

VocMoE: Vocabulary-Guided Mixture-of-Experts for Korean Medical Language Models

Sangyeon Cho^{1,2}, Jaeho Han¹, Mingyu Jeon¹, Junyeong Kim¹,

¹Department of Artificial Intelligence, Chung-Ang University, Republic of Korea,

²Korean Surgical Researcher Foundation, Republic of Korea

Correspondence: Junyeong Kim (junyeongkim@cau.ac.kr)

Abstract

In Korean medical LLMs, domain-trained models underperform General LLMs of comparable parameter scale. We attribute this *paradoxical failure* to a *cascading two-bottleneck* structure: tokenizer over-fragmentation (*Fragmentation bottleneck, B1*) compounded by a learning-stage representation-alignment failure (*Alignment bottleneck, B2*). Using the Korean medical domain as a case study, we diagnose this paradox and propose **VocMoE**, which is designed to address both bottlenecks: DVE selects domain vocabulary from a medical BERT via an information-theoretic score, and MoE uses the DVE output (the partition of token ids) as the vocabulary-conditioned, deterministic routing key of the MoE FFN. In this view, the vocabulary partition itself serves as a routing prior, shifting routing from hidden-state learning to vocabulary identity. MediKo-9B-A4B reaches simultaneous SOTA on KorMedMCQA (57.5, +14.8 pt), KMMLU Medical, and HAE-RAE, while MediKo-4B-A1B (active 1.69B) surpasses Qwen3-8B (8B) and KULLM3 (10.7B) on KorMedMCQA, updating the active-parameter efficiency frontier.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance on general natural-language understanding, and this success has driven the development of Domain LLMs specialized for professional domains such as medicine (PMC-LLaMA, Wu et al., 2024; Meditron, Chen et al., 2023; BioMistral, Labrak et al., 2024), science (Innovator; Liao et al., 2025), law (SaulLM; Colombo et al., 2024), and finance (BloombergGPT; Wu et al., 2023). These models typically inject domain knowledge via Continual Pre-Training (CPT) on domain-specific corpus and have achieved meaningful gains on English-language benchmarks.

However, a *paradoxical pattern* emerges in Korean medical LLM adaptation, where domain-fine-

tuned models underperform General LLMs of comparable parameter scale despite undergoing explicit domain-specific training. On Korean medical-token PPL and Top-1 (Table 1), English, multilingual, and Korean Medical LLMs all trail General LLMs of comparable parameter scale. This pattern is not explained by a simple data-quantity gap; rather, it points to a fundamental representational limitation that domain training alone does not resolve.

We attribute this limitation to a representational deficiency of non-English medical vocabulary, originating from a two-stage cascading bottleneck: (i) the tokenizer fragments domain vocabulary into excessive sub-words, and (ii) representations of these fragmented tokens fail to align sufficiently during training, eroding semantic precision. We use the Korean medical domain as a case study to diagnose this quantitatively and propose a framework potentially applicable to other non-English medical domains.

Fragmentation bottleneck (B1). Major LLM tokenizers fragment Korean medical terms roughly 2.9–4.6× more than Korean general words (Table 2a), with zero single-token representations observed across all five tokenizers. This over-fragmentation raises training and inference cost and induces *representational confusion* (semantically unrelated tokens partially sharing representations) — a phenomenon previously reported for byte-level fallback in Chinese-LLaMA (Cui et al., 2023) and Bllossom (Choi et al., 2024) and extending to BPE-based fragmentation.

Alignment bottleneck (B2). To test whether fragmentation affects representation quality, we extract Qwen3-8B contextual embeddings of the same medical term across English, French, German, Chinese, and Korean, computing the cosine similarity between each non-English embedding and its English counterpart (Table 2b). Korean—with highest fragmentation rate—shows lowest similarity, and similarity decreases monotonically with frag-

Group	Model	Med PPL↓	Med Top-1↑	Δ
Gen-Multi	Qwen3-8B (Yang et al., 2025)	5.41	62.9%	ref
Gen-Ko	KULLM3 (NLP & AI Lab, Korea University, 2024)	4.27	63.1%	+0.2
En-Med	OpenBioLLM-8B (Pal and Sankarasubbu, 2024)	8.68	54.5%	−8.4
Multi-Med	MMed-Llama3-8B (Qiu et al., 2024)	38.58	37.9%	−25.0
Ko-Med	Kor-Med-LLaMA3-8B (ih9511, 2024)	6.95	57.7%	−5.2

Table 1: Korean medical-token PPL and next-token Top-1 (DVE token set, 8,374 terms). Best model per category; Δ = Top-1 gap vs. Qwen3-8B. Full 14-baseline numbers in App. Table A15.

(a)				(b)						
Tokenizer	Med (Ko)	Gen (Ko)	Ratio	Domain	Category	EN-FR	EN-DE	EN-ZH	EN-KO	KO Frag
Qwen3	6.5	1.7	3.9×	General	—	0.788	0.709	0.611	0.494	1.3
LLaMA-3	6.8	1.5	4.6×							
Mistral	9.6	3.3	2.9×	Medical	Low frag (~ 2)	0.708	0.679	0.692	0.382	2.2
Gemma-2	6.4	1.7	3.7×		Mid frag (~ 4)	0.572	0.549	0.574	0.363	4.5
GPT-OSS	5.7	1.6	3.6×		High frag (~ 7)	0.612	0.639	0.514	0.344	7.3
					All	0.630	0.622	0.593	0.363	4.7

Table 2: **Bottleneck evidence.** (a) Fragmentation of Korean medical vs. general words across five LLM tokenizers (B1). (b) Qwen3-8B cross-lingual cosine to English, medical terms binned by Korean fragmentation ($\sim 2/4/7$ sub-words). EN-KO drops $0.494 \rightarrow 0.344$ monotonically with fragmentation while other languages stay ~ 0.6 (cascading B1 $\rightarrow$ B2).

mentation. Bottleneck 2 is thus not independent of Bottleneck 1 but cascadingly constrained by it.

Vocabulary expansion alone is insufficient. A direct approach is *Vocabulary Expansion* (Llama-2-Ko, Lee, 2024; EEEV, Kim et al., 2024b; AdaptiVocab, Nakash et al., 2025; VEGAD, Liu et al., 2024; BIllossom, Choi et al., 2024), which mitigates over-fragmentation by representing domain terms with appropriate sub-word granularity. However, most methods merge external vocabularies through simple set unions without explicit linguistic or domain specificity, and process expanded and existing tokens through the same learning pathway—causing knowledge to be bound onto unstable representations during training. Bottleneck 1 is thus alleviated but Bottleneck 2 re-emerges, requiring two conditions: (i) effective selection and expansion of domain vocabulary, and (ii) separation of learning pathways.

We propose **VocMoE** (Vocabulary-Guided Mixture-of-Experts), a framework designed to jointly address the two requirements identified. The first component, **DVE**, selects domain vocabulary candidates from KM-BERT (Kim et al., 2022) using an information-theoretic scoring function (Eq. 1), reducing the mean fragmentation of domain tokens from 4.5 to 1.6 (a 64% reduction). The second component, **MoDE**, introduces an FFN architecture that architecturally separates the forward paths of expanded and existing tokens. Instead of using a learned router, MoDE directly uses the

DVE-defined token-id partition as the routing key. We instantiate VocMoE in the Korean medical domain through the **MediKo** family, trained at three base scales: 1.1B-A0.6B, 4B-A1B, and 9B-A4B.

Our contributions are three-fold: **(1) Cascading two-bottleneck analysis.** We provide evidence that non-English medical adaptation failure can arise from a cascading interaction between tokenizer over-fragmentation and learning-stage representation misalignment. **(2) The VocMoE framework.** We propose vocabulary-conditioned MoE routing, where vocabulary ids directly determine routing paths, enabling architectural isolation of general-token representations. **(3) Simultaneous medical/general SOTA.** MediKo achieves state-of-the-art performance on both Korean medical and general benchmarks, overcoming the conventional domain-up / general-down trade-off.

2 Related Work

2.1 Domain Vocabulary Expansion

Continual Pre-Training (CPT) for domain adaptation (Gururangan et al., 2020; Ke et al., 2023) has produced specialized LLMs across diverse fields, including medicine (PMC-LLaMA, Wu et al., 2024; Meditron, Chen et al., 2023; BioMistral, Labrak et al., 2024), law (SaulLM; Colombo et al., 2024), and finance (BloombergGPT; Wu et al., 2023). For non-English domains, multiple studies have identified excessive tokenizer fragmentation as a primary limitation (Chinese-LLaMA,

Cui et al., 2023; Blossom, Choi et al., 2024), and Llama-2-Ko (Lee, 2024), EEVE (Kim et al., 2024b), AdaptiVocab (Nakash et al., 2025), and VEGAD (Liu et al., 2024) have attempted to mitigate it through *Vocabulary Expansion*. However, these approaches merge external vocabularies as simple set unions without explicit linguistic or domain specificity, and more critically, they process expanded and existing tokens through the same learning path, leaving the cascading training-stage deficit unresolved. Our **DVE** selects new tokens from the KM-BERT (Kim et al., 2022) medical vocabulary via an information-theoretic frequency-based score that blocks general-vocabulary noise, and uses its output (the token-id partition) as the routing key of MoDE, jointly addressing the cascading bottleneck at both the vocabulary and architectural levels. Full tokenizer retraining and vocabulary reallocation must re-establish alignment with the base weights from scratch, which makes a fair comparison under an identical CPT budget difficult, while domain adapters operate on an axis orthogonal to FFN path separation. Our experiments instead compare an expansion-only condition and a learned-routing condition under identical settings (§4.3, §4.4).

2.2 MoE for Domain Adaptation

Sparse Mixture-of-Experts (MoE) architectures were introduced by Switch Transformers (Fedus et al., 2022) and later advanced through Mixtral (Jiang et al., 2024), DeepSeekMoE (Dai et al., 2024), and Qwen3-MoE (Yang et al., 2025). Existing MoE-based domain adaptation approaches largely follow two directions. The first employs learned domain-aware routing to allocate domain tokens to specialized experts (Innovator Liao et al., 2025). The second uses modality-aware routing to mitigate multimodal task interference (MoME Shen et al., 2024, EvoMoE Wang et al., 2025, and SMAR Xia et al., 2025). Despite their differences, both approaches rely on learned routing, which can temporarily disrupt the base model’s representational alignment during early training. In contrast, we shift routing decisions from the learning domain to the vocabulary domain. Specifically, the DVE-defined token-id partition serves as a deterministic routing key, keeping the general-token forward path architecturally isolated throughout training. Unlike prior work that primarily uses MoE for active-sparsity optimization, VocMoE adopts the MoE structure for domain-knowledge adapta-

tion.

3 Method

The central design principle of VocMoE is to shift routing from learned hidden-state classifiers to the vocabulary partition itself. VocMoE consists of two components: **DVE** directly addresses Bottleneck 1 (tokenizer-level over-fragmentation) through vocabulary expansion, and **MoDE** addresses Bottleneck 2 (architecture-level learning-stage misalignment) by separating forward paths at the FFN level. Figure 1 illustrates the overall architecture (a), the MoDE FFN block (b), and the DVE scoring distribution (c).

3.1 Domain Vocabulary Expansion (DVE)

DVE is a vocabulary expansion procedure designed to directly mitigate the over-fragmentation imposed by the base tokenizer on Korean medical terminology. Candidate domain tokens are extracted from KmBERT (Kim et al., 2022), a Korean medical BERT model trained with BPE on a Korean medical corpus. After removing tokens already included in the base tokenizer (Qwen3), 14,090 candidate tokens remain.

To quantify the expansion value of each candidate token t , we define the following information-theoretic scoring function:

$$\text{Score}(t) = \text{Frag}(t) \cdot \log(1 + \text{Freq}(t)) \cdot \max(0, \text{PMI}(t)), \quad (1)$$

where $\text{Frag}(t)$ is the number of sub-words the base tokenizer decomposes t into, $\text{Freq}(t)$ is the medical-corpus frequency, and $\text{PMI}(t)$ is the pointwise mutual information (Church and Hanks, 1990) between the medical corpus and Korean Wikipedia; the three terms encode fragmentation cost, learnability, and domain specificity, respectively.

Applying this scoring, the top- $K = 5,000$ of the 8,374 candidates with $\text{Score} > 0$ are added to the tokenizer, reducing the mean fragmentation of candidate tokens from 4.5 to 1.6 (a **64% reduction**); the embeddings of newly added tokens are introduced as learnable parameters and initialized via *Mean initialization* for stable training.

3.2 Mixture-of-Domain-Experts (MoDE)

Even with vocabulary expansion, as long as the training path remains unified, the representation alignment failure of Bottleneck 2 re-emerges at the training stage. Standard MoE upcycling routes

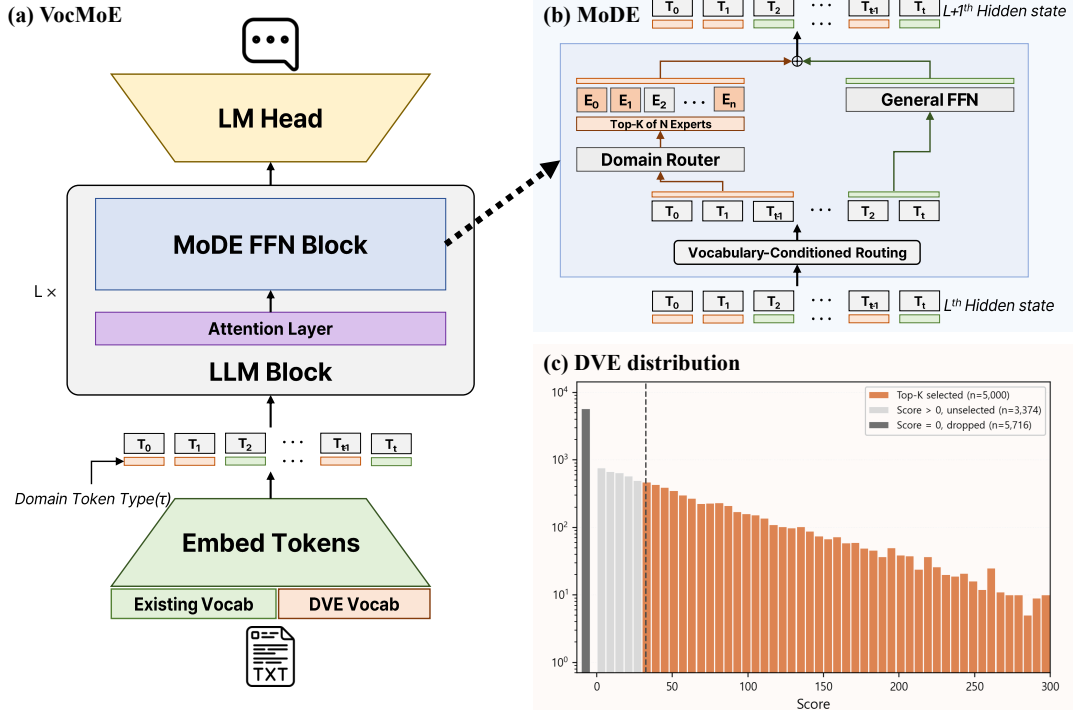


Figure 1: *VocMoE* architecture. (a) Tokens are embedded using the extended vocabulary (Existing + DVE) and routed by τ through L MoDE blocks. (b) **MoDE**. τ separates token paths according to token type: domain tokens are processed by the top- K medical experts, while general tokens pass through the preserved General FFN before the two paths are merged. (c) **DVE**. Distribution of candidate-token scores (Eq. 1). Tokens with Score = 0 are discarded, and the top- $K = 5,000$ tokens are added to the tokenizer.

the hidden state through a learned router across N experts, temporarily disrupting the representation alignment that the base model acquired during pre-training and forcing it to be re-aligned via domain training (§2). MoDE addresses this bottleneck by moving the routing decision from the learning domain to the vocabulary domain (Figure 1(b)). Our use of the MoE architecture differs from typical sparse MoE: the main objective is forgetting prevention via architectural isolation of the general-token path, while the MoE structure is restricted to providing specialization capacity for medical tokens.

The token type is determined deterministically from the vocabulary id:

$$\tau_t = \mathbf{1}[\text{id}_t \in \mathcal{V}_{\text{med}}], \quad (2)$$

where $\mathcal{V}_{\text{med}}$ is the set of medical token ids identified by DVE (newly added tokens plus base-vocabulary tokens flagged as medical by DVE scoring). General tokens ($\tau = 0$) and domain tokens ($\tau = 1$) are branched deterministically at the token id itself, without any learned classifier.

The FFN block of each transformer layer is then

branched according to the token type:

$$\text{out}_t = \begin{cases} \text{FFN}_{\text{gen}}(h_t) & \tau_t = 0, \\ \sum_{e \in \mathcal{T}_t} w_e \text{FFN}_e^{\text{med}}(h_t) & \tau_t = 1, \end{cases} \quad (3)$$

where $\mathcal{T}_t = \text{TopK}(R(h_t))$ is the set of top- K expert indices selected by the medical router $R(\cdot)$. FFN_{gen} shares the same dimension (intermediate_size) as the base model’s dense FFN, and is initialized by copying the base weights via Sparse Upcycling (Komatsuzaki et al., 2023). $\{\text{FFN}_e^{\text{med}}\}_{e=1}^{N_{\text{med}}}$ are N_{med} medical expert FFNs (with moe_intermediate_size), and the medical router $R(\cdot) = W_R h_t$ is applied only to domain tokens to select the top- K experts. Because the router is updated only on medical tokens, its learning burden is small. Existing sparse MoE and modality-conditioned routing methods rely on hidden-state-conditioned learned routers, causing transient disruption of the forward path during early training. In contrast, MoDE determines routing directly from token ids, maintaining fixed routing throughout training.

This architecture enforces two forms of isolation by construction. (i) **Forward isolation**. Be-

Domain	Source	Size (GB)	Language
Medical	Public	107.4	Ko+En
	Clinical inst.	0.8	Ko
General	NIKL	32.1	Ko
	Wiki	7.1	Ko
Total		147.4	

Table 3: Korean CPT corpus composition. Medical sources cover clinical text, papers, guidelines, and reports; General sources cover newspaper and encyclopedia. Sub-source breakdown in Appendix B (Table A3).

cause no medical-expert or router parameter appears on the FFN path of $\tau_t=0$ tokens, the general-token FFN forward path remains structurally preserved throughout training (the effect of shared self-attention on hidden states is analyzed in Appendix Figure A1). (ii) **Gradient isolation.** The loss generated by each token type produces zero gradient on the parameters of the other path, resulting in independent learning dynamics between general and medical tokens.

3.3 Training

Training Corpus. We use a **108.2 GB Korean medical corpus** (Medical/Ko, integrating and preprocessing AI-Hub and clinical-institution materials) together with a **39.2 GB Korean general corpus** (General/Ko: NIKL Modu newspaper, Korean Wikipedia, and Namuwiki) (Table 3). After preprocessing (deduplication and length filtering), the mixed corpus contains approximately **19.5B tokens**, with a token-level mix ratio of **Medical/Ko 76% / General/Ko 24%** set to preserve general Korean capability. A full sub-source breakdown is provided in Appendix B (Table A3).

Continual Pre-Training. We perform continual pre-training on the mixed corpus described above, exposing the model to approximately 19.5B training tokens. The training loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{aux}}, \quad \alpha = 0.001, \quad (4)$$

where $\mathcal{L}_{\text{CE}}$ is the standard next-token prediction cross-entropy loss and $\mathcal{L}_{\text{aux}}$ is a load-balancing auxiliary loss in the form of Switch Transformer (Fedus et al., 2022), applied only to medical tokens:

$$\mathcal{L}_{\text{aux}} = N_{\text{med}} \sum_{e=1}^{N_{\text{med}}} f_e \cdot p_e, \quad (5)$$

where f_e is the fraction of medical tokens assigned to expert e and p_e is the per-expert average of the

router probabilities over medical tokens. When no medical token is present in a batch, the medical router produces no output and $\mathcal{L}_{\text{aux}}$ computation is skipped.

To strengthen the weight-level isolation of the general FFN at the optimization dynamics level, we apply parameter grouping: general FFN parameters use $\eta_{\text{gen}} = 0.1 \eta$ (one tenth of the base learning rate), while the remaining parameters (medical experts, router, attention, embedding, lm_head) use the base rate η .

Supervised Fine-Tuning. After CPT, we apply LoRA (Hu et al., 2022)-based supervised fine-tuning to enhance the model’s instruction-following capability. The training data combine medical-domain QA and general-domain QA, jointly aligning domain expertise and general response quality. The detailed composition is described in Appendix C.

4 Experiments

4.1 Experimental Setup

MediKo Line-up. We train three MediKo models under the VocMoE framework, varying only the base model. MediKo-1.1B-A0.6B is built on Qwen3-0.6B with $N_{\text{med}}=8$ medical experts, top- $K=2$, and an MoE intermediate dimension of 768. MediKo-4B-A1B uses the same expert configuration on Qwen3-1.7B with an MoE intermediate dimension of 1280. MediKo-9B-A4B is built on Qwen3-4B with $N_{\text{med}}=16$, $K=4$, and the same MoE intermediate dimension. The first two share the expert configuration so that the effect of base scale can be isolated, while the 9B-A4B variant doubles both the expert pool and the top- K to verify capacity scaling.

Baseline Models.

The *General-Multi* group includes Qwen3 0.6B/1.7B/4B/8B (Yang et al., 2025), Llama-3.1-8B (Dubey et al., 2024), and Mistral-7B-v0.3 (Jiang et al., 2023). The *General-Ko* group includes KULLM3 (NLP & AI Lab, Korea University, 2024), SOLAR-10.7B (Kim et al., 2024a), and Polyglot-Ko-12.8B (Ko et al., 2023), which together measure how Korean-specialized training transfers to the medical domain. The medical baselines comprise OpenBioLLM-8B (Pal and Sankarababu, 2024) and BioMistral-7B (Labrak et al., 2024) as *En-Medical* cases of English medical CPT applied to Korean, MMed-Llama3-8B (Qiu et al., 2024) as a representative *Multi-Medical* case of Ko-

Group	Model	Active Params	KorMed	KMMLU-Med	EnMed	KoBEST	HAE-RAE
General-Multi	Qwen3-0.6B	0.6B	3.2 $\pm$ 0.3	22.9 $\pm$ 2.3	44.7 $\pm$ 1.0	48.2 $\pm$ 0.8	22.9 $\pm$ 1.3
	Qwen3-1.7B	1.7B	35.8 $\pm$ 0.9	36.6 $\pm$ 2.6	57.0 $\pm$ 1.0	48.6 $\pm$ 0.8	40.1 $\pm$ 1.5
	Qwen3-4B	4B	38.2 $\pm$ 0.9	47.1 $\pm$ 2.6	66.9 $\pm$ 1.0	53.1 $\pm$ 0.8	50.2 $\pm$ 1.6
	Qwen3-8B	8B	42.2 $\pm$ 0.7	52.4 $\pm$ 2.5	70.7 $\pm$ 0.9	64.0 $\pm$ 0.7	57.7 $\pm$ 1.6
	Llama-3.1-8B	8B	27.9 $\pm$ 0.8	39.2 $\pm$ 2.6	68.5 $\pm$ 1.0	54.4 $\pm$ 0.8	44.7 $\pm$ 1.6
	Mistral-7B-v0.3	7B	29.6 $\pm$ 0.9	27.6 $\pm$ 2.3	63.5 $\pm$ 1.0	51.8 $\pm$ 0.8	38.2 $\pm$ 1.6
General-Ko	KULLM3	10.7B	42.7 $\pm$ 0.9	39.1 $\pm$ 2.6	65.4 $\pm$ 1.0	67.7 $\pm$ 0.7	54.0 $\pm$ 1.6
	SOLAR-10.7B	10.7B	5.9 $\pm$ 0.4	31.4 $\pm$ 2.5	62.9 $\pm$ 1.0	53.5 $\pm$ 0.8	52.2 $\pm$ 1.6
	Polyglot-Ko-12.8B	12.8B	6.9 $\pm$ 0.5	25.9 $\pm$ 2.3	40.8 $\pm$ 1.0	64.9 $\pm$ 0.7	21.9 $\pm$ 1.3
En-Medical	OpenBioLLM-8B	8B	39.9 $\pm$ 0.9	36.4 $\pm$ 2.6	64.6 $\pm$ 1.0	53.1 $\pm$ 0.8	49.7 $\pm$ 1.6
	BioMistral-7B	7B	30.5 $\pm$ 0.9	30.1 $\pm$ 2.5	60.2 $\pm$ 1.0	48.4 $\pm$ 0.8	31.7 $\pm$ 1.5
Multi-Medical	MMed-Llama3-8B	8B	40.0 $\pm$ 0.9	30.4 $\pm$ 2.5	68.5 $\pm$ 1.0	47.8 $\pm$ 0.8	36.2 $\pm$ 1.5
Ko-Medical	Kor-Med-LLaMA3-8B	8B	38.9 $\pm$ 0.9	35.1 $\pm$ 2.6	65.3 $\pm$ 1.0	60.3 $\pm$ 0.8	47.8 $\pm$ 1.6
	hari-q3-8b	8B	27.9 $\pm$ 0.8	51.2 $\pm$ 2.5	<u>70.7</u> $\pm$ 0.9	<u>65.0</u> $\pm$ 0.7	<u>59.1</u> $\pm$ 1.5
Ours	MediKo-1.1B-A0.6B	0.59B	32.6 $\pm$ 0.8	33.3 $\pm$ 2.4	51.1 $\pm$ 1.0	53.4 $\pm$ 0.8	43.1 $\pm$ 1.5
	MediKo-4B-A1B	1.69B	44.8 $\pm$ 0.9	44.5 $\pm$ 2.6	60.8 $\pm$ 1.0	59.3 $\pm$ 0.8	50.2 $\pm$ 1.5
	MediKo-9B-A4B	4.5B	57.5 $\pm$ 0.9	53.0 $\pm$ 2.6	66.9 $\pm$ 1.0	63.9 $\pm$ 0.8	62.1 $\pm$ 1.6

Table 4: Main benchmark results — 5-benchmark Avg score, zero-shot. **Bold** = column-wise 1st; underline = 2nd. Medical-domain: KorMedMCQA (doctor, nurse, pharmacist, dentist licensing), KMMLU Medical (biology, health), English Medical (MedQA, PubMedQA). General-domain: KoBEST (5 reasoning sub-tasks), HAE-RAE (5 culture/language sub-tasks). Subscripts denote standard errors. Sub-task breakdowns are provided in Appendix D.

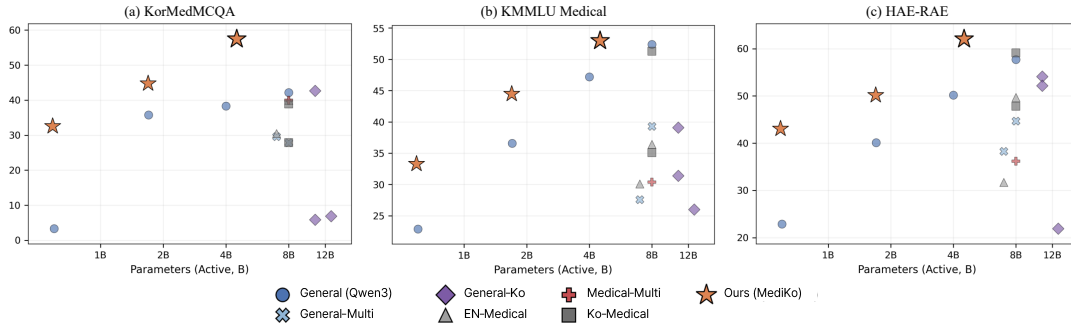


Figure 2: Active-parameter Pareto on KorMedMCQA / KMMLU Medical / HAE-RAE. X: active parameters (B, log scale); Y: benchmark Avg (%). MediKo (★, orange) updates the Pareto frontier on all three benchmarks.

rean medical-domain underperformance, and two *Ko-Medical* comparators that share our Korean-medical training strategy: Kor-Med-LLaMA3-8B (ih9511, 2024) and hari-q3-8b (Healthcare AI Research Institute (HARI), Seoul National University Hospital (SNUH), 2025).

Benchmarks. We evaluate medical capability on three benchmark groups. The first is *KorMedMCQA* (Kweon et al., 2024), a Korean medical licensing examination benchmark consisting of four sub-tasks (doctor, nurse, pharmacist, and dentist), evaluated with exact match accuracy. The second is KMMLU Medical (Son et al., 2024a), which includes biology and health subsets evaluated using accuracy. The third consists of English medical benchmarks, including MedQA 4-option (Jin et al., 2021) and PubMedQA (Jin et al., 2019). General Korean capability is evaluated on

KoBEST (Jang et al., 2022), which contains BoolQ, COPA, HellaSwag, SentiNeg, and WiC, as well as HAE-RAE (Son et al., 2024b), which covers general knowledge, history, loan words, rare words, and standard nomenclature. All evaluations are conducted in the zero-shot setting, and the three MediKo models are compared against 14 baselines.

4.2 Main Results

This section reports MediKo against all baselines on the five benchmarks. Table 4 consolidates the Avg score of each benchmark so that (i) medical-domain recovery and (ii) general Korean preservation can be read jointly; sub-task breakdowns are deferred to Appendix D.

Medical Domain Recovery. MediKo-9B-A4B reaches SOTA on both **KorMedMCQA** (+14.8 points over the previous best, KULLM3) and

#	CPT	DVE	MoDE	SFT	KorMed	KMMLU-Med	EnMed	KoBEST	HAE-RAE
1	-	-	-	-	38.2	47.1	66.9	53.1	50.2
2	✓	-	-	-	48.2	47.0	65.3	55.1	53.2
3	✓	✓	-	-	50.8	47.9	64.2	57.6	55.9
4	✓	✓	✓	-	54.2	48.5	63.0	60.8	58.0
5	✓	✓	✓	✓	57.5	53.0	66.9	63.9	62.1

Table 5: Component-wise ablation on a Qwen3-4B base. ✓ marks active components; row #4 is our CPT model (Ours, MediKo-9B-A4B) and row #5 adds SFT. All scores are zero-shot Avg.

KMMLU Medical, while MediKo-4B-A1B (active 1.69B) already surpasses Qwen3-8B (8B) and KULLM3 (10.7B) on KorMedMCQA (Figure 2). In contrast, the *adaptation failure* previously observed in Medical LLMs does not reappear under our architecture: while every existing medical-training paradigm (En/Multi/Ko-Medical) trails the corresponding General baselines (Table 1, §4.5.2), all three MediKo models exceed comparable active-parameter baselines across both medical and general axes, and active-parameter scaling translates monotonically into medical-domain accuracy across the entire MediKo lineup (Table 4). These results directly support the hypothesis that the cascading representational bottleneck can be substantially mitigated at the architectural level.

General Capability Preservation. Across all model scales, MediKo consistently improves KoBEST and HAE-RAE over the corresponding base models, demonstrating that general capability is preserved during medical-domain adaptation. Notably, MediKo-9B-A4B achieves a new HAE-RAE SOTA of 62.1, surpassing the strongest baseline, hari-q3-8b (59.1), even after medical specialization. Simultaneous SOTA on the medical (KorMedMCQA 57.5) and general (HAE-RAE 62.1) axes quantitatively demonstrates that the conventional domain-up / general-down trade-off of domain LLMs is substantially alleviated at the architectural level.

4.3 Ablation Studies

We perform a stage-wise ablation on the Qwen3-4B base to isolate each component’s contribution (Table 5).

Exposure to the Korean medical corpus alone (#2) substantially improves KorMed but barely affects KMMLU Medical, indicating that data alone does not recover domain representations. Adding DVE on top (#3) lifts the medical scores further, and the general-domain scores rise as well. Combining MoDE (#4) improves the medical scores

again while the general-domain scores continue to rise, so the two bottleneck-resolving components do not trade off against each other. Finally, SFT (#5) lifts every axis together, showing that post-training on top of MoDE does not damage the general path. The English medical benchmark declines slightly during CPT, which is an expected domain trade-off of Korean medical adaptation and recovers to the base level after SFT.

4.4 Routing Control

The stage-wise ablation above shows the effect of MoDE and SFT in sequence, but it does not separate the two factors. To control for this, we train a comparison model in which the base, corpus, training budget, and SFT are all held identical and only the routing mechanism is replaced (Table 6). A standard MoE with a learned router (std-MoE) falls below its base on the general benchmarks during CPT, so forgetting does occur. By contrast, MoDE preserves general capability under the same conditions. On the medical benchmarks the two routing schemes differ only marginally, which confirms that the contrast is confined to the general path. Both stages contribute: SFT improves the general benchmarks under either routing scheme, but it does not close the gap left by the routing choice, and on general ability MoDE without SFT already exceeds std-MoE with SFT. The preserved general capability therefore reflects vocabulary-conditioned separation during CPT together with the subsequent SFT alignment, rather than SFT alone.

Both models are trained under the same budget, and their comparable medical scores indicate that std-MoE is not simply under-trained. This control also provides direct evidence for the claim in §1 that the alignment bottleneck re-emerges when expanded and existing tokens share a single learning path.

Routing	SFT	KorMed	KMMLU-Med	EnMed	KoBEST	HAE-RAE
MoDE (ours)	-	54.2	48.5	63.0	60.8	58.0
std-MoE (learned router)	-	53.1	47.7	60.7	48.3	45.8
MoDE (ours)	✓	57.5	53.0	66.9	63.9	62.1
std-MoE (learned router)	✓	55.8	51.6	64.4	55.7	53.3

Table 6: Routing control. The base model, corpus, training budget, and SFT are identical, and only the FFN routing mechanism differs. **Bold** marks the better of the two routing schemes within each pair. All scores are zero-shot Avg.

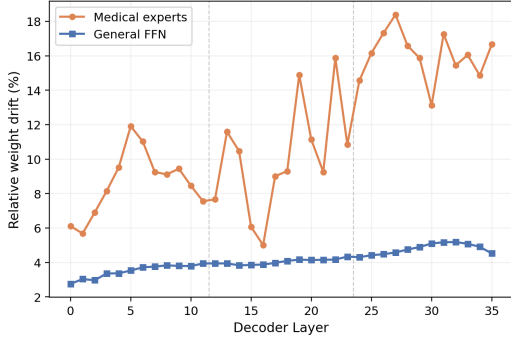


Figure 3: *Parameter-level Weight Drift on MediKo-9B-A4B*. Per-layer relative weight change $\|\Delta W\|/\|W\|$. Medical experts (5–18%) actively learn while the General FFN (2.7–5.2%) stays in a narrow stable band.

4.5 Analysis

This section complements the quantitative results with mechanism-level evidence, examining how VocMoE attains the joint medical/general gains.

4.5.1 Mechanism Analysis

We examine how the forward / gradient isolation promised in §3.2 manifests in the trained model along two axes: **parameter level** (Weight Drift, Figure 3) and **representation level** (t-SNE, Figure 4).

Parameter-level: Weight Drift. As shown in Figure 3, the General-FFN drift stays in a narrow stable band across all layers while the medical-experts drift is active and concentrated at the upper layers, indicating that domain specialization is sharper in the later layers. This separation is the joint effect of gradient isolation (§3.2) and parameter grouping (§3.3), demonstrating at the FFN-parameter level that medical learning signals do not act on the General FFN directly. The same separation pattern holds at every scale and General-FFN drift decreases monotonically as the base grows (Appendix Table A14).

Representation-level: t-SNE Specialization Pattern. We project hidden states of medical-corpus tokens at three layers of MediKo-9B-A4B via t-SNE (Figure 4). The two clusters partially overlap at the

embedding ($L=0$); at the middle layer ($L=18$), the activated medical experts pull medical tokens into a clearly separated cluster, visualizing MoDE’s semantic specialization; at the final layer ($L=36$), the clusters partially re-mix for LM-head vocabulary alignment, yet domain-specific sub-clusters persist in the local neighborhood. This “separation in the middle, integration at the output” pattern shows at the representation level that the medical gains and general preservation in §4.2 are two facets of the same architecture.

Synthesis. Together, these parameter- and representation-level observations reflect the two isolation properties of §3.2: gradient isolation manifests as the narrow drift band of the General FFN (Figure 3), while forward isolation, combined with the medical-router specialization, surfaces as the middle-layer cluster separation in t-SNE (Figure 4).

4.5.2 Medical Token Modeling Quality

We now test whether our framework addresses the *Korean medical underperformance* introduced in §1, in which existing Medical LLMs underperform comparable General LLMs in medical-token modeling. Using the same evaluation protocol as Table 1, we measure medical and general token perplexity (PPL) and Top-1 accuracy for both MediKo and all baselines. Full numerical results are provided in Appendix Table A15.

Reduction of the underperformance. The Qwen3 series exhibits normal scaling behavior on medical-token Top-1 accuracy, whereas existing Medical LLMs consistently underperform at comparable active-parameter scales. For example, MMed-Llama3-8B shows severe degradation, while BioMistral and OpenBioLLM both fall below Qwen3-8B. Both Korean medical baselines, KorMed-LLaMA3-8B and hari-q3-8b, trail Qwen3-8B (by 5.2 and 0.2 points respectively) despite explicit Korean medical-domain training. These results suggest that language- or domain-specialized training alone is insufficient to surpass the scaling behav-

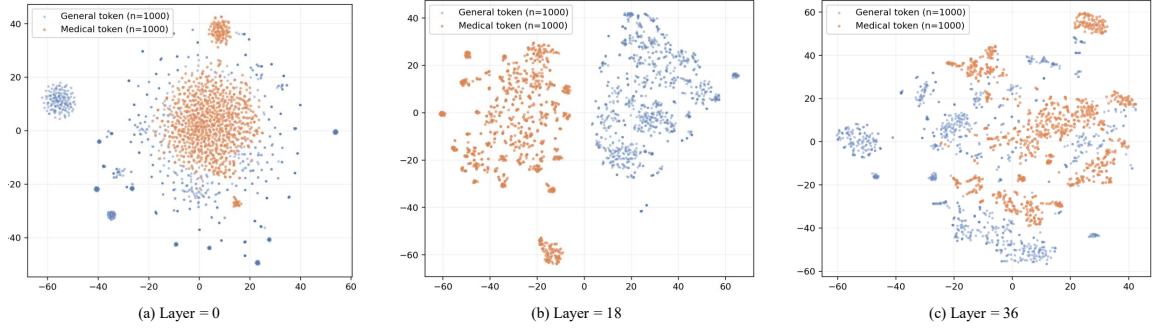


Figure 4: *Representation-level t-SNE on MediKo-9B-A4B*. Hidden-state progression of medical / general tokens at (a) L=0 (mixed), (b) L=18 (clear separation), and (c) L=36 (partial re-mixing for LM head alignment) — visualizing MoDE’s “separation in the middle, integration at the output” pattern, where domain specialization emerges in intermediate layers without disrupting final-layer vocabulary alignment.

ior of strong general LLMs. In contrast, all three MediKo models outperform comparable active-parameter baselines. MediKo-9B-A4B achieves medical Top-1 accuracy of **63.8%** with PPL **3.95** using only 4.5B active parameters, surpassing both Qwen3-8B (62.9% / 5.41) and the strongest Korean-general baseline, KULLM3 (63.1% / 4.27). Even the smallest MediKo-1.1B-A0.6B (active 0.59B) outperforms Qwen3-0.6B, indicating that MoDE improves the active-parameter efficiency frontier.

General-token preservation. MediKo-9B-A4B achieves the lowest general-token PPL among all evaluated models at **3.46**, corresponding to a 42% relative reduction from Qwen3-8B (5.96). Notably, medical-domain training improves general-token modeling rather than degrading it. This behavior is consistent with the KoBEST and HAE-RAE preservation results in §4.2, providing empirical support for MoDE’s forward isolation at the token level.

5 Conclusion

The *paradoxical failure* observed in Korean medical LLMs reflects a *cascading two-bottleneck*, which we hypothesize generalizes to other non-English medical settings. VocMoE addresses both: the DVE-defined vocabulary partition itself serves as the routing prior of MoDE, shifting catastrophic-forgetting prevention from training heuristics to the architectural level. MediKo attains SOTA on both Korean medical and general benchmarks, substantially alleviating the conventional domain-up / general-down trade-off.

Limitations

This study is validated on the Korean medical domain as a case study. We expect VocMoE to generalize to other languages and domains, but this

remains a hypothesis rather than an empirically verified result. At the architectural level, MoDE’s forward isolation is structurally complete at the FFN level, but the self-attention module is shared between general and domain tokens, so hidden states undergo partial mixing at the attention stage (quantified in Appendix Figure A1); introducing token-type-based isolation in the attention module as well is a natural follow-up direction.

On the inference side, although the active parameters of MediKo-9B-A4B (4.5B) are small relative to comparable dense models, the total memory (9.4B) is larger, so the GPU memory footprint at inference is higher than that of a dense model. Furthermore, because our main objective is forgetting prevention, general tokens pass through a dense FFN, giving an active ratio of about 50% — higher than the 10–30% typical of sparse MoE; this is our architectural choice, but it does impose an inference-compute trade-off. On the corpus side, despite our self-collected 108 GB Korean medical corpus, the total CPT token count (≈ 19.5 B) is still modest compared to English medical LLMs (PMC-LLaMA 75B / Meditron 48B), which reflects the inherent constraint of non-English medical corpus availability rather than something fully resolved by this work.

Ethics Statement

This model is designed as an assistive tool for medical information processing, summarization, and education, and does not replace clinical decision-making such as diagnosis or prescription. Any medical use must take place under the review and responsibility of qualified medical professionals, and because incorrect output from a medical-domain LLM can lead to harm, verification of every output

is required at use time. In addition, the model is specialized to Korean clinical environments, practice guidelines, and medical terminology, so generalization to other national or regional medical systems can introduce cultural and linguistic bias.

Acknowledgements

This work was partly supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2023-KH140182), and partly supported by the “Advanced GPU Utilization Support Program” funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. MEDITRON-70B: Scaling medical pre-training for large language models. *arXiv preprint arXiv:2311.16079*. 48B tokens medical CPT on LLaMA-2-70B.
- ChangSu Choi, Yongbin Jeong, Seoyoon Park, InHo Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and Kyung-Tae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. *arXiv preprint arXiv:2403.10882*. Bllosom: Korean vocab expansion on LLaMA2 + bilingual CPT + Korean LIMA SFT. Highlights byte-level over-fragmentation issue.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7B: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*. Mistral base + 30B legal tokens CPT.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*. Chinese vocab expansion on LLaMA; reports byte-token duplication problem (cited as motivation for VocMoE Bottleneck 1).
- Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Zhao, Xiangyue Sun, Shirong Ma, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Fine-grained expert segmentation, shared expert isolation. arXiv:2401.06066.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39. Sparse MoE, top-1 routing, load-balancing auxiliary loss. arXiv:2101.03961.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8342–8360.
- Healthcare AI Research Institute (HARI), Seoul National University Hospital (SNUH). 2025. hari-q3-8b. <https://huggingface.co/snuh/hari-q3-8b>. Hugging Face model card; Qwen3-8B-Base fine-tuned on Korean medical QA.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2106.09685.
- ih9511. 2024. Llama-3-Korean-8B continual pretrained on Korean–English medical translation corpus. https://huggingface.co/ih9511/llama-3-Korean-8B_koen_medical_translation. Hugging Face model card; Llama-3 base with Korean medical CPT.
- Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. KOBEST: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. ArXiv:2204.04541, anthology 2022.coling-1.325.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas,

- Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*. 8x7B MoE, top-2 routing.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421. ArXiv:2009.13081.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. ArXiv:1909.06146.
- Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Park, Hyunbyung Kim, Yongsoo Bak, Yongbeom Kim, Sungho Park, and 1 others. 2024a. SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT, Industry Track)*. ArXiv:2312.15166.
- Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024b. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*. Korean vocab expansion + curriculum CPT.
- Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. 2022. KM-BERT: A pre-trained language model for Korean medical natural language processing. *Scientific Reports*, 12. Korean medical BERT trained with BPE on Korean medical corpus.
- Hyunwoong Ko, Kichang Yang, Minhoo Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, and Sungho Park. 2023. A technical report for Polyglot-Ko: Open-source large-scale Korean language models. *arXiv preprint arXiv:2306.02254*.
- Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Dense to MoE conversion standard technique. arXiv:2212.05055.
- Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyoung Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. KorMedMCQA: Multi-choice question answering benchmark for Korean healthcare professional licensing examinations. *arXiv preprint arXiv:2403.01469*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics (ACL)*. Mistral base + PubMed Central CPT.
- Junbum Lee. 2024. Llama-2-Ko: Korean adapted llama 2 with vocabulary expansion. HuggingFace model: beomi/llama-2-ko-7b. <https://huggingface.co/beomi/llama-2-ko-7b>.
- Ning Liao, Xiaoxing Wang, Zehao Lin, Weiyang Guo, Feng Hong, Shixiang Song, Geng Yu, Zihua Zhao, Sitao Xie, Longxuan Wei, Xiangqi Jin, Xiaohan Qin, Jiale Ma, Kai Chen, Jiangchao Yao, Zhouhan Lin, Junchi Yan, Zhiyu Li, Feiyu Xiong, and 2 others. 2025. Innovator: Scientific continued pretraining with fine-grained MoE upcycling. *arXiv preprint arXiv:2507.18671*. Qwen2.5-7B upcycled to 53.3B-A13.3B (shared general expert + 64 scientific experts, top-8), 4-stage training paradigm.
- Chengyuan Liu, Shihang Wang, Lizhi Qing, Kun Kuang, Yangyang Kang, Changlong Sun, and Fei Wu. 2024. Gold panning in vocabulary: An adaptive method for vocabulary expansion of domain-specific LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. VEGAD: adaptive vocabulary expansion via gradient. arXiv:2410.01188.
- Itay Nakash, Nitay Calderon, Eyal Ben David, Elad Hoffer, and Roi Reichart. 2025. AdaptiVocab: Enhancing LLM efficiency in focused domains through lightweight vocabulary adaptation. In *Proceedings of the Conference on Language Modeling (COLM)*. Lightweight vocab adaptation for focused domains. arXiv:2503.19693.
- NLP & AI Lab, Korea University. 2024. KULLM3: Korean instruction-tuned large language model. <https://huggingface.co/nlpai-lab/KULLM3>. Hugging Face model card; fine-tuned from upstage/SOLAR-10.7B-v1.0.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLM-8B: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B>. Hugging Face model card; Saama AI Labs; built on Meta-Llama-3-8B.
- Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*. ArXiv:2402.13963.

- Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024. MoME: Mixture of multimodal experts for generalist multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. MoVE (vision experts) + MoLE (language experts), arXiv:2407.12709.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024a. KMMLU: Measuring massive multitask language understanding in Korean. *arXiv preprint arXiv:2402.11548*.
- Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaechol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2024b. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. ArXiv:2309.02706.
- Linglin Wang and 1 others. 2025. EvoMoE: Expert evolution in mixture of experts for multimodal large language models. *arXiv preprint arXiv:2505.23830*. Expert evolution + Dynamic Token-aware Router (DTR) for multimodal LLMs (arXiv:2505.23830). Full author list: Linglin Wang et al.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843. 75B tokens medical CPT on LLaMA.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*. 50B params, 50% finance + 50% general.
- Guoyang Xia, Yifeng Ding, Fengfa Li, Lei Ren, Wei Chen, Fangxiang Feng, and Xiaojie Wang. 2025. SMAR: Soft modality-aware routing strategy for moe-based multimodal large language models preserving language capabilities. *arXiv preprint arXiv:2506.06406*. KL divergence regularization for modality-aware expert specialization.
- An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*. Dense and MoE family 0.6B/1.7B/4B/8B/30B-A3B; base of MediKo lineup.

A Tokenization Examples

This appendix illustrates DVE’s selective compression at the word- and sentence-level.

A.1 Word-level Comparison

Table A1 directly compares the sub-word decomposition of five Korean medical terms and two general words under the Qwen3 (base) tokenizer and the DVE-expanded MediKo tokenizer. Clinical terms such as 심근경색 are fragmented down to the character level by the base tokenizer but collapse to single tokens under MediKo, whereas general vocabulary (학교, 컴퓨터) is tokenized identically by both, providing micro-scale evidence that DVE’s selective expansion does not perturb general vocabulary.

A.2 Sentence-level Comparison

Table A2 shows the sub-word sequences of one medical-dense Korean sentence (“심근경색 진단 되어 입원하였다”) under both tokenizers, in the Blossom-style side-by-side format. Only the clinical term 심근경색 is compressed to a single token (4→1), while the general verbs 되어/하였다 and mid-frequency tokens 진단/입원 retain the same token count (only the split position shifts), confirming at the sentence level that the word-level selectivity reported in Table A1 also holds in context.

B Training Corpus (Full Composition)

Table A3 provides the sub-source breakdown of the Korean medical and general corpora summarized in §3.3.

C Implementation Details

The implementation used for training and the released model weights are publicly available.¹ This appendix lists the training hyperparameters for the three paper-reported MediKo models. **CPT** is performed on the mixed Korean medical+general corpus described in Appendix B (~19.5B tokens), and **SFT** on a self-built Korean medical instruction dataset of 105,393 samples distilled from Gemini 2.5 Flash. All training runs use H200 GPUs with DeepSpeed ZeRO-2 and bf16 mixed precision.

¹<https://huggingface.co/collections/MediKo/mediko>

D Detailed Benchmark Results (Sub-task)

Tables A6–A10 present per-sub-task results that were collapsed into the main-results table (Table 4) by averaging across sub-tasks. All numbers are zero-shot, reported as mean $\pm$ standard error. **Bold** = column-wise 1st, underline = 2nd (matching the convention of Table 4); $\downarrow$ (Tables A9 and A10) marks a statistically meaningful drop vs. the same-scale General LLM.

E Additional Ablation

This appendix reports ablations on two DVE design choices — vocabulary scaling (K_{DVE}) and embedding initialization — measured on the MediKo-1.1B-A0.6B base.

Term	Qwen3 (base)	MediKo (DVE)	Red.	Pattern
심근경색 (myocardial infarction)	[심, 근, 경, 색] (4)	[심근경색] (1)	75%	full compression
갑상선기능저하증 (hypothyroidism)	[갑, 상, 선, 기, 능, 저, 하, 증] (8)	[갑상선기능저하증] (1)	88%	full compression
급성신부전 (acute renal failure)	[급, 성, 신, 부, 전] (5)	[급성신부전] (1)	80%	full compression
인슐린저항성 (insulin resistance)	[인, 술, 린, 저, 항, 성] (6)	[인슐린저항성] (1)	83%	full compression
학교 (school)	[학교] (1)	[학교] (1)	0%	no change (general)
컴퓨터 (computer; loanword)	[컴, 퓨, 터] (3)	[컴, 퓨, 터] (3)	0%	no change (DVE selective)

Table A1: Word-level sub-word decomposition before and after DVE expansion, measured on Qwen3-0.6B and MediKo’s DVE-expanded tokenizer. Bracketed sequences are per-token displays via the GPT-2 byte decoder; numbers in parentheses are total token counts. DVE compresses clinically dense Korean terms to single tokens (full compression) while leaving general vocabulary (e.g., 학교, 컴퓨터) unchanged, providing direct evidence of selective expansion.

Sentence: 심근경색 진단되어 입원하였다. (Diagnosed with myocardial infarction and hospitalized.)	
Model	Tokenization
Qwen3 (base)	[심, 근, 경, 색, ▷진, 단, 되어, ▷입, 원, 하였다, .] (11 tokens)
MediKo (DVE)	[심근경색, ▷, 진단, 되어, ▷, 입원, 하였다, .] (8 tokens, 27% reduction)

Table A2: Sentence-level tokenization comparison (Blossom-style; Choi et al., 2024). The leading ▷ marks a word-boundary (space) token. DVE compresses 심근경색 (4→1) while leaving general words (되어, 하였다) and split positions of mid-frequency tokens (진단, 입원) unchanged, providing micro-evidence consistent with the candidate-token average fragmentation reduction $4.5 \rightarrow 1.6$ (64%) reported in §3.1.

Domain	Sub-source	Size (GB)	Content
Medical (Ko)	AI-Hub Healthcare	37.8	medical info
	AI-Hub Clinical dialogue	32.4	clinical records
	AI-Hub Med textbook	5.6	textbook
	Healthcare articles	12.5	news article
	Medical/health news	1.3	news
	Medical papers	10.1	sci. paper
	Medical books	7.7	book
	Clinical-institution data	0.8	guidelines, reports
General (Ko)	NIKL Modu newspaper	32.1	newspaper
	Wikipedia (Ko)	1.3	encyclopedia
	Namuwiki	5.8	community wiki
Total		147.4	19.5B tokens

Table A3: Full sub-source breakdown of the training corpus. The Medical (Ko) corpus (108.2 GB) integrates AI-Hub public-domain sources and clinical-institution materials; the General (Ko) data is collected from public HuggingFace repositories. After preprocessing (deduplication and length filtering), the mixed corpus contains approximately 19.5B tokens at the Medical/Ko 76% : General/Ko 24% token ratio.

Model	Ep.	BS	Accum	LR	Warmup	Max seq	Gen FFN scale
MediKo-1.1B-A0.6B	2	16	8	2.0×10^{-4}	0.05	8192	$\times 0.1$
MediKo-4B-A1B	3	12	8	1.5×10^{-4}	0.05	8192	$\times 0.1$
MediKo-9B-A4B	3	10	10	1.0×10^{-4}	0.05	8192	$\times 0.1$

Table A4: CPT hyperparameters. Optimizer = AdamW (weight decay 0.0); scheduler = linear warmup + cosine decay. Load-balancing auxiliary loss coefficient $\alpha = 10^{-3}$, applied only to medical tokens (§3.3). General-FFN parameters use $\eta_{\text{gen}} = 0.1 \eta$ to suppress forgetting; all remaining parameters (medical experts, router, attention, embedding) use the base learning rate η .

Model	Ep.	BS	Accum	LR	Warmup	Max seq	LoRA r/ α /drop
MediKo-1.1B-A0.6B	3	12	5	2.0×10^{-4}	0.03	4096	64 / 128 / 0.10
MediKo-4B-A1B	3	12	8	1.0×10^{-4}	0.03	4096	64 / 128 / 0.05
MediKo-9B-A4B	3	8	8	7.5×10^{-5}	0.03	4096	64 / 128 / 0.05

Table A5: SFT hyperparameters. LoRA targets attention modules only (q/k/v/o projections); general FFN and medical experts remain frozen. Optimizer = AdamW (weight decay 0.1); scheduler = linear warmup + cosine decay.

Group	Model	Active Params	Doctor	Nurse	Pharm.	Dentist	Avg
General (Multi)	Qwen3-0.6B	0.6	1.8 $\pm$ 0.6	4.9 $\pm$ 0.7	3.0 $\pm$ 0.6	3.2 $\pm$ 0.6	3.2 $\pm$ 0.3
	Qwen3-1.7B	1.7	24.1 $\pm$ 2.0	46.7 $\pm$ 1.7	41.5 $\pm$ 1.7	30.8 $\pm$ 1.6	35.8 $\pm$ 0.9
	Qwen3-4B	4	29.4 $\pm$ 2.2	46.4 $\pm$ 1.7	48.1 $\pm$ 1.7	29.1 $\pm$ 1.6	38.2 $\pm$ 0.9
	Qwen3-8B	8	36.1 $\pm$ 1.8	48.2 $\pm$ 1.3	44.8 $\pm$ 1.4	<u>39.9$\pm$1.1</u>	42.2 $\pm$ 0.7
	Llama-3.1-8B	8	20.5 $\pm$ 1.9	35.3 $\pm$ 1.6	29.1 $\pm$ 1.5	26.9 $\pm$ 1.6	27.9 $\pm$ 0.8
	Mistral-7B-v0.3	7	25.1 $\pm$ 2.1	34.4 $\pm$ 1.6	34.8 $\pm$ 1.6	24.3 $\pm$ 1.5	29.6 $\pm$ 0.9
General (Ko)	KULLM3	10.7	39.1 $\pm$ 2.3	51.7 $\pm$ 1.7	46.7 $\pm$ 1.7	33.2 $\pm$ 1.6	42.7 $\pm$ 0.9
	SOLAR-10.7B	10.7	1.4 $\pm$ 0.6	12.2 $\pm$ 1.1	2.3 $\pm$ 0.5	7.8 $\pm$ 0.9	5.9 $\pm$ 0.4
	Polyglot-Ko-12.8B	12.8	7.4 $\pm$ 1.2	5.8 $\pm$ 0.8	8.6 $\pm$ 0.9	5.9 $\pm$ 0.8	6.9 $\pm$ 0.5
En-Medical	OpenBioLLM-8B	8	37.2 $\pm$ 2.3	44.2 $\pm$ 1.7	43.0 $\pm$ 1.7	35.0 $\pm$ 1.7	39.9 $\pm$ 0.9
	BioMistral-7B	7	25.1 $\pm$ 2.1	35.4 $\pm$ 1.6	33.8 $\pm$ 1.6	27.9 $\pm$ 1.6	30.5 $\pm$ 0.9
Multi-Medical	MMed-Llama3-8B	8	39.8 $\pm$ 2.4	43.0 $\pm$ 1.7	44.2 $\pm$ 1.7	33.2 $\pm$ 1.6	40.0 $\pm$ 0.9
Ko-Medical	Kor-Med-LLaMA3-8B	8	34.5 $\pm$ 2.3	45.3 $\pm$ 1.7	44.6 $\pm$ 1.7	31.3 $\pm$ 1.6	38.9 $\pm$ 0.9
	hari-q3-8b	8	23.2 $\pm$ 2.0	30.9 $\pm$ 1.6	36.4 $\pm$ 1.6	21.1 $\pm$ 1.4	27.9 $\pm$ 0.8
Ours	MediKo-1.1B-A0.6B	0.59	33.0 $\pm$ 2.0	36.0 $\pm$ 1.5	32.2 $\pm$ 1.5	29.3 $\pm$ 1.5	32.6 $\pm$ 0.8
	MediKo-4B-A1B	1.69	41.8 $\pm$ 2.1	52.6 $\pm$ 1.7	45.0 $\pm$ 1.6	39.7 $\pm$ 1.6	44.8 $\pm$ 0.9
	MediKo-9B-A4B	4.5	48.8$\pm$2.4	68.5$\pm$1.6	66.0$\pm$1.6	46.9$\pm$1.8	57.5$\pm$0.9

Table A6: KorMedMCQA sub-task accuracy (zero-shot, exact match).

Group	Model	Active Params	Biology	Health	Avg
Gen. (Multi)	Qwen3-0.6B	0.6	20.8 $\pm$ 1.3	25.0 $\pm$ 4.3	22.9 $\pm$ 2.3
	Qwen3-1.7B	1.7	29.2 $\pm$ 1.4	44.0 $\pm$ 5.0	36.6 $\pm$ 2.6
	Qwen3-4B	4	37.3 $\pm$ 1.5	57.0 $\pm$ 5.0	47.1 $\pm$ 2.6
	Qwen3-8B	8	40.8 $\pm$ 1.6	64.0$\pm$4.8	52.4 $\pm$ 2.5
	Llama-3.1-8B	8	31.5 $\pm$ 1.5	47.0 $\pm$ 5.0	39.2 $\pm$ 2.6
	Mistral-7B-v0.3	7	28.1 $\pm$ 1.4	27.0 $\pm$ 4.5	27.6 $\pm$ 2.3
Gen. (Ko)	KULLM3	10.7	28.2 $\pm$ 1.4	50.0 $\pm$ 5.0	39.1 $\pm$ 2.6
	SOLAR-10.7B	10.7	26.7 $\pm$ 1.4	36.0 $\pm$ 4.8	31.4 $\pm$ 2.5
	Polyglot-Ko-12.8B	12.8	26.9 $\pm$ 1.4	25.0 $\pm$ 4.3	25.9 $\pm$ 2.3
En-Med	OpenBioLLM-8B	8	28.7 $\pm$ 1.4	44.0 $\pm$ 5.0	36.4 $\pm$ 2.6
	BioMistral-7B	7	27.2 $\pm$ 1.4	33.0 $\pm$ 4.7	30.1 $\pm$ 2.5
Multi-Med	MMed-Llama3-8B	8	28.8 $\pm$ 1.4	32.0 $\pm$ 4.7	30.4 $\pm$ 2.5
Ko-Med	Kor-Med-LLaMA3-8B	8	28.2 $\pm$ 1.4	42.0 $\pm$ 5.0	35.1 $\pm$ 2.6
	hari-q3-8b	8	40.5 $\pm$ 1.6	62.0 $\pm$ 4.8	51.2 $\pm$ 2.5
Ours	MediKo-1.1B-A0.6B	0.59	33.7 $\pm$ 1.4	32.9 $\pm$ 4.6	33.3 $\pm$ 2.4
	MediKo-4B-A1B	1.69	31.4 $\pm$ 1.4	57.6 $\pm$ 5.0	44.5 $\pm$ 2.6
	MediKo-9B-A4B	4.5	42.8$\pm$1.5	<u>63.1$\pm$5.0</u>	53.0$\pm$2.6

Table A7: KMMLU Medical sub-task accuracy (zero-shot).

Group	Model	Active Params	MedQA	PubMedQA	Avg
Gen. (Multi)	Qwen3-0.6B	0.6	32.4±1.3	57.0±1.6	44.7±1.0
	Qwen3-1.7B	1.7	42.0±1.4	72.1±1.4	57.0±1.0
	Qwen3-4B	4	57.3±1.4	76.5±1.3	66.9±1.0
	Qwen3-8B	8	64.3±1.3	77.1±1.3	70.7±0.9
	Llama-3.1-8B	8	61.1±1.4	75.8±1.4	68.5±1.0
	Mistral-7B-v0.3	7	49.8±1.4	77.3±1.3	63.5±1.0
Gen. (Ko)	KULLM3	10.7	53.9±1.4	76.9±1.3	65.4±1.0
	SOLAR-10.7B	10.7	54.8±1.4	70.9±1.4	62.9±1.0
	Polyglot-Ko-12.8B	12.8	27.5±1.2	54.0±1.6	40.8±1.0
En-Med	OpenBioLLM-8B	8	58.0±1.4	71.2±1.4	64.6±1.0
	BioMistral-7B	7	46.5±1.4	73.9±1.4	60.2±1.0
Multi-Med	MMed-Llama3-8B	8	60.4±1.4	76.6±1.3	68.5±1.0
Ko-Med	Kor-Med-LLaMA3-8B	8	56.2±1.4	74.4±1.4	65.3±1.0
	hari-q3-8b	8	64.8±1.3	76.5±1.3	<u>70.7±0.9</u>
Ours	MediKo-1.1B-A0.6B	0.59	44.4±1.3	57.8±1.5	51.1±1.0
	MediKo-4B-A1B	1.69	51.4±1.4	70.2±1.4	60.8±1.0
	MediKo-9B-A4B	4.5	59.5±1.4	74.2±1.4	66.9±1.0

Table A8: English Medical sub-task accuracy (zero-shot).

Group	Model	Active Params	BoolQ	COPA	HSwag	SentiNeg	WiC	Avg
Gen. (Multi)	Qwen3-0.6B	0.6	50.4±1.3	52.9±1.6	36.4±2.1	52.6±2.5	48.8±1.4	48.2±0.8
	Qwen3-1.7B	1.7	51.8±1.3	56.9±1.6	36.4±2.1	49.4±2.5	48.8±1.4	48.6±0.8
	Qwen3-4B	4	53.7±1.3	63.2±1.5	42.8±2.2	56.9±2.5	48.8±1.4	53.1±0.8
	Qwen3-8B	8	61.1±1.3	69.6±1.5	45.2±2.2	89.9±1.5	<u>54.0±1.4</u>	64.0±0.7
	Llama-3.1-8B	8	52.8±1.3	68.1±1.5	43.4±2.2	58.7±2.5	48.8±1.4	54.4±0.8
	Mistral-7B-v0.3	7	57.7±1.3	59.3±1.6	42.6±2.2	50.4±2.5	48.8±1.4	51.8±0.8
Gen. (Ko)	KULLM3	10.7	88.0±0.9	71.4±1.4	45.6±2.2	84.4±1.8	48.8±1.4	67.7±0.7
	SOLAR-10.7B	10.7	50.6±1.3	57.6±1.6	43.0±2.2	67.8±2.4	48.8±1.4	53.5±0.8
	Polyglot-Ko-12.8B	12.8	56.9±1.3	78.5±1.3	<u>49.0±2.2</u>	91.2±1.4	48.8±1.4	64.9±0.7
En-Med	OpenBioLLM-8B	8	62.9±1.3	63.4±1.5	41.6±2.2	49.1±2.5	48.7±1.4	53.1±0.8
	BioMistral-7B	7	50.6±1.3	55.0±1.6	40.6±2.2	46.9±2.5	48.8±1.4	48.4±0.8
Multi-Med	MMed-Llama3-8B	8	50.2±1.3	55.6±1.6	33.4±2.1	50.9±2.5	48.8±1.4	47.8±0.8 ↓
Ko-Med	Kor-Med-LLaMA3-8B	8	57.8±1.3	67.0±1.5	42.2±2.2	85.6±1.8	48.8±1.4	60.3±0.8
	hari-q3-8b	8	64.0±1.3	69.3±1.5	44.8±2.2	<u>90.4±1.5</u>	56.2±1.4	<u>65.0±0.7</u>
Ours	MediKo-1.1B-A0.6B	0.59	52.2±1.3	59.4±1.6	43.4±2.2	62.0±2.5	50.0±1.4	53.4±0.8
	MediKo-4B-A1B	1.69	61.2±1.3	67.2±1.5	43.8±2.2	74.3±2.2	49.8±1.4	59.3±0.8
	MediKo-9B-A4B	4.5	<u>83.1±1.0</u>	<u>73.7±1.4</u>	49.7±2.2	64.0±2.4	48.8±1.4	63.9±0.8

Table A9: KoBEST sub-task accuracy (zero-shot). ↓ marks a meaningful drop vs. a same-scale General LLM.

Group	Model	Active Params	GenKnow	History	Loan	Rare	StdNom	Avg
Gen. (Multi)	Qwen3-0.6B	0.6	19.3±3.0	14.9±2.6	26.0±3.4	27.4±2.2	26.8±3.6	22.9±1.3
	Qwen3-1.7B	1.7	29.6±3.5	26.6±3.2	40.8±3.8	53.1±2.5	50.3±4.1	40.1±1.5
	Qwen3-4B	4	39.2±3.7	35.6±3.5	58.0±3.8	64.4±2.4	53.6±4.0	50.2±1.6
	Qwen3-8B	8	38.6±3.7	46.3±3.6	65.7±3.7	70.6±2.3	67.3±3.8	57.7±1.6
	Llama-3.1-8B	8	37.5±3.7	38.3±3.5	46.8±3.9	56.0±2.5	45.1±4.0	44.7±1.6
	Mistral-7B-v0.3	7	36.4±3.6	29.3±3.3	40.2±3.8	44.2±2.5	41.2±4.0	38.2±1.6
Gen. (Ko)	KULLM3	10.7	36.4±3.6	50.0±3.7	57.4±3.8	67.7±2.3	58.8±4.0	54.0±1.6
	SOLAR-10.7B	10.7	39.8±3.7	50.0±3.7	53.2±3.9	61.0±2.4	56.9±4.0	52.2±1.6
	Polyglot-Ko-12.8B	12.8	18.8±3.0	22.9±3.1	23.1±3.2	22.7±2.1	22.2±3.4	21.9±1.3
En-Med	OpenBioLLM-8B	8	36.9±3.6	41.5±3.6	61.5±3.8	56.8±2.5	51.6±4.0	49.7±1.6
	BioMistral-7B	7	24.4±3.2	24.5±3.1	37.9±3.7	37.0±2.4	34.6±3.9	31.7±1.5
Multi-Med	MMed-Llama3-8B	8	38.1±3.7	28.2±3.3	34.9±3.7	45.9±2.5	34.0±3.8	36.2±1.5↓
Ko-Med	Kor-Med-LLaMA3-8B	8	39.8±3.7	36.2±3.5	49.1±3.9	61.5±2.4	52.3±4.0	47.8±1.6↓
	hari-q3-8b	8	42.1±3.7	47.3±3.7	67.5±3.6	70.9±2.3	68.0±3.8	59.1±1.5
Ours	MediKo-1.1B-A0.6B	0.59	45.0±3.5	39.8±3.3	38.0±3.4	55.9±2.5	36.8±3.6	43.1±1.5
	MediKo-4B-A1B	1.69	42.3±3.0	49.8±3.5	48.6±3.8	59.3±2.5	51.0±4.0	50.2±1.5
	MediKo-9B-A4B	4.5	47.2±3.8	61.7±3.5	62.7±3.7	73.3±2.2	65.4±3.9	62.1±1.6

Table A10: HAE-RAE sub-task accuracy (zero-shot). ↓ marks a meaningful drop vs. a same-scale General LLM. **Bold** on Avg = SOTA within the corresponding (Ours) row.

K_{DVE}	KorMed	KMMLU-Med	EnMed	KoBEST	HAE-RAE
2,000	31.2	32.4	50.9	52.4	42.6
5,000 (default)	32.6	33.3	51.1	53.4	43.1
8,000	31.7	33.1	50.2	52.2	42.9

Table A11: Effect of varying the DVE top- K vocabulary expansion size. $K = 2,000$ under-covers domain vocabulary and degrades medical benchmarks (KorMed 31.2, KMMLU-Med 32.4); $K = 8,000$ relaxes the specificity threshold so general tokens slip in, marginally lowering general benchmarks (KoBEST 52.2, HAE-RAE 42.9). $K = 5,000$ is best on all five benchmarks, confirming it as the appropriate specificity threshold.

Init	KorMed	KMMLU-Med	EnMed	KoBEST	HAE-RAE
Random ($\mathcal{N}(0, 0.02)$)	31.8	32.9	50.5	52.7	42.8
KmBERT-projected	32.2	33.4	50.8	52.1	42.6
Mean (default)	32.6	33.3	51.1	53.4	43.1

Table A12: Embedding-initialization strategy for newly added DVE tokens. Mean (the average of constituent sub-word embeddings) aligns naturally with the base BPE space and is best on four of five benchmarks. KmBERT-projected explicitly injects domain semantics but loses base alignment (HAE-RAE 42.6 vs. Mean 43.1); Random lacks initial alignment, lowering KorMed (31.8 vs. 32.6). Mean balances domain learnability with base preservation.

F Expert Configuration

We select the number of experts N_{med} and top- K so that the expert capacity stays modest relative to the size of the domain vocabulary, and we verify this choice at the 9B scale after the fact. General tokens pass through the dense General FFN, which is independent of N_{med} and K , so only the medical path is affected and we compare medical-token PPL. Each configuration is trained on a subset of the CPT corpus under an identical budget, so we compare the relative ranking across configurations rather than absolute PPL. The adopted configuration ($N_{\text{med}} = 16$, $K = 4$) attains the global minimum of the grid, and both axes form a U shape, confirming that increasing N_{med} or K further does not help.

N_{med}	top- $K=2$	top- $K=4$	top- $K=8$
8	5.42	5.03	4.96
16	5.04	4.25	4.67
32	4.80	4.98	4.85

Table A13: Expert configuration probe at the 9B scale. Cells report medical-token PPL (lower is better) under an identical reduced training budget per configuration.

G Analysis Supplements

This appendix reports supplementary data referenced from §4.5: (i) Weight-Drift mean by scale (Table A14), (ii) hidden-state cosine-similarity layer plot across three scales (Figure A1), and (iii) full numerical Med/Gen token PPL and Top-1 with all 14 baselines (Table A15).

G.1 Weight Drift Mean Across Scales

Table A14 aggregates the layer-mean weight-drift magnitudes of §?? across all three MediKo scales, with medical experts and the General FFN reported separately. The General-FFN drift decreases monotonically as the base grows (5.7% $\rightarrow$ 5.4% $\rightarrow$ 4.1%), whereas the medical-experts drift stays substantially larger across all scales. This widening Medical/General gap indicates that the $\eta_{\text{gen}} = 0.1 \eta$ scaling preserves the general path more strongly at larger base models, even as the medical path continues to specialize.

G.2 Layer-wise Cosine Similarity

Figure A1 overlays the layer-wise base-vs-MediKo cosine similarity across the three scales, with normalized layer depth on the x-axis to make trajec-

ries comparable despite differing layer counts (29 for the 1.1B and 4B bases, 37 for the 9B base). The three curves share an almost identical shape, indicating that the forward-isolation property established in §3.2 holds invariantly as the base model is scaled up.

G.3 Medical/General Token PPL & Top-1

Table A15 reports the full numerical values for medical and general token PPL and next-token Top-1 across all 14 baselines plus the three MediKo models, complementing the figure-based summary in §4.5.2. MediKo-9B-A4B is column-wise 1st on all four metrics, and KULLM3 emerges as the only baseline that surpasses Qwen3-8B itself (Med Top-1 63.1%, Gen PPL 3.90); all four Medical-specialized baselines (En/Multi/Ko-Medical) trail both, reproducing the paradox first reported in Table 1 at the full-baseline scale.

H Case Study

This appendix qualitatively illustrates MediKo-9B-A4B’s outputs across three medical scenarios — (1) pharmacology, (2) emergency-medicine diagnostic reasoning, and (3) a mixed medical+general scenario — demonstrating that the quantitative gains in §4.5 (PPL, Top-1) translate to natural-language generation that simultaneously preserves Korean fluency and medical accuracy.

Model	Med exp.	Gen FFN	Ratio
MediKo-1.1B-A0.6B	22.5%	5.7%	3.9×
MediKo-4B-A1B	8.6%	5.4%	1.6×
MediKo-9B-A4B	11.6%	4.1%	2.8×

Table A14: Layer-mean of $\|\Delta W\|/\|W\|$ between the upcycled and CPT-finished checkpoints. Medical experts use the base LR η while General FFN uses $\eta_{\text{gen}} = 0.1 \eta$ (§3.3). **General-FFN drift decreases monotonically with scale** (5.7% $\rightarrow$ 5.4% $\rightarrow$ 4.1%), indicating that forgetting prevention strengthens at larger base.

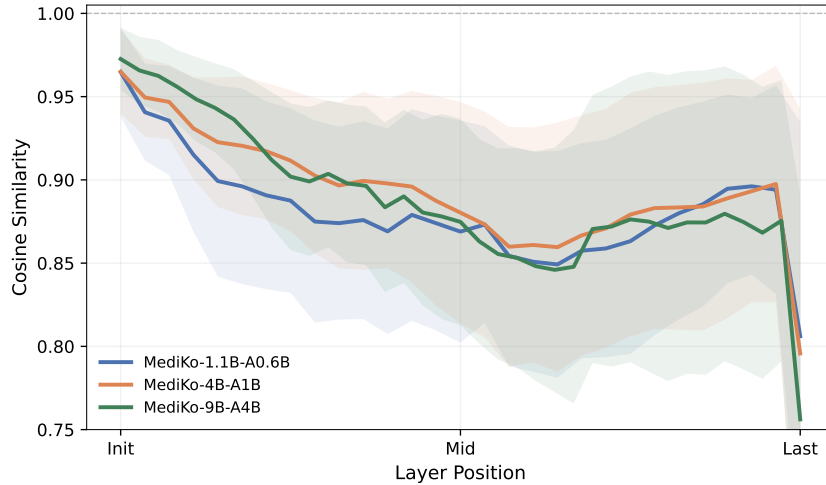


Figure A1: Layer-wise cosine similarity between base (Qwen3) and MediKo hidden states on a general Korean corpus (three scales overlaid; shaded band = ± 1 std). All scales follow the same trajectory — a gradual decline from ~ 0.97 (Init) to ~ 0.86 (Mid), a brief recovery, and a dip just before the `lm_head`. This scale-invariant pattern indicates that MoDE’s forgetting prevention is consistent across model sizes.

Group	Model	Med PPL ↓	Med Top-1 ↑	Gen PPL ↓	Gen Top-1 ↑
General (Multi)	Qwen3-0.6B	10.47	52.2%	11.66	50.4%
	Qwen3-1.7B	7.46	58.2%	8.35	56.3%
	Qwen3-4B	6.56	60.5%	7.27	58.8%
	Qwen3-8B	5.41	62.9%	5.96	61.1%
	Llama-3.1-8B	6.16	58.6%	6.96	56.5%
	Mistral-7B-v0.3	4.96	61.4%	4.66	<u>67.5%</u>
General (Ko)	KULLM3	4.27	<u>63.1%</u>	<u>3.90</u>	67.0%
	SOLAR-10.7B	8.08	<u>54.8%</u>	<u>4.69</u>	63.1%
	Polyglot-Ko-12.8B	7.72	51.9%	5.22	60.5%
En-Medical	BioMistral-7B	9.09	53.8%	6.41	59.2%
	OpenBioLLM-8B	8.68	54.5%	8.28	56.1%
Multi-Medical	MMed-Llama3-8B	38.58	37.9%	16.37	47.5%
Ko-Medical	Kor-Med-LLaMA3-8B	6.95	57.7%	7.52	57.1%
	hari-q3-8b	5.44	62.7%	6.14	60.1%
Ours	MediKo-1.1B-A0.6B	5.13	58.7%	4.28	64.5%
	MediKo-4B-A1B	<u>4.55</u>	61.2%	3.93	66.5%
	MediKo-9B-A4B	3.95	63.8%	3.46	69.0%

Table A15: Source data for §4.5.2: medical / general token PPL and Top-1 prediction accuracy on a Korean medical corpus (same protocol as Table 1; DVE token set defines medical tokens). **Bold** = column-wise 1st, underline = 2nd. **MediKo-9B-A4B achieves the lowest PPL and highest Top-1 on all four metrics simultaneously**, with KULLM3 emerging as the strongest baseline (Med Top-1 63.1%, Gen PPL 3.90) — surpassing Qwen3-8B itself. All four Medical-specialized baselines (En/Multi/Ko-Medical) trail both Qwen3-8B and KULLM3, confirming the Korean medical adaptation failure.

Role	Content
Q1	항생제를 장기간 복용하면 어떤 부작용이 발생할 수 있나요? (What side effects may occur with long-term antibiotic use?)
A1	항생제 장기 복용 시 (i) 항생제 내성균 발생 (세균이 저항력을 획득해 향후 치료 효과가 떨어질 수 있음), (ii) 장내 유익균 교란으로 인한 위장 장애 (설사·복통), (iii) 알러지 반응 (발진·가려움·호흡곤란, 심하면 쇼크), (iv) 간·신장 기능 부담, (v) 비타민 K·B군 결핍, (vi) 두통·어지럼증 등 신경계 영향이 보고됩니다. 의사의 지시에 따라 정확한 용량·기간을 지켜 복용하고, 이상 증상이 나타나면 즉시 의료 전문가와 상담해야 합니다. (Long-term antibiotic use may cause (i) antibiotic resistance (bacteria acquiring resistance and reducing future treatment efficacy), (ii) gastrointestinal disturbance from disruption of gut microbiota (diarrhea, abdominal pain), (iii) allergic reactions (rash, itching, dyspnea, shock in severe cases), (iv) hepatic and renal burden, (v) vitamin K and B-group deficiency, and (vi) neurological effects such as headache and dizziness. Patients should follow the prescribed dosage and duration, and consult a medical professional immediately if adverse symptoms appear.)
Q2	50대 남성이 갑작스러운 흉통과 호흡곤란을 호소합니다. 어떤 질환을 의심해야 하나요? (A 50-year-old man presents with sudden chest pain and dyspnea. Which conditions should be suspected?)
A2	이러한 증상은 (i) 심근경색 (Myocardial Infarction): 심장 근육 혈류 차단으로 발생하는 급성 흉통과 호흡곤란, (ii) 협심증: 심근 허혈로 인한 흉통이 휴식 시에도 나타나며 호흡곤란이 동반, (iii) 폐색전증 (Pulmonary Embolism): 다리 정맥에서 떨어진 혈전이 폐혈관을 막아 발생하는 갑작스러운 흉통·호흡곤란 등 생명을 위협하는 질환을 시사합니다. 즉시 의료 전문가의 진료가 필요하며, 심전도·심초음파·혈액 검사를 통해 정확한 진단과 치료가 이루어져야 합니다. (These symptoms suggest life-threatening conditions including (i) myocardial infarction (MI): acute chest pain and dyspnea due to interrupted coronary blood flow, (ii) angina pectoris: ischemic chest pain that may occur at rest and is accompanied by dyspnea, and (iii) pulmonary embolism (PE): sudden chest pain and dyspnea caused by a clot from a leg vein blocking a pulmonary vessel. Immediate medical evaluation is required, with diagnosis and treatment guided by ECG, echocardiography, and blood tests.)
Q3	우리 아이가 38.5도 열이 나는데 어떻게 하면 좋을까요? 학교는 보내야 하나요? (My child has a 38.5°C fever. What should I do? Should they go to school?)
A3	38.5°C 발열은 우선 충분한 휴식과 수분 섭취로 관리하시고, 의사의 처방에 따라 아세트아미노펜·이부프로펜 같은 해열제를 투여할 수 있습니다. 발열이 있는 동안에는 등교를 보류하고 학교에 상황을 알리는 것이 권장됩니다. 다만 고열이 지속되거나 구토·설사·호흡곤란·경련 등이 동반되면 즉시 응급실을 방문하시기 바랍니다. (A 38.5°C fever should be managed first with adequate rest and hydration, and antipyretics such as acetaminophen or ibuprofen may be given per a physician's instruction. The child should stay home from school during the fever and the parents are advised to inform the school. If high fever persists or is accompanied by vomiting, diarrhea, dyspnea, or seizures, immediate emergency care is required.)

Table A16: Zero-shot generation examples from MediKo-9B-A4B across three medical scenarios: (Q1) pharmacology, (Q2) emergency-medicine diagnostic reasoning, and (Q3) a mixed medical+general scenario. Each Korean response is shown verbatim with a small-italic English translation. The outputs demonstrate medical accuracy (specific differential diagnoses, drug names, and clinical guidance) while preserving Korean fluency and a natural conversational tone.