Escaping the Tokenizer Trap: Probabilistic Semantic Routing for Dense Retrieval in Specialized Domains

Siu Hin Ng[1,2,3], Yung-Chun Chang[4], Chao-Lin Liu[1,3], Wei-Yun Ma[1,2], Wen-Lian Hsu[5]
[1]Social Network and Human-Centered Computing (SNHCC),
Taiwan International Graduate Program (TIGP), [2]Academia Sinica,
[3]National Chengchi University, [4]Taipei Medical University, [5]Asia University
{ng_siuhin, ma}@iis.sinica.edu.tw, changyc@tmu.edu.tw,
chaolin@nccu.edu.tw, wenhsu@asia.edu.tw

Abstract

Dense retrievers can blur the distinctions between short, jargon-heavy records that share most of their vocabulary. We study Lex-TM, a training-free layer that interpolates dense cosine similarity with the Jensen–Shannon similarity of topic distributions from an LDA model whose topic–word prior is scaled by a Herfindahl–Hirschman (HHI) exclusivity score. On NTCIR-13 MedWeb, cast as retrieving tweets that share a query's exact symptom-label set, Lex-TM reaches an English MRR of 0.8151 against 0.7081 for BM25+dense reciprocal rank fusion, a significant paired difference; in Japanese and Chinese the differences from fusion are not significant. With Qwen3-Embedding-0.6B, routing improves on dense retrieval significantly in English and Chinese. Three findings bound these gains. SPLADE++ and BGE-M3's multi-vector mode match or exceed Lex-TM. In English and Chinese the gain depends on the label strings in the indexed tweets: without them, routing no longer improves on dense retrieval. The HHI prior brings no benefit: on short texts it becomes a rarity weight, and a symmetric prior performs as well across five Gibbs seeds. Outside MedWeb, routing matches dense retrieval on Chinese medical QA and lowers nDCG@10 on ChatDoctor, FiQA and long biomedical abstracts. Code, seeds and a changelog of fixed bugs are released.

1 Introduction

Retrieval-augmented generation (RAG) grounds language models in retrieved evidence (Lewis et al., 2020; Asai et al., 2024), so errors made by the retriever become errors made by the generator. Current retrievers are mostly dense bi-encoders: multilingual encoders such as BGE-M3 (Chen et al., 2024), embedders built on small language models such as Qwen3-Embedding (Zhang et al., 2025), and retrievers built on modern bidirectional encoders such as ModernBERT (Warner et al., 2025). Each maps a query and a document to a vector and ranks documents by cosine similarity.

In specialized collections of short records, many documents share most of their words. Patient reports mention the same few complaints (fever, runny nose, cough), and what separates one report from another is a small number of distinguishing terms. A dense encoder summarizes each record as a single vector, which can be dominated by the shared content, so a query can land nearest to a record that is close but wrong. We call this the *tokenizer trap*. Zero-shot dense retrievers are known to trail lexical matching on many specialized corpora (Thakur et al., 2021), and the trap is one plausible cause.

One might expect the problem, and the benefit of fixing it, to be larger in Chinese and Japanese, where word segmentation produces dense overlaps of short character sequences. Our results do not show a uniform benefit for these languages. Several language-specific factors pull in different directions: how the segmenter treats the distinguishing terms, the strength of the lexical baseline, the quality of the dense encoder in each language, and document length (Section 6.4).

Stronger retrievers address the trap by learning finer-grained representations: learned sparse models such as SPLADE (Formal et al., 2021, 2022), late-interaction models such as ColBERT (Khattab and Zaharia, 2020), and domain-adapted encoders. These require training, larger indexes, or both. Fusing lexical and dense rankings (Cormack et al., 2009) is training-free, and filtering retrieved passages with a language model (Asai et al., 2024) adds inference cost at query time.

We study a lighter alternative: keep the dense encoder frozen and add a second similarity channel computed from topic distributions. Lex-TM fits a latent Dirichlet allocation (LDA) model (Blei et al., 2003; Griffiths and Steyvers, 2004) to the collection with a Dirichlet prior scaled by

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 320–332
November 6-10, 2026 ©2026 Association for Computational Linguistics

a Herfindahl–Hirschman exclusivity score, and interpolates dense cosine similarity with the Jensen–Shannon similarity between the topic distributions of the query and the document. The topic model runs on a CPU and needs no labeled data. Figure 1 shows both channels on a MedWeb query. Our contributions are:

- A training-free routing layer that interpolates a frozen dense retriever with topic-distribution similarity, with an HHI-scaled asymmetric topic–word prior (Section 3).
- An evaluation on MedWeb in three languages against lexical, dense, fusion, learned-sparse and multi-vector baselines, under two encoders and with paired significance tests, together with further medical, non-medical, open-domain and long-document collections (Section 5). Lex-TM improves significantly over fusion in English, and the differences in Japanese and Chinese are not significant; SPLADE++ and BGE-M3's multi-vector mode match or exceed it, while BGE-M3's learned-sparse mode does not.
- An analysis of when routing helps and why (Section 6). The gain comes from routing through topics, not from the prior, which on short texts reduces to a rarity weight akin to inverse document frequency. In English and Chinese the topics lock onto the label strings that the indexed tweets contain: without them, routing no longer improves on dense retrieval. Outside MedWeb, routing does not help. We also state precisely how MedWeb queries are constructed: each query is a label string that every relevant document contains.

2 Related Work

Dense, sparse and late-interaction retrieval. Dense passage retrieval (Karpukhin et al., 2020) replaced lexical matching with BM25 (Robertson and Zaragoza, 2009) as the default first stage for RAG; current encoders include BGE-M3 (Chen et al., 2024), which also produces lexical weights and multi-vector representations, and LLM-based embedders such as Qwen3-Embedding (Zhang et al., 2025). BEIR showed that dense models often trail BM25 outside their training domain (Thakur et al., 2021); larger encoders narrow the gap (Ni et al., 2022). Learned sparse retrievers (Formal et al., 2021, 2022) and late-interaction retrievers (Khattab and Zaharia, 2020) keep term-level evidence that single-vector encoders compress away. Reciprocal rank fusion (RRF) (Cormack et al., 2009) is the standard training-free way to combine lexical and dense rankings, and we use it as our main baseline.

Topic models for retrieval. LDA (Blei et al., 2003) fitted by collapsed Gibbs sampling (Griffiths and Steyvers, 2004) was used for ad-hoc retrieval by interpolating topic-based document models with query likelihood (Wei and Croft, 2006); Lex-TM's topic channel follows the same idea but interpolates with a frozen dense encoder and compares topic distributions with a Jensen–Shannon distance. Wallach et al. (2009) found that an asymmetric Dirichlet prior over document–topic distributions helps LDA, whereas an asymmetric prior over topic–word distributions brings no real benefit; our HHI-scaled prior is of the second kind, and Section 6.2 finds no benefit from it either. Short texts are a known weak spot for LDA (Yan et al., 2013). Embedding-based topic models such as BERTopic (Grootendorst, 2022) and CTM (Bianchi et al., 2021) produce topics as an output rather than as a retrieval signal.

3 Method

Let $\mathcal{D}$ be a collection of N documents over a vocabulary $\mathcal{V}$, modeled with K topics.

3.1 HHI-Scaled Topic–Word Prior

For a term v let $c_{v,d}$ be its count in document d and $c_v = \sum_d c_{v,d}$. With $P(d \mid v) = c_{v,d}/c_v$, the Herfindahl–Hirschman index measures how concentrated the term is across documents,

$$\mathrm{HHI}_v = \sum_{d \in \mathcal{D}} P(d \mid v)^2, \tag{1}$$

and its min–max normalization over the vocabulary gives the exclusivity score $E_v \in [0, 1]$. Terms with $E_v > \tau$ receive amplified prior mass,

$$\beta_v = \begin{cases} \beta_{\text{base}}\,(1 + \lambda E_v) & \text{if } E_v > \tau, \\ \beta_{\text{base}} & \text{otherwise,} \end{cases} \tag{2}$$

and the vector $\vec{\beta}$ replaces the symmetric β in the topic–word term of the collapsed Gibbs conditional,

$$P(z_i{=}k \mid \cdot) \propto \frac{n^{(w_i)}_{k,-i} + \beta_{w_i}}{n_{k,-i} + \sum_v \beta_v} \left(n^{(k)}_{d,-i} + \alpha\right). \tag{3}$$

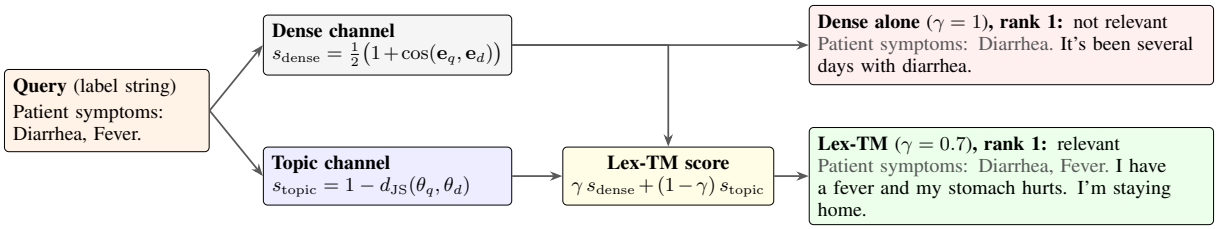


Figure 1: A real case from the English MedWeb evaluation (BGE-M3). Documents are indexed as their label string (grey) followed by the tweet. Ranking by the dense channel alone ($\gamma = 1$) puts a document with a different label set first, and the first relevant document is at rank 6; the Lex-TM score ($\gamma = 0.7$) ranks a document with the query's exact label set first.

Setting $\tau \geq 1$ recovers standard LDA with a symmetric prior. If every term occurs at most once per document, $P(d \mid v) = 1/\mathrm{df}_v$ for each of the df_v documents containing v, so $\mathrm{HHI}_v = 1/\mathrm{df}_v$ exactly and E_v is a decreasing function of document frequency: the exclusivity score is then a rarity score, like inverse document frequency. Section 6.2 shows that MedWeb is in this regime.

3.2 Topic Inference

Gibbs sampling yields topic–word distributions ϕ_k. A document or query x with term counts $n_{x,w}$ receives the topic distribution

$$\theta_x(k) \propto \alpha + \sum_w n_{x,w}\, \phi_{k,w}, \tag{4}$$

a single fold-in step applied to documents at indexing time and to queries at search time.

3.3 Routing Score

With a frozen dense encoder producing normalized embeddings $\mathbf{e}_q, \mathbf{e}_d$, the score of document d for query q is

$$\begin{aligned} s(q,d) = \gamma_q\, \tfrac{1}{2}\big(1 + \cos(\mathbf{e}_q, \mathbf{e}_d)\big) \\ + (1-\gamma_q)\big(1 - d_{\text{JS}}(\theta_q, \theta_d)\big), \end{aligned} \tag{5}$$

where d_{JS} is the Jensen–Shannon distance, the square root of the Jensen–Shannon divergence computed with natural logarithms (as in SciPy). $\gamma = 1$ recovers dense retrieval and $\gamma = 0$ ranks by topic similarity alone. Two rules adjust the weight γ_q per query. *Entropy adaptation* lowers it when the query's topic distribution is concentrated, $\gamma_q = \gamma - 0.2\,(1 - H(\theta_q)/\log K)$, clipped to $[\gamma - 0.2, \gamma]$; our main configuration uses it, although it changes MRR by at most 0.0009 (seed 42; Table 3). A *sparsity gate* sets $\gamma_q = 1$ when the query has fewer than δ in-vocabulary terms, because a topic distribution inferred from two or three words is close to uniform. We apply the gate ($\delta = 5$) to the open-domain sets with short factoid queries and not to MedWeb, whose label-string queries have 2–7 content terms: with $\delta = 5$ the gate would switch routing off for 21 of the 39 MedWeb queries. Appendix A gives the full procedure (Algorithm 1).

4 Experimental Setup

4.1 MedWeb and Its Label-Set Protocol

NTCIR-13 MedWeb (Wakamiya et al., 2017) provides 2,560 pseudo-tweets in each of English, Japanese and Chinese (training and test portions combined; the partitions are parallel), each annotated with eight binary symptom labels: influenza, diarrhea, hay fever, cough, headache, fever, runny nose and cold. MedWeb has no queries or relevance judgments, so we derive both from the labels. A tweet's positive labels are rendered as a canonical label string, for example "Patient symptoms: Hayfever, Runnynose." ("Routine health status." if no label is positive), with fixed Japanese and Chinese equivalents. Each distinct label string is a query, and every tweet with exactly that label set is relevant. Each partition has 39 label sets, with 1 to 725 relevant tweets (median 6; 21 sets have fewer than ten). Each tweet is indexed as its label string followed by the tweet text.

Two consequences follow, and they matter for reading the MedWeb results. First, every relevant document contains its query verbatim, and documents with overlapping label sets contain most of it: the task is to rank an exact label-set match above near misses such as {Hayfever, Fever, Runnynose} for the query {Hayfever, Runnynose}. This is the saturation setting we target, but it measures discrimination between label sets, not clinical relevance. Second, the label strings carry much of the lexical signal: removing them from the indexed text low-

ers Pure BM25 from 0.6566 to 0.1398 in English (0.5855 to 0.1906 in Japanese, 0.7579 to 0.2435 in Chinese); Section 6.3 repeats this control for the dense and routed systems. After preprocessing, a document has 9.2, 8.4 and 7.8 content tokens in English, Japanese and Chinese, including its label string, and 5.9, 5.1 and 4.8 without it; before filtering, an English tweet has 12.9 words, and Japanese and Chinese tweets have 13.4 and 11.2 segmenter tokens, punctuation included.

4.2 Other Collections

ChatDoctorRetrieval (Enevoldsen et al., 2025; Li et al., 2023) pairs 5,591 patient questions with physician answers (5,545 answers, one relevant each; we evaluate the 5,583 questions whose answer survives chunking); CmedqaRetrieval (Xiao et al., 2024; Zhang et al., 2018) pairs Chinese medical questions with answers (dev split, 3,999 questions, full corpus of 100,001 answers); FiQA-2018 (Maia et al., 2018) is financial question answering, our domain beyond medicine. MIRACL-zh (Zhang et al., 2023) (a 100k-passage subset) and subsamples of Mr. TyDi (Zhang et al., 2021) (Japanese, Thai) test short open-domain queries; TREC-COVID (Roberts et al., 2020) and NFCorpus (Boteva et al., 2016) from BEIR (Thakur et al., 2021) test long biomedical documents; two news and two review collections are reported in Appendix H.

4.3 Systems

Pure BM25 (Okapi, k_1=1.5, b=0.75) uses the same tokenizer as the topic model: NLTK with lemmatization for English, MeCab with IPAdic for Japanese and jieba for Chinese, keeping tokens of at least three Latin or two CJK characters after stop-word removal. *Pure Dense* uses BGE-M3; our stronger encoder is Qwen3-Embedding-0.6B with its query instruction (documents are not prompted). *RRF* fuses the BM25 and dense rankings with k=60. As learned-sparse and late-interaction baselines we use BGE-M3's lexical-weight and multi-vector outputs (the latter scored ColBERT-style, as the mean over query tokens of the maximum dot product) and their combination with its dense output using the weights recommended with the model (0.4/0.2/0.4); for English we add SPLADE++ (Formal et al., 2022). *LDA routing* is Lex-TM with $\tau = 1.1$, which switches off the HHI amplification; it runs without entropy adaptation, and Table 3 compares it with Lex-TM under the same setting, which isolates the prior's contribution.

4.4 Protocol and Statistics

We report MRR@20 (reciprocal rank of the first relevant document within the top 20, zero otherwise) on MedWeb, and nDCG@10 on ChatDoctor, CmedqaRetrieval, FiQA, TREC-COVID and NFCorpus. MedWeb results cover all 39 label sets of each partition. The submitted version reported results on 32 test label sets after choosing τ on the other seven; this version evaluates every label set (Appendix B gives the history). Neither hyperparameter was chosen strictly out of sample: $\gamma = 0.7$ was fixed after early sweeps over $\gamma \in \{0.7, 0.9, 1.0\}$ on all label sets, and $\tau = 0.8$ was chosen on the seven validation sets, which the reported results include. Table 3 therefore reports both sweeps in full; τ moves English and Chinese MRR by at most 0.0014 and Japanese MRR by 0.0124, so its choice cannot drive the results. We use $\gamma = 0.7$ on every collection, $K = 20$ topics on MedWeb and 50 elsewhere, $\lambda = 2$, $\alpha = 0.1$, $\beta_{\text{base}} = 0.01$ and 500 Gibbs sweeps with seed 42 (Section 6.1 repeats the MedWeb runs with four further seeds). We report 95% bootstrap confidence intervals (Efron and Tibshirani, 1993) and, for Lex-TM against each baseline, a paired bootstrap interval for the mean per-query difference (10,000 resamples) together with a paired randomization test (Smucker et al., 2007) (Appendix C); we call a difference significant when the paired interval excludes zero and the randomization test gives $p < 0.05$. With 39 queries per partition, intervals are wide.

5 Results

5.1 MedWeb

Table 1 gives the main results. In English, Lex-TM reaches an MRR of 0.8151 against 0.7081 for RRF (+15.1% relative); the paired interval for the difference excludes zero and the randomization test gives $p = 0.001$. It also exceeds Pure BM25 and Pure Dense by wide margins. In Japanese Lex-TM leads RRF by 0.018 and in Chinese it trails RRF by 0.030; neither difference is significant. In Chinese, Pure BM25 (0.7579) scores higher than BGE-M3 dense retrieval and, though not significantly, Lex-TM, for reasons discussed in Section 6.4.

Learned-sparse and late-interaction baselines. Two of the stronger baselines requested in review do better than Lex-TM. BGE-M3's multi-vector

	English	Japanese	Chinese
Lexical, learned-sparse and late-interaction baselines			
Pure BM25	0.6566	0.5855	0.7579
BGE-M3 learned sparse	0.5687	0.6935	0.5944
BGE-M3 multi-vector (late interaction)	0.8387	0.8036	0.7786
BGE-M3 dense + sparse + multi-vector	0.7475	0.7434	0.7158
SPLADE++	0.8631	–	–
Encoder: BGE-M3			
Pure Dense	0.6596 [0.52, 0.79]	0.6323 [0.50, 0.76]	0.6266 [0.49, 0.76]
BM25 + Dense RRF (k=60)	0.7081 [0.58, 0.83]	0.6883 [0.55, 0.82]	0.7716 [0.65, 0.88]
LDA routing (τ=1.1, symmetric prior)*	0.8147 [0.70, 0.92]	0.7062 [0.57, 0.83]	0.7415 [0.61, 0.86]
Lex-TM (τ=0.8)	0.8151 [0.70, 0.92]	0.7063 [0.57, 0.83]	0.7420 [0.62, 0.86]
Δ vs. RRF	+0.107 [+0.041, +0.181]	+0.018 [−0.063, +0.104]	−0.030 [−0.095, +0.031]
Δ vs. Pure BM25	+0.159 [+0.055, +0.270]	+0.121 [−0.005, +0.247]	−0.016 [−0.117, +0.083]
Δ vs. LDA routing	+0.000 [−0.038, +0.039]	+0.000 [+0.000, +0.000]	+0.001 [−0.002, +0.003]
Encoder: Qwen3-Embedding-0.6B			
Pure Dense	0.7436 [0.62, 0.85]	0.7949 [0.68, 0.90]	0.6849 [0.55, 0.81]
BM25 + Dense RRF (k=60)	0.7944 [0.67, 0.91]	0.7604 [0.63, 0.88]	0.7965 [0.68, 0.90]
Lex-TM (τ=0.8)	0.8515 [0.75, 0.94]	0.8175 [0.71, 0.92]	0.8241 [0.71, 0.92]
Δ vs. Pure Dense	+0.108 [+0.045, +0.178]	+0.023 [−0.021, +0.069]	+0.139 [+0.064, +0.229]
Δ vs. RRF	+0.057 [+0.012, +0.118]	+0.057 [−0.014, +0.143]	+0.028 [−0.051, +0.108]

Table 1: MedWeb label-set retrieval: MRR@20 over the 39 label sets of each partition; brackets give 95% bootstrap intervals. Δ rows give Lex-TM's mean per-query difference with a paired bootstrap interval (10,000 resamples). The first block does not depend on the encoder used with Lex-TM; SPLADE++ (checkpoint naver/splade-cocondenser-ensembledistil) is English-only. *Without entropy adaptation; Table 3 compares the two priors with all other settings equal.

mode, scored by late interaction, reaches 0.8387, 0.8036 and 0.7786, above Lex-TM in all three languages and significantly so in Japanese (Lex-TM minus multi-vector: −0.097 [−0.171, −0.035]), and SPLADE++ reaches 0.8631 in English; the English differences are not significant (Appendix C). Lex-TM does beat BGE-M3's learned-sparse mode significantly in English and Chinese, and its combined mode in English (+0.068 [+0.014, +0.137]). Late interaction and SPLADE++ therefore avoid most of the tokenizer trap on their own. What Lex-TM offers instead is a small index: one dense vector and one K-dimensional topic vector per document ($K = 20$ on MedWeb), where late interaction stores one vector per token.

Stronger encoder. With Qwen3-Embedding-0.6B, dense MRR rises to 0.7436, 0.7949 and 0.6849, and Lex-TM reaches 0.8515, 0.8175 and 0.8241: routing adds +0.108, +0.023 and +0.139 over dense retrieval, significantly in English and Chinese, and in English it also beats Qwen3-based RRF significantly (+0.057 [+0.012, +0.118]). The topic channel is identical under both encoders (it never sees the embeddings). Its gain shrinks where the encoder improves most (Japanese: +0.074 over BGE-M3, +0.023 over Qwen3), so part of what routing adds to BGE-M3 corrects errors that a better encoder avoids; it is largest in Chinese, where the Qwen3 dense model is weakest.

5.2 Beyond MedWeb

Table 2 reports the other collections, and Appendix F adds the learned-sparse and late-interaction baselines. Routing does not help outside MedWeb. On CmedqaRetrieval Lex-TM matches dense retrieval (0.3343 vs. 0.3372 nDCG@10). On ChatDoctor, where on average only 16% of a question's distinct content words occur in its answer, it lowers nDCG@10 from 0.5806 to 0.4933, and on FiQA-2018, our non-medical collection, from 0.4025 to 0.2887; the symmetric prior changes these results by less than 0.01. On long biomedical abstracts routing also hurts: Lex-TM falls below dense retrieval on TREC-COVID (0.4069 vs. 0.5699) and NFCorpus (0.2575 vs. 0.2900). A likely cause is that long abstracts mixing several topics yield diffuse topic distributions that add noise to the dense score; these are earlier-pipeline runs, and we did not test this explanation. On the open-domain sets the sparsity gate hands short factoid queries to the dense model, and Lex-TM matches dense retrieval on MIRACL-zh and Mr. TyDi-ja. On Thai it falls below the dense baseline; Thai lacks a segmenter comparable to MeCab or jieba, and we did not rerun these earlier-pipeline

Collection	Type	Metric	BM25	Dense	RRF	Lex-TM	Δ vs. Dense
ChatDoctor (en)	medical Q&A	nDCG@10	0.3325	0.5806	0.4967	0.4933	−0.087
CmedqaRetrieval (zh)	medical QA	nDCG@10	0.1568	0.3372	0.2595	0.3343	−0.003
FiQA-2018 (en)	finance QA	nDCG@10	0.2122	0.4025	0.3374	0.2887	−0.114
TREC-COVID (en)‡	biomedical	nDCG@10	0.4485	0.5699	0.6603	0.4069	−0.163
NFCorpus (en)‡	biomedical	nDCG@10	0.2661	0.2900	0.3079	0.2575	−0.033
MIRACL (zh)‡	open domain	MRR@20*	0.2035	0.4474	0.3240†	0.4479	+0.001
Mr. TyDi (ja)‡	open domain	MRR@20*	0.4003	0.7724	0.5896†	0.7726	+0.000
Mr. TyDi (th)‡	open domain	MRR@20*	0.4929	0.8259	0.6774†	0.7811	−0.045

Table 2: Collections beyond MedWeb (BGE-M3, $\gamma = 0.7$). ‡Earlier-pipeline runs, not repeated for this version (Appendix G); they use a sparsity gate ($\delta = 5$ on the open-domain sets, $\delta = 3$ on TREC-COVID and NFCorpus). *Each relevant passage is scored as a separate query–passage pair. †RRF with $k = 600$ (earlier baseline script; all other RRF results use $k = 60$).

Configuration	EN	JA	ZH
Lex-TM (final: γ=0.7, τ=0.8)	0.8151	0.7063	0.7420
without entropy adaptation	0.8142	0.7063	0.7420
… and τ=0.5	0.8156	0.6938	0.7409
… and τ=1.1 (no HHI)	0.8147	0.7062	0.7415
with query smoothing 0.1*	0.7837	0.6980	0.7241
γ=1.0 (Pure Dense)	0.6596	0.6323	0.6266
γ=0.9	0.7243	0.6797	0.6835
γ=0.8	0.7376	0.6900	0.7173
γ=0.7	0.8142	0.7063	0.7420
γ=0.6	0.8260	0.7106	0.7580
γ=0.5	0.8229	0.7136	0.8000
γ=0 (topic only)	0.7824	0.5597	0.7708

Table 3: Ablations on MedWeb (BGE-M3, MRR@20). The γ sweep is without entropy adaptation at $\tau = 0.8$. *Earlier diagnostic run with entropy adaptation; smoothing blends θ_q with the uniform distribution and is off in the final configuration.

results to analyze the gap.

Efficiency. Routing needs no GPU training: the topic model is fitted by a Gibbs sampler on the CPU. In the earlier pipeline, the 100k-passage MIRACL subset took three hours end to end, of which the routing step took three minutes after embedding, and a 187,176-document airline-review collection (Appendix H) took 24 hours, of which 19 were spent on dense embeddings and topic distributions.

6 When Does Routing Help?

6.1 Hyperparameters

Table 3 separates the components. The interpolation weight matters most: English MRR rises from 0.6596 with dense retrieval alone to 0.7243 at $\gamma = 0.9$ and 0.8142 at $\gamma = 0.7$, and it peaks at $\gamma = 0.6$ (0.8260); in Chinese, $\gamma = 0.5$ reaches 0.8000, above RRF, although neither difference is significant. We keep $\gamma = 0.7$, which was fixed before these sweeps: choosing γ on the same label sets would overstate the gain. The exclusivity threshold barely matters: across $\tau \in \{0.5, 0.8, 1.1\}$, English and Chinese MRR change by at most 0.0014 and Japanese by 0.0124, and switching the prior off entirely ($\tau = 1.1$) matches the final configuration to within 0.0005 on every partition with seed 42. Entropy adaptation changes MRR by at most 0.0009 with seed 42. All three effects are smaller than the variation across Gibbs seeds: over 5 seeds, Lex-TM's MRR varies more than it does across τ in every language (by up to 0.0147, in Japanese), and averaged over seeds the HHI prior is no better than the symmetric one (mean difference without adaptation +0.0005 in English, −0.0066 in Japanese and +0.0004 in Chinese; standard deviations 0.0020, 0.0063 and 0.0013). Smoothing the query's topic distribution towards uniform, used in an earlier configuration, lowered MRR in all three languages and is off in the final configuration.

6.2 Why the Prior Is Inert on MedWeb

MedWeb documents are short enough that terms rarely repeat within a document: 93.5%, 92.5% and 91.8% of term–document pairs have a count of one (EN, JA, ZH). HHI is therefore essentially $1/\mathrm{df}$ (Section 3.1); its Spearman correlation with document frequency is -0.9994 (EN, JA) and -0.9996 (ZH). At $\tau = 0.8$ the amplified terms are exactly those that occur in two documents (255 of 909 English terms, 285 of 921 Japanese, 307 of 921 Chinese); at $\tau = 0.5$ they are the terms that occur in two or three documents and a handful that occur in four. None of the 11–13 distinct query terms per language is amplified: the rarest occurs in 130 (EN), 241 (JA) and 280 (ZH) documents, and query terms have $E_v \leq 0.013$. Amplifying the

prior mass of the rarest terms reorders the lower-ranked words of a topic but keeps the label words that head it: matched by their ten top words, the topics with and without the prior share 7.5 (English), 7.7 (Japanese) and 7.0 (Chinese) of them on average (Table 4). The gain in Table 1 thus comes from routing through topics that recover MedWeb's label structure, in English and Chinese largely from the label strings themselves (Section 6.3), and not from the HHI prior. On the three new collections, whose documents are longer, the prior did not help either, and on ChatDoctor it slightly hurt (Appendix F).

6.3 Query Types and Document Structure

Query types. In English, the gain over RRF comes from label sets with two or more labels (English MRR 0.897 vs. 0.737 with two labels, 0.729 vs. 0.613 with three to five) and from sets with fewer than 100 relevant tweets; on the English single-label sets and the nine largest sets, dense retrieval, RRF and Lex-TM score the same (Appendix D). Because MRR weights every label set equally, the 21 sets with fewer than ten relevant tweets carry more than half of the metric. On the open-domain sets, short factoid queries are routed to the dense model by the sparsity gate; without the gate, routing lowered MRR on all three (MIRACL-zh 0.4246, Mr. TyDi-ja 0.7643, -th 0.7513, against dense 0.4474, 0.7724 and 0.8259).

Label strings. Removing the label strings from the indexed MedWeb text changes Lex-TM from 0.8151 to 0.3808 in English, against 0.3723 for dense retrieval and 0.2017 for RRF (Appendix E). Without the label strings, routing no longer improves on dense retrieval in English (+0.008) or Chinese (+0.002), and the same holds with Qwen3-Embedding. Japanese is the exception: the difference, +0.108, is larger than with the label strings, but it misses significance ($p = 0.060$). In English and Chinese, then, the topics recover MedWeb's label structure because the labels are written into the documents. Without them, routing still beats RRF significantly in English and Japanese, because BM25 collapses (to 0.1398 in English).

Document structure. Routing helps on MedWeb, whose documents are short and fall into discrete categories (7.8–9.2 content tokens per document, 39 label sets). It hurts on TREC-COVID and NFCorpus, whose documents are long (92 and 132 content tokens) and mix topics, and on two question–answer collections, ChatDoctor and FiQA; on CmedqaRetrieval it matches dense retrieval. These observations are consistent with four conditions raised during review, which we adopt as Lex-TM's operating envelope: a specialized domain; heavy lexical overlap with a few distinguishing terms; documents that each concentrate on one topic; and documents long enough for topic inference but short enough not to mix topics. The label-free control suggests a fifth: in English and Chinese, the gain depends on category words that appear verbatim in the documents. Our experiments do not test the conditions separately.

6.4 Multilingual Differences

Four factors separate the partitions. *Tokenization of the distinguishing terms.* Our preprocessing drops single-character CJK tokens. Every Chinese symptom name is a multi-character jieba word (花粉症 "hay fever", 鼻涕 "runny nose", 发烧 "fever"), but MeCab returns the Japanese 熱 (fever) as one character and splits 花粉症 into 花粉+症. Fever therefore disappears from Japanese queries and label strings in the lexical and topic channels (tweets keep compounds such as 微熱 and 高熱), the cough label 咳・たん is reduced to たん (phlegm), and 24 of the 39 Japanese label strings share their bag of words with another label string (27 distinct bags of words, against 39 in English and Chinese). This caps what BM25 and the topic channel can do in Japanese, and Japanese BM25 is also the most sensitive to how ties are broken (0.5447–0.6347 MRR, depending on the order of equal scores). *Strength of the lexical baseline.* Intact label words make Chinese label strings easy to match exactly, so Pure BM25 is stronger in Chinese (0.7579) than in English and Japanese (0.6566, 0.5855); RRF inherits this, while Lex-TM does not use BM25, so with BGE-M3 fusion stays ahead in Chinese. *Dense encoder quality.* BGE-M3's dense MRR is lowest in Chinese and Japanese; with Qwen3-Embedding the same topic channel lifts Chinese most. *Document length.* Chinese documents are the shortest (7.8 content tokens against 8.4 and 9.2), which leaves the least evidence for topic inference. Taken together, the Chinese result with BGE-M3 appears to reflect the base encoder and an unusually strong lexical baseline more than document length, since with Qwen3-Embedding routing gains most in Chinese; the Japanese result is limited by tokenization.

Topic	**LDA routing** ($\tau = 1.1$, symmetric prior)	**Lex-TM** ($\tau = 0.8$)
EN, *diarrhea*	diarrhea, patient, symptom, stomach, morning, hurt, since, might	diarrhea, patient, symptom, stomach, hurt, work, lately, morning
EN, *no symptom*	health, status, routine, flu, shot, bird, season, get	status, health, routine, flu, get, shot, season, going
JA, *diarrhea*	下痢 (diarrhea), 症状 (symptom), 臨床 (clinical), お腹 (stomach), 最近 (lately), 痛い (hurts)	下痢 (diarrhea), 臨床 (clinical), 症状 (symptom), お腹 (stomach), やばい (awful), 痛い (hurts)
ZH, *hay fever*	花粉症 (hay fever), 症状 (symptom), 鼻涕 (runny nose), 特定 (specific), 临床 (clinical), 春天 (spring)	花粉症 (hay fever), 症状 (symptom), 特定 (specific), 春天 (spring), 好像 (seems), 明年 (next year)

Table 4: Top words of matching topics without and with the HHI prior ($K = 20$, seed 42); for each topic of LDA routing, the Lex-TM topic shown is the one that shares most of its ten top words. Topics follow MedWeb's label sets, and several top words come from the label strings themselves (*patient, symptom*; *routine, health, status*; 臨床/症状; 临床/症状).

7 Conclusion

Lex-TM adds a topic channel to a frozen dense retriever without any neural training. On MedWeb's label-set retrieval task it improves English MRR over BM25+dense fusion significantly, and the differences in Japanese and Chinese are not significant; with Qwen3-Embedding, routing still improves over dense retrieval, significantly in English and Chinese. Three findings bound these gains. Two retrievers that keep term-level evidence, BGE-M3's multi-vector mode and SPLADE++, match or exceed Lex-TM. In English and Chinese the topics recover MedWeb's label structure from the label strings in the documents, and without them the gain over dense retrieval disappears. And routing did not transfer: it matched or trailed dense retrieval on every other collection. The HHI prior, which on short texts reduces to a rarity weight, brings no benefit. Future work includes relevance judgments beyond label sets, free-text queries against unlabeled records, phrase-aware priors, choosing γ per query from the query's own statistics, and topic routing on top of late interaction.

Limitations

Label-set relevance. MedWeb queries are label strings, every relevant document contains its query, and relevance is defined by identical label sets rather than by clinical judgment. Our MedWeb results measure discrimination between overlapping label sets; performance with free-text queries against unlabeled records is untested. **Label strings.** In English and Chinese the gain over dense retrieval disappears when the label strings are removed from the documents. **Stronger retrievers.** BGE-M3's multi-vector mode and SPLADE++ match or exceed Lex-TM on MedWeb; Lex-TM's advantage over them is index size, not accuracy. **Small query sets.** Each MedWeb partition yields 39 queries, so confidence intervals are wide, and with the label strings in the documents the improvement over fusion is significant only in English. **The prior.** On MedWeb the HHI prior reduces to a rarity weight and has no measurable effect; we have not shown a benefit from it on any collection. **Operating envelope.** Routing needs short, single-topic documents from a specialized domain with heavy lexical overlap; outside that regime it matched or trailed dense retrieval in our experiments. **Bag of words.** Topic routing cannot represent multi-word anchors such as *maculopapular rash*; abbreviations mitigate this only partly, and phrase-aware priors are future work. **Tokenization.** Results depend on the segmenter and on the length filter: in Japanese it removes the symptom name 熱 (fever) and most of the cough label, and Thai lacks a segmenter of comparable maturity. **Tuning and seeds.** $\gamma = 0.7$ was fixed after sweeps on all MedWeb label sets and is used on every collection; τ was chosen on seven validation label sets, which the reported MedWeb results include. The main results use one Gibbs seed; Section 6.1 repeats MedWeb with five. **Encoders.** We test BGE-M3 and Qwen3-Embedding-0.6B; larger embedders and domain-adapted dense retrievers are untested. **Static vocabulary.** The prior and topics are computed on a fixed collection and must be refitted as documents are added. **Earlier pipeline.** The TREC-COVID, NFCorpus, MIRACL and Mr. TyDi results come from an earlier version of the pipeline (Appendix G).

Ethical Considerations

MedWeb consists of pseudo-tweets written for the shared task and contains no personal health information; the other collections are public research benchmarks used under their terms. Retrieval errors in medical applications can mislead clinicians and patients. Lex-TM is a retrieval component for research: we have not evaluated it with clinical users or linked it to patient outcomes, and any medical use of retrieval-augmented generation requires review by qualified professionals.

Acknowledgments

We thank the area chair and the three reviewers for their comments, which led to the stronger baselines, additional collections and analyses in this version. This study was supported by the National Science and Technology Council of Taiwan under grants NSTC 114-2410-H-038-034-MY3, NSTC 113-2627-M-A49-002, and NSTC 114-2221-E-004-001-MY2.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 759–766.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.

Gordon V Cormack, Charles LA 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*, pages 758–759.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

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)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model

Meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6):e40895.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9844–9855.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM)*, pages 623–632.

Maosong Sun, Jingyang Li, Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, and Zhiyuan Liu. 2016. THUCTC: An efficient Chinese text classifier. http://thuctc.thunlp.org/.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13: MedWeb task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems 22*, pages 1973–1981.

Canhui Wang, Min Zhang, Liyun Ru, and Shaoping Ma. 2008. Automatic online news topic ranking using media focus and user attention based on aging theory. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1033–1042.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for Chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Implementation Details

Code, preprocessing scripts, seeds, per-query outputs and a changelog of every issue we found and fixed are available at https://github.com/mnsh0409/Lex-TM. Table 5 lists the settings. Topic inference uses the one-step fold-in of Eq. (4) for documents and queries alike; Gibbs sampling runs for at most 500 sweeps; from sweep 50 on, it would stop early if the log-likelihood changed by less than 10^{-4} between checks 50 sweeps apart, which never happened on MedWeb, so every MedWeb model uses 500 sweeps. Documents longer than the chunk size are split into overlapping chunks and scored per chunk; every MedWeb document is a single chunk. Algorithm 1 summarizes the pipeline.

Setting	Value
Topics K	20 (MedWeb), 50 (others)
α, β_{base}, λ	0.1, 0.01, 2.0
τ	0.8
γ; adaptation	0.7; entropy, max. 0.2
Gate δ	5 (open-domain sets only)
Gibbs sweeps; seed	500 (max.); 42 (43–46 in §6.1)
Vocabulary	min. document frequency 2
Chunk size	512 tokens
Dense encoders	BGE-M3; Qwen3-Emb.-0.6B
RRF k	60
Bootstrap resamples	10,000 (CIs); 10,000 (paired)

Table 5: Hyperparameters of the current pipeline. The earlier-pipeline runs differ as described in Appendix G.

Algorithm 1 Lex-TM routing

Require: collection $\mathcal{D}$, query q, weight γ, gate δ

1: compute E_v and $\vec{\beta}$ (Eqs. 1–2); fit LDA by Gibbs sampling
2: $\theta_d \leftarrow$ fold-in (Eq. 4); $\mathbf{e}_d \leftarrow$ encoder, for all d
3: $\theta_q \leftarrow$ fold-in; $\mathbf{e}_q \leftarrow$ encoder
4: $\gamma_q \leftarrow 1$ if q has $< \delta$ in-vocabulary terms, else entropy-adapted γ
5: rank $\mathcal{D}$ by $s(q, d)$ (Eq. 5)

B Choosing τ

The review version split MedWeb's 39 label sets into 32 test and 7 validation sets. Summed over the three partitions, validation MRR favored $\tau = 0.8$ (2.396 against 2.347 for $\tau = 0.5$; Table 6), although $\tau = 0.5$ scored higher on the Japanese and Chinese validation sets. That version's headline English result (0.8312 against RRF 0.7276 on the 32 test sets) used $\tau = 0.5$; with $\tau = 0.8$ its English test-set MRR was 0.8024. This version uses $\tau = 0.8$ for every partition and reports all 39 label sets. With seven validation sets the choice is noisy, and here it is immaterial: on all 39 label sets, τ changes MRR by at most 0.0014 in English and Chinese and 0.0124 in Japanese (Table 3).

Partition	$\tau = 0.5$	$\tau = 0.8$
English	0.7440	0.8681
Japanese	0.7347	0.7262
Chinese	0.8681	0.8016
Sum	2.347	2.396

Table 6: Validation MRR on seven label sets.

C Paired Tests and Learned-Sparse Baselines

Table 7 reports Lex-TM's mean per-query difference from each system on MedWeb with a paired bootstrap interval and the p-value of a paired randomization test.

D Query-Type Breakdown

Table 8 splits the English label sets by the number of relevant tweets and by the number of labels in the set.

E Label-Free Control

Table 9 repeats the MedWeb evaluation with the label strings removed from the indexed documents; queries and relevance are unchanged. Topic models are refitted on the label-free text. One English tweet ("Flu crisis.") is too short to survive chunking without its label string, so the English label-free index has 2,559 documents.

F Full Results on the New Collections

Table 10 lists every system on the three collections run for this version. The multi-vector and combined BGE-M3 modes re-score the union of the top 100 dense and top 100 sparse candidates, and were not run on CmedqaRetrieval; SPLADE++ is English-only.

G Earlier-Pipeline Results

The TREC-COVID and NFCorpus results in Table 2 come from the submitted version: 150-token chunks, $K = 50$, 300 Gibbs sweeps, a sparsity gate with $\delta = 3$, the top 100 chunks max-pooled per document and scored with the official BEIR nDCG@10. Their BM25 baselines tokenized queries (and, for NFCorpus, documents) by whitespace only, so the BM25 column is a weak baseline,

	English		Japanese		Chinese	
Lex-TM vs.	Δ [95% CI]	p	Δ [95% CI]	p	Δ [95% CI]	p
Encoder: BGE-M3						
RRF	+0.107 [+0.041, +0.181]	0.001	+0.018 [−0.063, +0.104]	0.671	−0.030 [−0.095, +0.031]	0.438
Pure BM25	+0.159 [+0.055, +0.270]	0.007	+0.121 [−0.005, +0.247]	0.070	−0.016 [−0.117, +0.083]	0.759
Pure Dense	+0.155 [+0.076, +0.248]	<0.001	+0.074 [+0.029, +0.128]	0.002	+0.115 [+0.053, +0.189]	<0.001
LDA routing	+0.000 [−0.038, +0.039]	0.966	+0.000 [+0.000, +0.000]	1.000	+0.001 [−0.002, +0.003]	0.751
BGE-M3 sparse	+0.246 [+0.138, +0.359]	<0.001	+0.013 [−0.083, +0.108]	0.805	+0.148 [+0.045, +0.251]	0.007
BGE-M3 multi-vector	−0.024 [−0.089, +0.034]	0.529	−0.097 [−0.171, −0.035]	0.004	−0.037 [−0.081, −0.001]	0.112
BGE-M3 all modes	+0.068 [+0.014, +0.137]	0.037	−0.037 [−0.080, −0.004]	0.077	+0.026 [−0.005, +0.067]	0.240
SPLADE++	−0.048 [−0.170, +0.069]	0.460	–	–	–	–
Encoder: Qwen3-Embedding-0.6B						
Pure Dense	+0.108 [+0.045, +0.178]	0.002	+0.023 [−0.021, +0.069]	0.346	+0.139 [+0.064, +0.229]	<0.001
RRF	+0.057 [+0.012, +0.118]	0.034	+0.057 [−0.014, +0.143]	0.178	+0.028 [−0.051, +0.108]	0.503

Table 7: Paired comparisons on MedWeb (MRR@20): Lex-TM's mean per-query difference from each system, with a paired bootstrap 95% interval (10,000 resamples) and the p-value of a two-sided paired randomization test (10,000 sign flips). The BGE-M3 sparse, multi-vector and SPLADE++ baselines are compared with Lex-TM on BGE-M3.

English label sets	n	**Dense**	**RRF**	**Lex-TM**
≥100 relevant	9	1.000	1.000	1.000
10–99 relevant	9	0.889	0.861	1.000
<10 relevant	21	0.416	0.517	0.657
0–1 labels	8	0.875	0.875	0.875
2 labels	13	0.716	0.737	0.897
3–5 labels	18	0.523	0.613	0.729

Table 8: MRR@20 by query type (English, BGE-M3). The single no-symptom set is counted with the one-label sets.

Label-free MedWeb	**EN**	**JA**	**ZH**
Pure BM25	0.1398	0.1906	0.2435
BGE-M3 learned sparse	0.1564	0.4684	0.4264
BGE-M3 multi-vector	0.3445	0.5214	0.4463
BGE-M3 all modes	0.3673	0.4811	0.4604
SPLADE++	0.3818	–	–
Encoder: BGE-M3			
Pure Dense	0.3723	0.4089	0.4416
RRF	0.2017	0.3181	0.3500
LDA routing	0.3689	0.5182	0.4817
Lex-TM	0.3808	0.5166	0.4439
Δ vs. Pure Dense	+0.008	+0.108	+0.002
p	0.871	0.060	0.967
Encoder: Qwen3-Embedding-0.6B			
Pure Dense	0.4269	0.4809	0.4245
Lex-TM	0.4490	0.5390	0.4421
Δ vs. Pure Dense	+0.022	+0.058	+0.018
p	0.662	0.242	0.716

Table 9: MRR@20 without label strings in the documents. Δ is Lex-TM's mean per-query difference from dense retrieval with the same encoder; p is from the paired randomization test.

and we draw no conclusion from it. The MIRACL and Mr. TyDi results use the sparsity gate ($\delta = 5$) and pair-level MRR, and their RRF column was computed by an earlier baseline script with $k = 600$. All earlier-pipeline runs use $\tau = 0.5$ and no entropy adaptation. These runs predate several fixes documented in the repository's changelog; we did not rerun them, and they serve only to show where routing does not help.

H News and Review Collections

Table 11 lists four title-query collections from the earlier pipeline, in which each document's title is a query and the documents that share it are relevant. On THUCNews all systems approach ceiling performance and routing lowers MRR slightly (0.9801 against 0.9963 for dense retrieval); on Sogou it lowers MRR by 0.050 relative to dense retrieval and by 0.098 relative to BM25. On airline passenger reviews in English (187,176 reviews) and Chinese (2,663), routing lowers MRR well below dense retrieval, with or without the HHI prior ($\tau = 1.1$: 0.0896 and 0.4007). These collections lie outside the operating envelope of Section 6.3; BM25 and RRF were not run on the review collections.

nDCG@10	ChatDoctor	Cmedqa	FiQA
Pure BM25	0.3325	0.1568	0.2122
BGE-M3 dense	0.5806	0.3372	0.4025
RRF (k=60)	0.4967	0.2595	0.3374
LDA routing	0.5008	0.3354	0.2880
Lex-TM	0.4933	0.3343	0.2887
BGE-M3 learned sparse	0.4463	0.2440	0.2621
BGE-M3 multi-vector	0.5278	–	0.4094
BGE-M3 all modes	0.5856	–	0.4249
SPLADE++	0.4764	–	0.3445
Queries	5,583	3,999	648
Documents	5,545	100,001	57,600

Table 10: nDCG@10 on the collections run for this version ($K = 50$, $\tau = 0.8$, $\gamma = 0.7$).

Dataset	BM25	Dense	RRF†	Lex-TM
THUCNews (zh)	0.9974	0.9963	0.9985	0.9801
Sogou (zh)	0.9490	0.9004	0.9267	0.8506
Airline reviews (en)	–	0.1917	–	0.0893
Airline reviews (zh)	–	0.5902	–	0.3904

Table 11: MRR on news (Sun et al., 2016; Wang et al., 2008) and review collections (earlier pipeline, BGE-M3, $\tau = 0.5$; †RRF with $k = 600$).