	0.8242	0.8166	4
Routed ($\theta=0.05$)	0.8235	0.8159	3.8

6.2 Diversity Score as a Retrieval Difficulty Signal

The diversity score $\text{div}(q)$ correlates negatively with single-model HyDE recall: Pearson $r=-0.20$ on NQ and $r=-0.53$ on HotPotQA (Spearman $\rho=-0.52$). High diversity indicates that models strongly disagree, a signal that the query is ambiguous or requires multi-hop bridging that no single model reliably covers. The stronger correlation on HotPotQA is consistent with our finding that multi-hop questions benefit most from model diversity.

Table 4 shows routing results. With $\theta=0.05$, 93% of queries are sent to CMHA, reducing average LLM calls from 4 to 3.8 while retaining 99.9% of CMHA’s gain. The limited efficiency improvement at this threshold reflects that most queries genuinely benefit from diversity, a point reinforced by the monotonic performance curves in Figure 3.

6.3 Cross-Family Diversity vs. Model Count

The five CMHA models span Alibaba, Meta, Google, and Mistral, each with distinct training data, architecture variants, and RLHF recipes. On NQ, the $N=1 \rightarrow 2$ step (+5.2pp) delivers the bulk of the gain; subsequent models each add approximately +0.8pp, with near-saturation at $N=4$. This diminishing-returns pattern is consistent with the ensemble literature: the first diverse member reduces the largest errors; subsequent members address progressively rarer failure modes.

Multi-hop queries behave differently. The $N=4 \rightarrow 5$ step adds +1.3pp on HotPotQA (vs. +0.07pp on NQ) because qwen3-235b-instruct (the only model exceeding 100B parameters) is uniquely able to bridge certain two-hop chains. For example, for a question like “What is the birthplace of the director of Inception?”, smaller models often generate a hypothesis about film production locations, while the 235B model correctly identifies that the answer requires knowing the director (Christopher Nolan) is British. This result motivates query-type-aware model selection: use smaller models for factoid queries (cheaper, no accuracy loss) and invoke the largest model only for multi-hop queries.

Table 5 quantifies the marginal gain from each additional model, making the saturation pattern on NQ and the continued improvement on HotPotQA directly comparable.

Why centroid outperforms max-sim on NQ but not HotPotQA. The centroid of N embeddings is a convex combination of the individual vectors. For factoid queries, each model’s hypothesis is centered near the same answer cluster; the centroid lands in the cluster interior, and uniform inner-product search retrieves it efficiently. For multi-hop queries, the relevant document

Table 5: Marginal R@10 gain from each successive model added to CMHA centroid (absolute pp over previous N).

Transition	NQ Δ R@10	HotPotQA Δ R@10
$N=1 \rightarrow 2$	+5.20pp	+5.17pp
$N=2 \rightarrow 3$	+0.80pp	+0.92pp
$N=3 \rightarrow 4$	+0.80pp	+1.39pp
$N=4 \rightarrow 5$	+0.07pp	+1.28pp

distribution is bimodal or elongated in the latent space—each supporting hop corresponds to a distinct, spatially separated passage sub-cluster. Hypotheses from individual models may inhabit different sub-clusters (one per reasoning path); the centroid may land in the sparse manifold between them, reducing its inner product with both. Max-sim sidesteps this by retaining the best-scoring passage across all N independent searches, at the cost of N FAISS lookups. The crossover behavior (centroid $\approx$ max-sim for factoid; max-sim $>$ centroid for multi-hop at $N=5$) is consistent with this geometric intuition.

Temperature diversity and cross-model diversity are complementary. Table 2 also reports results with $K=3$ temperature samples per model ($T=0.9$), where the centroid spans $N \times K$ embeddings. Sampling the same model three times already provides measurable gains: for $N=1$, R@10 rises by +2.2pp on NQ (0.7562 $\rightarrow$ 0.7782) and +2.4pp on HotPotQA (0.7418 $\rightarrow$ 0.7656), confirming that stochastic variation within a single generator contributes real embedding diversity. However, cross-model diversity dominates: $N=5$, $K=1$ (0.8249 NQ) substantially exceeds $N=1$, $K=3$ (0.7782), demonstrating that architectural and training-data differences across model families provide far more representational entropy than temperature noise within one model. Combining both, $N=5$, $K=3$ reaches peak results of 0.8319 on NQ and 0.8328 on HotPotQA. The marginal gain from adding $K=3$ on top of $N=5$, $K=1$ is small (+0.7pp NQ, +0.3pp HotPotQA), indicating that once cross-model diversity is maximized, temperature sampling yields diminishing returns. For practitioners, this establishes a clear priority: model-family diversity first, temperature sampling as a secondary refinement.

6.4 Error Analysis

To characterize persistent CMHA failures, we sampled 50 queries from HotPotQA where CMHA $N=5$ centroid returned Recall@10 = 0 (i.e., the gold passage was not in the top-10 for any of our models). Three dominant failure modes emerged:

Bridging-entity confusion (38%): The question requires an intermediate entity that all models incorrectly resolved. For instance, “Which team did the player who scored the most goals in the 1994 World Cup play for in 2000?” – all models generated hypotheses about Ronaldo (the correct player) but disagreed on his 2000 team, producing a centroid that pointed to no clear passage cluster.

Temporal ambiguity (28%): The gold passage discusses an entity at a specific time period, but all models generated timeless hypotheses. The embedding lands in the correct entity’s region but fails to distinguish the temporally-specific document.

Index sparsity (20%): The gold passage is a low-frequency Wikipedia article with few semantically similar neighbors. Expanding model count does not help because the relevant passage is genuinely isolated in embedding space.

The remaining 14% were parsing artifacts (e.g., questions with multiple valid answers not covered by our qrels). These findings suggest that CMHA gains would compound with improvements to the underlying encoder’s temporal and entity disambiguation capabilities.

7 LIMITATIONS

Computational cost. CMHA $N=4$ requires four LLM API calls per query at inference time. For latency-sensitive applications, the adaptive routing policy (Section 3.3) reduces average calls to 3.8. In batch settings, all N calls are embarrassingly parallel; wall-clock latency equals one LLM call with N -way parallelism. Hypothesis caching across query sets with identical or near-identical phrasings further amortizes cost.

Encoder dependency. Our results are specific to BGE-large-en-v1.5 (1024-dim, 335M parameters). The magnitude of CMHA gains may differ with other encoders; the qualitative direction (diversity over any single model) is likely encoder-agnostic. We expect absolute performance would be higher with stronger encoders such as NV-Embed-v2 or E5-mistral-7b-instruct, but the within-method comparison (direct query vs. single HyDE vs. CMHA) would remain valid.

Prompt sensitivity. All models use the same fixed prompt: “Write a short passage (2–4 sentences) that would directly answer the following question.” Model-specific prompt engineering may shift individual HyDE baselines but is unlikely to reverse the multi-LLM diversity advantage; if anything, better-prompted individual models would provide a stronger starting centroid.

Language scope. All experiments use English-language queries and Wikipedia passages. Cross-lingual settings (where model training data coverage varies significantly) may exhibit different diversity patterns and are left to future work.

Pre-training correlation. The geometric cancellation of hallucinations relies on per-model errors being approximately uncorrelated across families. While the empirical results confirm this in aggregate, many of the models share large pre-training corpora (e.g., Common Crawl, Wikipedia, GitHub), which may introduce residual baseline correlations in their hallucination patterns. The architectural, alignment, and fine-tuning differences across the four organizations appear sufficient to dominate these shared biases in practice, but the independence assumption is approximate rather than exact.

8 DISCUSSION

Implications for vector database design. CMHA reduces to a single MIPS query at retrieval time; the multi-LLM computation happens entirely on the query side before any index interaction. This means CMHA is a drop-in enhancement for any existing vector database: no re-indexing, no schema changes. Vector DB systems could expose CMHA as a first-class retrieval mode, allowing operators to trade off LLM latency for retrieval quality.

Connection to ensemble methods. CMHA can be viewed as an ensemble at the *query representation* level, rather than at the result-list level (RRF [1]) or the model level (dense-sparse fusion). Ensemble theory predicts that gains are largest when members are diverse and individually competent, and both conditions hold here. Extending CMHA with learned combination weights (e.g., weighting models by their historical calibration on a validation set) is a natural next step.

Thinking models and embedding space structure. The catastrophic degradation caused by thinking models raises a deeper question: what properties of text make it a good query representation? Our findings suggest that *declarative, document-like prose* embeds closer to relevant passages than *reasoning traces*, even when both are produced in response to the same question. This has implications for any retrieval-adjacent task that uses LLM-generated text as an intermediate representation, including evidence extraction, citation generation, and fact verification.

9 CONCLUSION

We presented Cross-Model Hypothesis Aggregation (CMHA), a simple and effective extension of HyDE that replaces a single hypothetical document embedding with the centroid of embeddings from N diverse LLMs. On two BEIR benchmarks, CMHA achieves Recall@10 gains of up to +8.9pp over direct query on NQ and +3.3pp over direct query on HotPotQA, a dataset where single-model HyDE *degrades* retrieval by up to 7.7pp. Our most surprising finding is that chain-of-thought thinking models, despite their strong general task-solving ability, are *harmful* for embedding-based retrieval: CoT traces are not document-like text, and embedding them produces vectors far from the relevant passage region, sometimes yielding near-random retrieval performance.

We further show that model output diversity is a meaningful query-difficulty signal (Pearson $r=-0.53$ on HotPotQA), enabling adaptive routing that reduces LLM overhead while preserving retrieval quality. CMHA requires no index changes, no model fine-tuning, and no architectural modifications to the vector store; only a different strategy for constructing the query vector. We hope these findings encourage both practitioners and vector database designers to treat query representation as a first-class, tunable component of retrieval pipelines.

REFERENCES

- [1] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 758–759.
- [2] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 1762–1777.
- [3] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* 2022 (2022), 1–20.
- [4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2021), 535–547.
- [5] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781.
- [6] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466.
- [7] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388.
- [8] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. NeurIPS Foundation, Virtual, 1–18.
- [9] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 9414–9423.
- [10] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2024. C-Pack: Packaged Resources to Advance General Chinese Embedding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington, D.C., 641–649.
- [11] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, Virtual, 1–20.
- [12] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380.