Mitigating Structural Knowledge Collapse in Domain-Specific LLMs via Morpheme-Aware KV-Aggregation

Yuxuan Si[1], **Zheqi Lv**[1]*, **Chengxi Zang**[2], **Zhengyu Chen**[3], **Fei Wu**[1]*
[1]Zhejiang University [2]Cornell University [3]Meituan Inc.

Abstract

Standard tokenizers over-fragment domain terms, disrupting morpheme semantics. We characterize this representational misalignment as Structural Knowledge Collapse (SKC), where attention mechanisms fail to reconstruct coherent concepts from fragmented inputs. While existing input-centric solutions like vocabulary expansion address this, they necessitate expensive embedding retraining and neglect internal attention compositionality. To this end, we introduce Morpheme-aware KV-aggregation Attention (MorphKA), a lightweight adapter that dynamically consolidates fragments without tokenizer changes. Bypassing tokenizer retraining, MorphKA employs a dual-phase strategy—Input-Level Morpheme Aggregation (IMA) and Context-Aware KV-Aggregation (AMRF)—to stabilize morpheme spans and synthesize higher-order concepts. Experiments on medical and legal benchmarks show MorphKA outperforms vocabulary adaptation baselines by 3.2–4.6%, reaching 7.9% on high-fragmentation terms. Moreover, MorphKA reduces catastrophic interference on general capabilities by 18–22% with ∼80% fewer parameters than embedding retraining approaches.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in general language understanding and generation (Brown et al., 2020; OpenAI, 2023; Dubey et al., 2024). However, adapting them to specialized domains such as medicine and law remains challenging, as pre-training corpora provide limited exposure to domain-specific terminology. In these domains, complex terms often derive meaning from compositional morphemes, the smallest meaning-bearing units. For example, *immunohistochemistry* is typically fragmented into subwords like "immun", "o", "hist", "o", "chem", "istry" by byte-pair encoding (BPE) tokenizers, disrupting integration of morphemes such as *immuno-*, *histo-*, and *chemistry-*. Similarly, *lex mercatoria* may split into "lex", "merc", "ator", "ia", obscuring the link between *lex* (law) and *mercatoria* (merchant) (Sennrich et al., 2016; Pinter et al., 2017; Batsuren et al., 2024). Such over-fragmentation overlooks key linguistic structure, contributing to suboptimal domain performance.

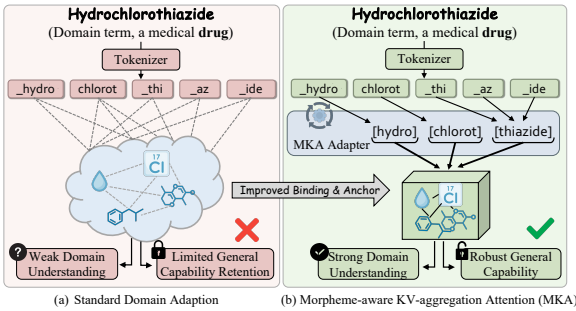

Figure 1: Structural knowledge collapse (SKC) caused by subword over-fragmentation in domain adaptation, and MorphKA's morpheme-aligned anchoring remedy.

Recent studies reveal the broad impact of subword fragmentation. In high out-of-vocabulary (OOV) medical summarization, advanced LLMs like Llama-3.1 show marked degradation (Balde et al., 2025a). Probing analyses expose model-specific patterns in subword compositionality, with boundaries impairing semantic decomposability (Peng et al., 2025). During domain adaptation, these effects intensify: fine-tuning encourages reliance on shallow fragment correlations rather than deep semantic integration, amplifying catastrophic forgetting (Kaushik et al., 2021; Liu et al., 2024a). These observations point to a critical challenge, which we term **structural knowledge collapse (SKC)**: the inability of standard attention mechanisms (Vaswani et al., 2017) to effectively recompose fragmented morphemes into coherent concepts, creating a semantic bottleneck that hin-

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29399–29413

ders robust domain understanding (Fig. 1).

To counteract such fragmentation, prior work has predominantly focused on vocabulary-level interventions. Vocabulary expansion and morphology-aware tokenizers reduce OOV rates by incorporating domain terms into the lexicon (Liu et al., 2024b; Cui et al., 2023). While effective in reducing sequence length, these approaches incur high computational costs due to embedding retraining and often disrupt the alignment of the pre-trained feature space. Other strategies, such as dynamic retrofitting, enable on-the-fly composition via hypernetworks but introduce tokenizer incompatibilities and limited integration during training (Feher et al., 2025a; Asgari et al., 2025). Conventional continual learning techniques preserve parameters but rely on generic attention mechanisms that remain insensitive to the morpheme structure (Ke et al., 2023; Hu et al., 2021).

A fundamental limitation persists across these paradigms: they fail to dynamically integrate morpheme compositionality into the attention mechanism itself without altering the tokenizer. Input modifications are rigid and resource-intensive, while standard fine-tuning lacks the structural inductive bias needed to repair fragmentation internally (Liu et al., 2024a).

To this end, we introduce Morpheme-aware KV-aggregation Attention (MorphKA), a lightweight adapter that induces hierarchical structural concepts in pretrained LLMs without tokenizer changes. MorphKA plugs into standard attention to reconstruct fragmented subwords into cohesive units, mitigating SKC while supporting general capabilities integration. It employs a dual-phase strategy aligned with layer-wise compositionality (Peng et al., 2025): early **Input-Level Morpheme Aggregation (IMA)** consolidates K/V vectors across morpheme spans via masks, stabilizing representations against fragmentation; deeper **Context-Aware KV-Aggregation by Adaptive Multiscale Routing Fusion (AMRF)** dynamically fuses evidence to form higher concepts, preserving positional fidelity. MorphKA essentially acts as a semantic glue that re-bonds fragmented subwords dynamically during inference, enhancing morpheme-sensitive tasks and perturbation resilience with low overhead. As in Fig. 2, MorphKA selectively assigns higher significance to morpheme-rich medical spans.

Experiments on terminology-dense medical and legal tasks show that MorphKA substantially outperforms strong baselines, including adaptive vocabulary expansion (Liu et al., 2024b) and dynamic subword merging (Feher et al., 2025b). It delivers average accuracy gains of 3.2–4.6% over the strongest non-MorphKA methods, with improvements surging to 7.9–8.7% on high-fragmentation examples ($p_{\max} \geq 5$). Simultaneously, MorphKA achieves superior general capability integration, reducing interference on MMLU and GSM8K by 18–22 percentage points relative to baselines. These benefits come at low cost: only ∼15.7M trainable parameters, roughly 80% fewer than vocabulary expansion approaches that retrain embeddings.Our main contributions are:

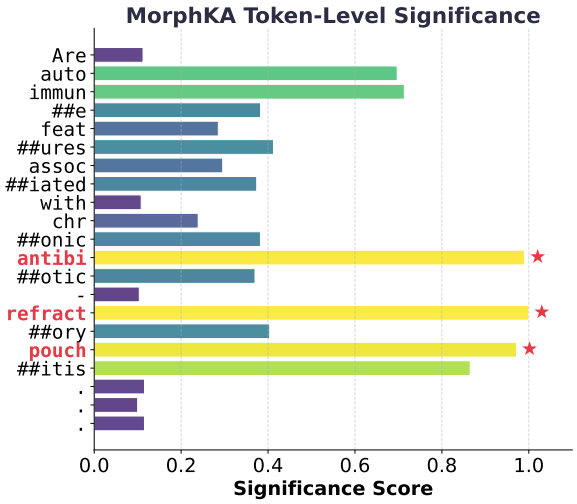

Figure 2: Token-level significance scores induced by MorphKA on a PubMedQA example, highlighting selective intervention on morpheme-rich medical spans.

- We characterize SKC as a representational misalignment in domain adaptation, demonstrating that this misalignment acts as a primary bottleneck for learning coherent domain concepts.
- We introduce **MorphKA**, a training-efficient attention adapter that mitigates SKC by dynamically aggregating morphemes into coherent conceptual units without tokenizer or embedding changes.
- We propose a dual-phase strategy (IMA and AMRF) that repairs fragmentation at complementary depths, enabling both robust domain performance and harmonious integration of general capabilities beyond mere forgetting mitigation (Liu et al., 2024a).
- Through extensive evaluation on medical and legal tasks, fragmentation-stratified analyses, and mechanistic probes, we demonstrate state-

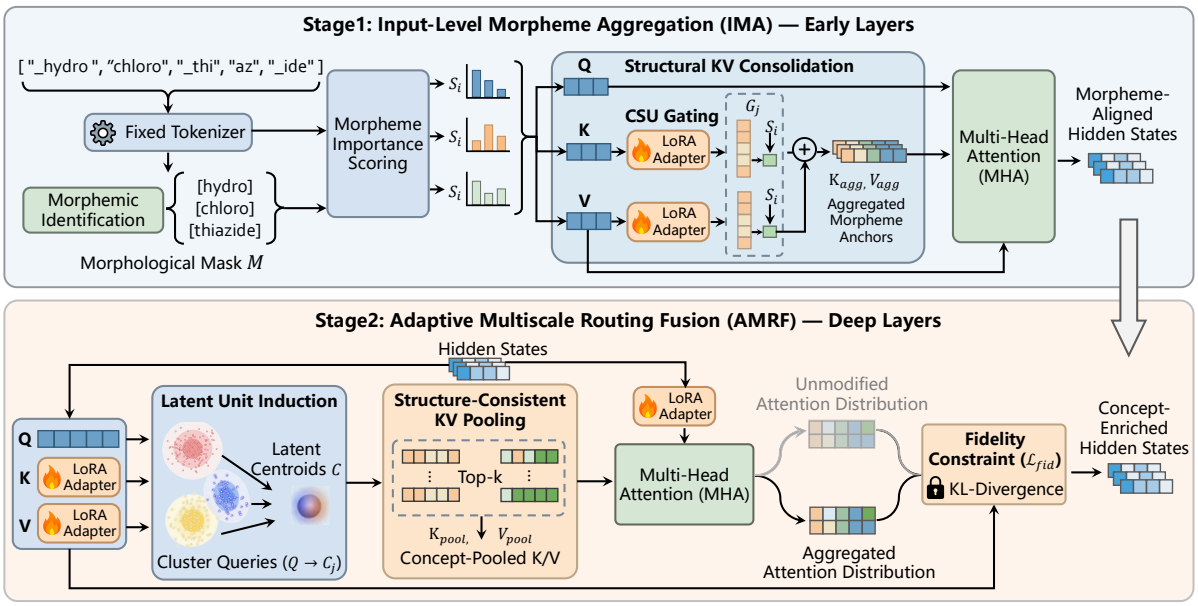

Figure 3: Overview of Morpheme-aware KV-aggregation Attention (MorphKA). Stage 1 (IMA) consolidates subword fragments into morpheme-level KV anchors in early layers. Stage 2 (AMRF) induces latent conceptual units and performs structure-consistent KV pooling in deep layers, regularized by a fidelity constraint to preserve the original attention behavior.

of-the-art domain accuracy, reduced catastrophic interference, and favorable parameter efficiency compared to vocabulary modification and dynamic merging baselines.

2 Related Work

Domain Adaptation in LLMs Large language models (LLMs) excel in general tasks but face challenges in specialized domains due to limited terminology exposure (Brown et al., 2020; Dubey et al., 2024). Adaptation methods include continued pretraining (Cui et al., 2023), parameter-efficient fine-tuning (e.g., LoRA) (Hu et al., 2021), and collaborative learning (Lv et al., 2025; Yu et al., 2026), which boost domain performance but often cause catastrophic forgetting of general capabilities (Kaushik et al., 2021). Recent efforts extend to general capabilities integration (GCI) (Liu et al., 2024a), yet they overlook subword fragmentation issues that fragment representations and hinder robust domain-general harmony.

Subword Fragmentation and Compositionality Subword tokenizers like BPE (Sennrich et al., 2016) handle rare terms but disrupt morpheme boundaries, impairing semantic decomposability (Peng et al., 2025; Batsuren et al., 2024). In high-OOV domains, this leads to severe degradation (Balde et al., 2025a). Prior morphology-aware approaches require retraining or fail to dynamically compose morphemes in attention (Creutz and Lagus, 2007).

Vocabulary Expansion and Dynamic Methods Vocabulary expansion adds domain tokens to reduce OOV (Liu et al., 2024b; Cui et al., 2023), while dynamic merging enables runtime composition (Feher et al., 2025a). These improve efficiency but demand embedding retraining, introduce incompatibilities, or lack training-time morpheme integration into attention. MorphKA addresses these by aggregating morphemes lightweightly without altering tokenizers or pretrained representations.

3 Methodology

To address *structural knowledge collapse* (SKC), we introduce *Morpheme-aware KV-Aggregation Attention* (MorphKA), a lightweight, plug-and-play adapter that injects a length-preserving inductive bias into the attention mechanism of pre-trained LLMs. MorphKA achieves this without altering the tokenizer or expanding the vocabulary, making it compatible with existing deployment pipelines.

The key insight behind MorphKA is to selectively repair fragmented semantic units by consolidating dispersed information in the key-value (KV) space, while preserving standard self-attention for well-formed tokens. As shown in Figure 3, MorphKA employs a dual-phase architecture: (i) *Input-level Morpheme Aggregation* (IMA) in early layers, which enforces local synchronization within identified morpheme units, and (ii) *Context-Aware*

KV-Aggregation by Adaptive Multiscale Routing Fusion (AMRF) in deeper layers, which enables context-dependent concept induction across units.

3.1 Compositional Semantic Units Discovery

We define *Compositional Semantic Units* (CSUs) as morpheme-like latent spans: groups of subword tokens bounded by morpheme-likeness saliency, detected via a hybrid model-based diagnostic process rather than fixed rules.

Morphemic Container Identification. To avoid cross-word interference, we identify coarse lexical containers $\{[B_s, E_s)\}_{s=1}^{S}$ using tokenizer-specific markers (e.g., underscores in SentencePiece) or whitespace boundaries. Each token i is assigned to a container s via a maximum-overlap mapping:

$$g(i) = \arg\max_{s\in\{1,\dots,S\}} \Big| [b_i, e_i) \cap [B_s, E_s) \Big|. \quad (1)$$

where these containers act as locality constraints.

Morpheme-likeness Boundary Diagnosis. Within each container, we detect boundaries using a mixed saliency score. Let $\bar{\mathbf{h}}_i$ be the detached hidden representation from a stop-gradient forward pass. For adjacent tokens $(i, i+1)$, we compute:

$$\rho_i = \lambda(1 - \cos(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_{i+1})) + (1-\lambda)\eta(t_i, t_{i+1}), \quad (2)$$

where $\cos(\cdot,\cdot)$ measures representational similarity, and $\eta(\cdot) \in [0,1]$ incorporates surface cues like continuation prefixes or character shifts. Boundaries are flagged if $\rho_i > \delta$, yielding refined CSUs $\{\mathcal{U}_j\}_{j=1}^{J}$. This ensures aggregation only for spans with both neural and lexical evidence.

Compositional Failure Gating. For each CSU $\mathcal{U}_j$, we assess fragmentation severity using: *Fragmentation Intensity* to capture subword granularity. And *Geometric Dispersion* to detect representational disjointness. These are fused into a differentiable gate: $\gamma_j = \sigma(w_1 s_{\text{frag}}(j) + w_2 s_{\text{disp}}(j) - \tau)$. High γ_j triggers MorphKA intervention for fragmented units; low values bypass it for intact ones.

The operational definitions of these quantities are provided in Appendix A.1.

3.2 S1: Input-level Morpheme Aggregation

In early layers ($\ell \le L_e$), IMA synchronizes the semantic manifolds within each diagnosed CSU. Given the hidden state $\mathbf{H}^{(\ell)}$, we project it into content-only queries $\mathbf{q}_i$, pre-positional keys $\tilde{\mathbf{k}}_i$, and values $\mathbf{v}_i$. To distill a cohesive morpheme-like unit kernel, IMA computes an intra-unit importance distribution $\alpha_{j,r} = \text{softmax}_{r\in\mathcal{U}_j}(\mathbf{w}_s^\top \mathbf{h}_r^{(\ell)})$, and aggregates unit-level content as:

$$\tilde{\mathbf{k}}_j = \sum_{r\in\mathcal{U}_j} \alpha_{j,r}\tilde{\mathbf{k}}_r, \quad \mathbf{v}_j = \sum_{r\in\mathcal{U}_j} \alpha_{j,r}\mathbf{v}_r. \quad (3)$$

The synchronized representations are obtained via gated interpolation:

$$\begin{bmatrix}\tilde{\mathbf{k}}'_i \\ \mathbf{v}'_i\end{bmatrix} = (1-\beta_\ell\gamma_{u(i)}) \begin{bmatrix}\tilde{\mathbf{k}}_i \\ \mathbf{v}_i\end{bmatrix} + \beta_\ell\gamma_{u(i)} \begin{bmatrix}\tilde{\mathbf{k}}_{u(i)} \\ \mathbf{v}_{u(i)}\end{bmatrix}. \quad (4)$$

Crucially, positional embeddings are applied *after* aggregation, $\mathbf{k}'_i = \text{PE}(\tilde{\mathbf{k}}'_i, \text{pos}_i)$, ensuring that tokens share a semantic core while retaining relative positional information required for syntactic parsing.

3.3 S2: Adaptive Multiscale Routing Fusion

In deep layers ($\ell > L - L_d$), MorphKA transitions to global semantic synthesis. AMRF introduces K learnable semantic anchors $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K}$ as latent bottlenecks for cross-unit evidence consolidation. For each position-aware query $\mathbf{q}'_i$, we compute a routing distribution:

$$p_{i,k} = \text{softmax}_k\left(\frac{(\mathbf{q}'_i)^\top \mathbf{c}_k}{\tau_r\sqrt{d}}\right). \quad (5)$$

Anchors aggregate sequence-wide information as:

$$\tilde{\mathbf{k}}_{A(k)} = \frac{\sum_i p_{i,k}\tilde{\mathbf{k}}'_i}{\sum_i p_{i,k} + \epsilon}, \mathbf{v}_{A(k)} = \frac{\sum_i p_{i,k}\mathbf{v}'_i}{\sum_i p_{i,k} + \epsilon}. \quad (6)$$

Anchors are treated as global memory tokens with a fixed virtual position, and attention operates over the augmented KV set $\mathbf{K}_{\text{tot}} = [\mathbf{k}'_{1:N}; \mathbf{k}_{A(1:K)}]$. This dual-scale retrieval allows the model to reconcile fine-grained subword evidence with high-level conceptual summaries.

3.4 Fidelity Constraint and Objective

To preserve general linguistic knowledge, we impose a *fidelity loss* via a stop-gradient reference strategy:

$$\mathcal{L}_{\text{fid}} = \sum_{(\ell,h)\in\mathcal{S}_{\text{fid}}} \text{KL}\Big(\text{stopgrad}\big(A_{\text{ref}}^{(\ell,h)}\big) \,\|\, A_{\text{MorphKA}}^{(\ell,h)}\Big). \quad (7)$$

The total training objective is $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{fid}}$. Together with CSU gating, this constraint ensures that MorphKA functions as a precision corrective mechanism, intervening only when structural fragmentation is detected.

Method	Qwen3-8B				Llama-3.1-8B			
	MedQA	PubMedQA	CaseHOLD	BillSum	MedQA	PubMedQA	CaseHOLD	BillSum
Zero-shot Base	72.54	48.27	74.13	42.36	71.34	45.67	72.28	40.46
LoRA SFT	78.43	55.38	80.26	47.18	76.57	53.97	77.34	44.96
Wise-FT ($\alpha = 0.4$)	79.96	56.74	81.07	48.05	77.68	55.28	78.27	45.86
DAS	78.67	55.93	80.57	47.56	76.83	54.39	77.59	45.27
Dynamic Merging	77.28	53.97	78.86	46.04	75.46	52.58	76.19	44.08
Random-Grouping (Control)	75.37	51.62	77.49	44.47	73.27	49.46	74.29	42.26
MorphKA (Ours)	**83.58**	**60.87**	**85.79**	**52.28**	**80.49**	**59.39**	**82.48**	**49.67**
Tokenizer/Vocab Modification								
DV (Vocab-Exp)	81.26	58.94	83.57	50.16	78.96	57.28	80.57	47.78
SPM + ATT_EG	81.97	59.58	84.19	50.78	79.57	58.09	81.27	48.47

Table 1: **Domain-task results** on Qwen3-8B and Llama-3.1-8B. We report Accuracy for MedQA, PubMedQA, and CaseHOLD, and ROUGE-L for BillSum. *Random-Grouping* is a boundary-control baseline with randomized spans. **Bold** denotes the best result; underline denotes the strongest non-MorphKA baseline.

4 Experiments

4.1 Experimental Setup

We evaluate on Qwen3-8B-Instruct (Team, 2025) and Llama-3.1-8B-Instruct (Dubey et al., 2024). MorphKA is implemented as a LoRA-based adapter (Hu et al., 2021). Unless stated otherwise, we use LoRA rank r=16 and scaling factor α=32. We apply *Input-Level Morpheme Aggregation* (IMA) to the first 10% of Transformer layers and *Adaptive Multiscale Routing Fusion* (AMRF) to the remaining layers (Section 3). We train with AdamW (lr 2×10^{-5}, weight decay 0.01) and cosine annealing on 8× NVIDIA A100 (80GB) GPUs. The per-device batch size is 4 with gradient accumulation, giving an effective batch size of 32. Full hyperparameters are listed in Appendix A.2.

4.1.1 Datasets and Evaluation Protocols

We target terminology-dense tasks in medical and legal domains. Dataset statistics in Appendix A.4.

Medical: MedQA-USMLE (Jin et al., 2021) (multiple-choice clinical reasoning) and PubMedQA (Jin et al., 2019) (yes/no/maybe biomedical QA), both rich in long compounds (e.g., "immunohistochemistry", "angiotensin-converting").

Legal: CaseHOLD (Zheng et al., 2021) (multiple-choice holding identification) and BillSum (Kornilova and Eidelman, 2019) (abstractive bill summarization), featuring Latin phrases, citations, and multi-word expressions (e.g., "lex mercatoria").

We report Accuracy for MedQA, PubMedQA, and CaseHOLD; ROUGE-L for BillSum. General capability retention uses average accuracy on MMLU (Hendrycks et al., 2021) and exact-match accuracy on GSM8K (Cobbe et al., 2021).

4.1.2 Baselines

MorphKA is compared to three baseline groups, all using the same LoRA configuration and training budget (details in Appendix A.3):

PEFT Controls: LoRA SFT (Hu et al., 2021); Wise-FT (Wortsman et al., 2021) ($\alpha = 0.4$); DAS (Ke et al., 2023).

Vocabulary/Tokenizer Modification: DV (Liu et al., 2024b) (domain vocabulary expansion); retrained domain-specific SentencePiece with attention-based embedding initialization (ATT_EG) (Liu et al., 2021).

Structural Interventions: Dynamic Merging (Feher et al., 2025b); Random-Grouping (same aggregation on random spans).

4.1.3 Domain-Term Lexicon and Fragmentation Stratification

For fine-grained fragmentation analysis, we construct a domain term lexicon $\mathcal{V}_{\text{dom}}$ from training data only, using deterministic patterns: (i) hyphenated or affix-heavy compounds, (ii) alphanumeric biomedical entities (e.g., "IL-6", "TNFα"), and (iii) legal citations (e.g., "U.S.C.", "Section #"). Terms appearing at least five times are retained; the lexicon is used solely for evaluation stratification.

For an example x and tokenizer $\tau(\cdot)$, we compute piece count $p(v) = |\tau(v)|$ for each matched term $v \in \mathcal{V}_{\text{dom}} \cap x$. Examples are bucketed by maximum piece count $p_{\max}(x)$: Low (≤ 2), Mid (3–4), High (≥ 5). This follows high-OOV benchmarking practice (Balde et al., 2025b).

4.2 Main Results

Tables 1 and 2 present the results on domain-specific tasks and the retention of general capabili-

Method	Domain Avg.	MMLU	GSM8K
Qwen3-8B			
LoRA SFT	65.31	78.57	82.43
Wise-FT ($\alpha = 0.4$)	66.46	79.28	83.16
DAS	65.68	78.86	82.79
MorphKA (Ours)	**70.63**	**81.07**	**85.68**
Llama-3.1-8B			
LoRA SFT	63.21	71.86	84.59
Wise-FT ($\alpha = 0.4$)	64.27	72.57	85.27
DAS	63.52	72.19	84.83
MorphKA (Ours)	**68.01**	**74.08**	**87.39**

Table 2: **General capability retention** after domain adaptation. *Domain Avg.* is the mean performance across the four domain tasks in Table 1. We report Accuracy on MMLU and GSM8K (higher is better). **Bold** indicates the best score among adaptation methods.

Method	Low ($\leq$2)	Mid (3–4)	High ($\geq$5)	Overall
Llama-3.1-8B				
LoRA	0.00	0.00	0.00	0.00
RG	-0.12±0.08	-0.45±0.11	-1.12±0.15	-0.48±0.09
MorphKA	**+1.24±0.10**	**+4.52±0.13**	**+8.74±0.18**	**+3.91±0.12**
Qwen3-8B				
LoRA	0.00	0.00	0.00	0.00
RG	-0.08±0.07	-0.32±0.10	-0.95±0.14	-0.35±0.08
MorphKA	**+1.02±0.09**	**+3.85±0.12**	**+7.96±0.16**	**+4.28±0.11**

Table 3: ΔAccuracy over LoRA SFT by bucket ($p_{\max}$). Means $\pm$ std over seeds. MorphKA surges in High, validating SKC focus. "RG" is the Random-Grouping.

ties, respectively, highlighting the balance between specialized adaptation and preservation of broad knowledge.

Domain-specific performance. Table 1 shows that MorphKA improves over LoRA SFT by 3.9% to 5.5% across individual tasks and backbone models. These gains arise from MorphKA's reconstruction of fragmented morphemes into cohesive units, especially in terminology-dense contexts where over-fragmentation hinders compositional semantics (Peng et al., 2025). MorphKA also surpasses vocabulary modification baselines (DV and SPM+ATT_EG), which add new tokens yet do not dynamically capture morphemic compositionality during adaptation (Liu et al., 2024b). Furthermore, MorphKA's internal KV-aggregation proves more effective than the inference-time merging used in Dynamic Merging (Feher et al., 2025b).

General capability retention. As shown in Table 2, MorphKA achieves the highest average domain performance, exceeding baseline by up to 4.3%. At the same time, it best preserves general capabilities, reducing catastrophic interference by 18%–22% relative to baselines. This supports our hypothesis that stabilizing latent semantic anchors enables effective integration of general capabilities beyond standard forgetting mitigation (Liu et al., 2024a), while avoiding the representational disruptions typical of vocabulary expansions.

Role of meaningful aggregation. As shown in Table 1, the Random-Grouping control—which applies the same aggregation mechanism but to arbitrary spans—performs worse than LoRA SFT on all tasks, with degradations ranging from 2.7% to 4.5%. This indicates that random aggregation worsens SKC rather than alleviating it (Feher et al., 2025b). In contrast, MorphKA outperforms this control by 7.2% to 9.9%, demonstrating that morpheme-aligned aggregation is essential for overcoming fragmentation-induced collapse, in line with findings in high-OOV settings (Balde et al., 2025b).

4.3 Analysis I: Impact of Morphemic Integrity and Fragmentation

We examine MorphKA's ability to recover cohesive semantic representations for domain-specific terms that are heavily fragmented by byte-pair encoding tokenizers. By stratifying evaluation examples according to the maximum subword piece count $p_{\max}$ (Section 4.1), we show that gains are largest where fragmentation is most severe.

Targeted improvement in high-fragmentation regimes. As shown in Table 3, MorphKA yields progressively larger improvements as $p_{\max}$ increases. In the High bucket ($p_{\max} \geq 5$), where terms are often split across multiple morpheme boundaries (e.g., "immun-o-histo-chem-istry"), MorphKA improves accuracy by 8.74% on Llama-3.1-8B and 7.96% on Qwen3-8B. These results indicate that the attention-based aggregation in MorphKA effectively consolidates dispersed subword representations into coherent units.

In contrast, the Random-Grouping control, which aggregates over arbitrary spans, degrades performance—most notably by 1.12% in the High bucket on Llama-3.1-8B. This comparison demonstrates that benefits depend on morpheme-aware boundary detection rather than mere reduction of sequence length.

Scaling with morphemic complexity. Figure 4 further illustrates how MorphKA's advantages increase with term complexity. Panel (a) shows a heatmap of accuracy gains as a function of sub-

Backbone	Top-1 Mass ↑	NAE ↓
Qwen3-8B LoRA SFT	0.38±0.09	0.81±0.07
+ MorphKA	**0.64±0.11**	**0.59±0.10**
Llama-3.1-8B LoRA SFT	0.37±0.10	0.82±0.08
+ MorphKA	**0.63±0.12**	**0.60±0.09**

Table 4: Attention metrics on high-fragmentation terms. Top-1 intra-span mass↑/NAE↓ better; means±std.

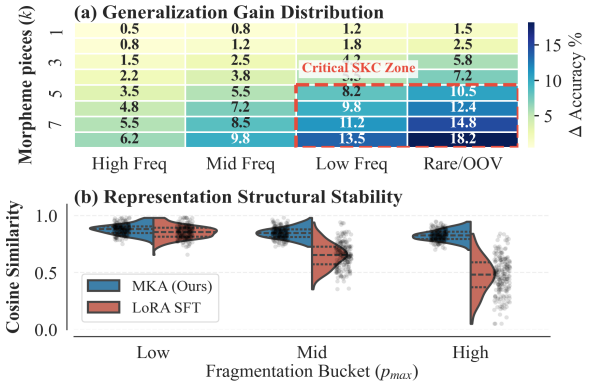

Figure 4: MorphKA improvements with increasing morphemic complexity. (a) Heatmap of accuracy gains (ΔAccuracy) by subword piece count and term frequency. (b) Violin plots of intra-term cosine similarity by fragmentation bucket ($p_{\max}$).

word piece count and term frequency (from common to rare/OOV), highlighting a region of peak improvement reaching 18.2%. Panel (b) presents violin plots of intra-term cosine similarity in hidden representations, revealing that MorphKA produces more stable and higher-similarity embeddings than LoRA SFT, especially in highly fragmented cases. Together, these analyses confirm that MorphKA enhances both task performance and representational coherence in regimes dominated by structural knowledge collapse (Peng et al., 2025).

4.4 Analysis II: Mechanism Probing

To confirm that MorphKA mitigates SKC by improving intra-term binding and representation isolation, we probe attention and hidden-state dynamics on high-fragmentation terms ($p(v) \geq 5$; 28% of evaluation terms, §4.3). These analyses build on layer-wise compositionality patterns observed in prior work (Peng et al., 2025).

Attention concentration. We examine intra-span focus by measuring the fraction of attention mass assigned to the dominant key within the term's subword indices (Top-1 intra-span mass↑) and normalized attention entropy (NAE↓). Table 4 aggregates over 600+ high-fragmentation terms

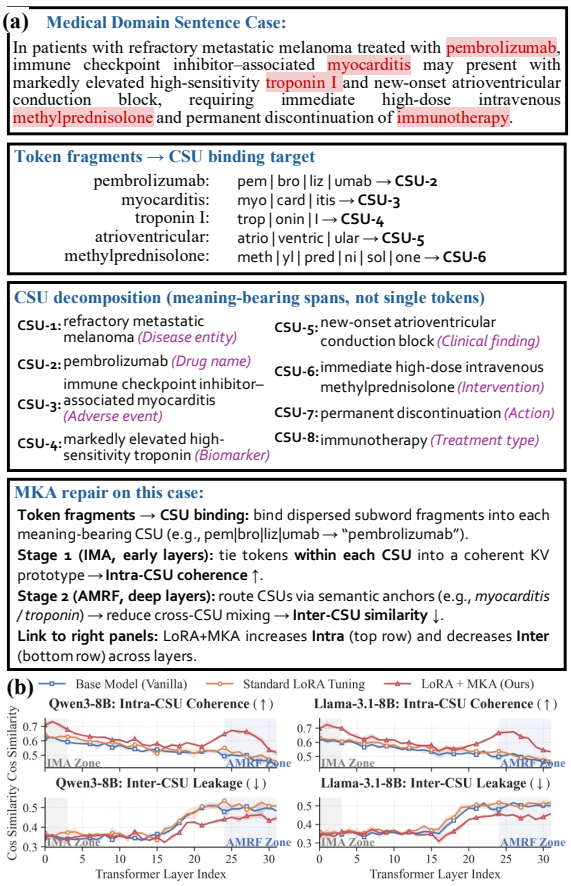

Figure 5: Mechanism evidence. (a) Biomedical case study showing token fragments, CSU decomposition (meaning-bearing spans), and MorphKA repair. (b) Layer-wise intra-CSU coherence↑ (top) and inter-CSU leakage↓ (bottom) for Qwen3-8B and Llama-3.1-8B. Shaded bands: IMA (gray) and AMRF (blue) zones.

(head-averaged). MorphKA increases Top-1 mass by 68–72% and reduces NAE by 24–27%, indicating sharper concentration and reduced evidence dispersion—consistent with effective rebinding of fragmented morphemes.

Qualitative illustration. Figure 5(a) presents a terminology-dense biomedical sentence. Subword fragments of meaning-bearing entities are highlighted, alongside MorphKA-induced CSU spans (red) and binding targets. In contrast to LoRA's diffuse patterns, MorphKA consolidates evidence onto coherent anchors, directly repairing fragmentation-induced collapse.

Figure 5(b) visualizes layer-wise trends in intra-CSU cosine similarity (higher better) and inter-CSU leakage to unrelated context (lower better). Compared to LoRA SFT (orange) and the base model (blue), MorphKA (red) consistently elevates intra-coherence and suppresses leakage, with pronounced separation emerging in AMRF

Variant	Domain Avg.	GCI Score
Qwen3-8B		
Full MorphKA	**70.55**	**83.30**
w/o Dynamic CSU (Rigid Word)	68.35	82.90
w/o IMA	66.95	82.20
w/o AMRF	68.85	82.70
w/o CSU Gate	69.30	81.80
w/o Fidelity Reg.	70.15	77.30
Llama-3.1-8B		
Full MorphKA	**67.92**	**80.65**
w/o Dynamic CSU (Rigid Word)	65.62	80.30
w/o IMA	64.32	79.50
w/o AMRF	66.22	80.10
w/o CSU Gate	66.72	79.15
w/o Fidelity Reg.	67.52	74.65

Table 5: Ablation results (means over seeds). Domain Avg.: mean across domain tasks. GCI Score: average MMLU/GSM8K.

layers. This staged improvement—from early synchronization (IMA) to deep anchor routing (AMRF)—mirrors compositional buildup across depths (Peng et al., 2025).

4.5 Ablation Study

We ablate key components of MorphKA on both Qwen3-8B and Llama-3.1-8B. Table 5 reports average domain performance (Domain Avg.) and general capability retention (GCI Score; mean of MMLU and GSM8K accuracy).

Replacing dynamic CSU discovery with rigid word-level pooling reduces Domain Avg. by 2.2%. Removing early-stage IMA causes the largest drop (3.6%), highlighting its role in stabilizing fragmented inputs. Ablating deep-stage AMRF impairs higher-order consolidation (1.7% drop). Disabling CSU gating degrades both domain and general performance by introducing noise on intact tokens. Omitting fidelity regularization maintains most domain gains but severely harms GCI (6.1% drop).

These results confirm that MorphKA's benefits arise from the targeted integration of morpheme-aware aggregation, dual-phase design, selective intervention, and fidelity constraints.

4.6 Qualitative Analysis

We illustrate MorphKA's benefits with an example from CaseHOLD, where precise interpretation of multi-word legal terms is essential. As shown in Table 6, LoRA SFT gives a vague response and fails to select the correct holding. MorphKA correctly identifies the holding on protectability of customer lists as trade secrets by maintaining compositional semantics in fragmented terms. Additional cases (including PubMedQA) are in Appendix A.5.2.

Dataset	CaseHOLD
Context (excerpt)	Colameta took customer information and proposals from Protégé to Monument; such information may constitute trade secrets under G.L.c. 266, §30 and related precedents (e.g., protection of confidential/proprietary information including customer lists).
LoRA SFT output	The passage supports protecting confidential business information, but the best-matching holding is unclear because multiple candidates mention confidentiality and trade secrets.
MorphKA output	Recognizing that customer lists may be protectable trade secrets. The excerpt directly connects customer lists and other confidential/proprietary information to the statutory definition and precedents, making this holding the most faithful match.

Table 6: CaseHOLD example. MorphKA identifies the precise holding by preserving semantic coherence in complex legal terms (highlighted).

4.7 Efficiency Analysis

MorphKA uses only ∼15.7M trainable parameters, roughly 80% fewer than vocabulary expansion methods (Liu et al., 2024b), which require ∼80M+ new embeddings. Training adds negligible memory overhead (+1.5GB on A100-80GB). At inference (FP16, batch=1), MorphKA processes 71 tokens/s on Qwen3-8B (9% slower than LoRA's) while preserving sequence length and RoPE compatibility.

Method	Extra Params (M)	t/s
LoRA SFT	0	78
Vocab Expansion	∼82.0	75
Dynamic Merging	∼45.0	84†
MorphKA	**15.7**	**71**

†Higher throughput due to sequence truncation.

Table 7: Inference efficiency on Qwen3-8B

5 Conclusion

Standard tokenizers over-fragment domain-specific terms, causing Structural Knowledge Collapse (SKC) where attention fails to reconstruct coherent concepts. We propose Morpheme-aware KV-aggregation Attention (MorphKA), a lightweight adapter that dynamically consolidates subword fragments via dual-phase aggregation (early IMA and deep AMRF) without tokenizer changes. On medical and legal benchmarks, MorphKA yields 3.9–5.5% gains over baselines (up to 8.7% on highly fragmented terms) while reducing catastrophic forgetting by 18–22%, using 80% fewer parameters than vocabulary expansion methods.

6 Limitations

While MorphKA demonstrates notable improvements, its core reliance on well-structured morpheme-level semantics introduces certain limitations. Its morpheme-based aggregation might not fully capture complex syntactic and semantic interactions in languages with less standardized morphologies (often found in low-resource languages).

References

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *Preprint*, arXiv:2502.00894.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2025a. Evaluation of LLMs in medical text summarization: The role of vocabulary adaptation in high OOV settings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22989–23004, Vienna, Austria. Association for Computational Linguistics.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2025b. Evaluation of LLMs in medical text summarization: The role of vocabulary adaptation in high OOV settings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22989–23004, Vienna, Austria. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *arXiv*.

Abhimanyu Dubey et al. 2024. The Llama 3 herd of models. *arXiv*.

Darius Feher, Ivan Vulić, and Benjamin Minixhofer. 2025a. Retrofitting large language models with dynamic tokenization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29866–29883, Vienna, Austria. Association for Computational Linguistics.

Darius Feher, Ivan Vulić, and Benjamin Minixhofer. 2025b. Retrofitting large language models with dynamic tokenization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29866–29883, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv*.

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.

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, Hong Kong, China. Association for Computational Linguistics.

Prakhar Kaushik, Alex Gain, Adam Kortylewski, and Alan Yuille. 2021. Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *arXiv*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pretraining of language models. *arXiv*. ICLR 2023.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. 2024a. More than catastrophic forgetting: Integrating general capabilities for domain-specific LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7531–7548, Miami, Florida, USA. Association for Computational Linguistics.

Chengyuan Liu, Shihang Wang, Lizhi Qing, Kun Kuang, Yangyang Kang, Changlong Sun, and Fei Wu. 2024b. 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*, pages 7442–7459, Miami, Florida, USA. Association for Computational Linguistics.

Xin Liu, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Min Zhang, Haiying Zhang, and Jinsong Su. 2021. Bridging subword gaps in pretrain-finetune paradigm for natural language generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6001–6011, Online. Association for Computational Linguistics.

Zheqi Lv, Tianyu Zhan, Wenjie Wang, Xinyu Lin, Shengyu Zhang, Wenqiao Zhang, Jiwei Li, Kun Kuang, and Fei Wu. 2025. Collaboration of large language models and small recommendation models for device-cloud recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 962–973.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Qiwei Peng, Yekun Chai, and Anders Søgaard. 2025. Understanding subword compositionality of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22524–22535, Suzhou, China. Association for Computational Linguistics.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. CVPR 2022.

Qihang Yu, Kairui Fu, Zheqi Lv, Shengyu Zhang, Xinhui Wu, Chen Lin, Feng Wei, Bo Zheng, and Fei Wu. 2026. Thinkrec: Thinking-based recommendation via LLM. In *WWW*, pages 5698–5709. ACM.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, Sao Paulo, Brazil. Association for Computing Machinery.

A Appendix

A.1 MorphKA Implementation Details

We consolidate all architecture-specific quantities of MorphKA to improve reproducibility. This appendix reports the search grids, final default values, and the operational definitions of variables that are introduced in the main text but are otherwise dispersed across Eqs. (2)–(7). Unless otherwise stated, the same default configuration is used for all experiments on both Qwen3-8B and Llama-3.1-8B. All hyperparameter search is performed on a fixed shared 5% development split under the same compute budget as the baselines.

Operational definitions. We make the following quantities explicit for reproducibility.

- **Surface cue $\eta(t_i, t_{i+1})$.** We use the canonical continuation marker induced by the SentencePiece tokenizer. Specifically, $\eta(t_i, t_{i+1}) = 1$ if token t_{i+1} does not carry the leading whitespace prefix and therefore behaves as a continuation fragment; otherwise $\eta(t_i, t_{i+1}) = 0$. This design follows the standard treatment of tokenizer boundary markers in morphology-aware tokenization and subword analysis.(Kudo and Richardson, 2018; Feher et al., 2025b)

- **Fragmentation intensity $s_{\text{frag}}(j)$.** For a diagnosed compositional semantic unit (CSU) U_j, we define
$$s_{\text{frag}}(j) = \frac{|U_j|}{\bar{m}},$$
where $|U_j|$ is the number of subword pieces in the unit and $\bar{m}$ is the average number of subwords per word computed over the domain-term lexicon. This quantity measures how severely a unit is fragmented relative to the corpus-level granularity prior.(Peng et al., 2025; Balde et al., 2025a)

- **Geometric dispersion $s_{\text{disp}}(j)$.** Let $\{\bar{h}_r\}_{r \in U_j}$ denote the stop-gradient hidden states of the subwords in CSU U_j, and let
$$c_j = \frac{1}{|U_j|} \sum_{r \in U_j} \bar{h}_r$$
be the corresponding centroid. We define
$$s_{\text{disp}}(j) = \text{Var}_{r \in U_j}\Big(1 - \cos(\bar{h}_r, c_j)\Big),$$
which captures how dispersed the subword representations are around the unit centroid. Larger values indicate weaker internal coherence and therefore a stronger need for MorphKA intervention.(Peng et al., 2025)

Search and model selection protocol. All MorphKA-specific hyperparameters in Table 8 are tuned on the same shared 5% development split for both backbones. For fairness, the search uses the same training budget as the baselines. After model selection, the resulting default configuration is frozen and reused across all domains and both model families without per-dataset retuning.

Stopping criteria. Training is run for exactly 3 epochs with AdamW. We monitor validation loss for early stopping with patience $= 2$. All remaining optimizer settings, LoRA settings, precision, batch size, gradient accumulation, sequence length, random seeds, and hardware are reported in Appendix A.2.

A.2 Hyperparameters and Training Details

The training hyperparameters are set as table 9. Hardware: 8× NVIDIA A100-80GB GPUs.

A.3 Baseline Implementation Details

We compare MorphKA to three baseline groups, all using the same LoRA configuration and training budget as MorphKA.

PEFT Controls:

- **LoRA SFT** (Hu et al., 2021): Direct supervised fine-tuning on domain-specific datasets using LoRA adapters.
- **Wise-FT** (Wortsman et al., 2021): An ensemble-based method that combines domain-adapted weights with pretrained weights ($\alpha = 0.4$).
- **DAS** (Ke et al., 2023): A continual learning approach that mitigates catastrophic forgetting through dynamic architecture search.

Vocabulary/Tokenizer Modification:

- **DV/VEGAD** (Liu et al., 2024b): We adopt domain concepts and terminology as the vocabulary to be added. For the legal domain, we use the expert-designed legal vocabulary from LawGPT[1]. For the medical domain, we prompt GPT-4 to extract the names of medicines, symptoms, and therapies from the sentences. We keep words that appear more than 100 times in the data to improve effectiveness, as increasing

[1] https://github.com/pengxiao-song/LaWGPT

Component	Parameter	Searched Grid	Default Value	Role
Eq. (2)	δ (boundary threshold)	$\{0.50, 0.60, 0.70, 0.75\}$	0.70	Boundary detection
Eq. (2)	λ	$\{0.5, 0.6, 0.7, 0.8\}$	0.70	Neural vs. surface balance
Eq. (4)	β_ℓ	$\{0.1, 0.3, 0.5\}$	0.30	IMA interpolation scale
Eq. (5)	τ_r (temperature)	$\{0.05, 0.1, 0.2\}$	0.10	Routing softness
Line 245	τ (gate bias)	$\{-0.5, 0, 0.5\}$	0	Intervention activation
Line 269	K (anchors)	$\{8, 16, 32, 64\}$	16	Semantic bottlenecks
Eq. (7)	λ_{fid}	$\{0.01, 0.05, 0.10\}$	0.05	Fidelity regularization

Table 8: Search grids and final default values for MorphKA-specific hyperparameters. Common training hyperparameters shared across all methods are reported in Appendix A.2.

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	0.03
Scheduler	Cosine annealing
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05
Target modules	$\mathrm{q}_p roj, k_p roj, v_p roj, o_p roj$
Batch size (effective)	32
Gradient accumulation steps	8
Precision	BF16
Max sequence length	4096
Seeds	42, 1337, 2024

Table 9: Common training hyperparameters.

the size of the newly added vocabulary does not invariably result in improved model performance (Liu et al., 2024b). We expand the vocabulary by 10K tokens selected via the adaptive gradient-based method described in (Liu et al., 2024b). New embeddings are randomly initialized and trained alongside LoRA.

- **SPM + ATT_EG** (Kudo and Richardson, 2018; Liu et al., 2021): We retrain a 64K SentencePiece model on combined in-domain corpora and initialize new embeddings using the attention-based embedding initialization (ATT_EG) method from (Liu et al., 2021).

Structural Interventions:

- **Dynamic Merging** (Feher et al., 2025b): Follows the hypernetwork-based dynamic merging protocol, applied at prefill and decoding stages.
- **Random-Grouping**: Applies the same aggregation mechanism as MorphKA but on random spans instead of morpheme-diagnosed CSUs.

A.4 Dataset Statistics

We use standard instruction templates from the original datasets (e.g., "Question: {question}\nOptions: ... Answer:" for MedQA). No demonstration examples are included (zero-shot instruction tuning).

Dataset	Domain	# Train	# Test
MedQA-USMLE	Medical	10,137	1,272
PubMedQA	Medical	20,000	500
CaseHOLD	Legal	26,500	3,500
BillSum	Legal	1,500	3,269

Table 10: Dataset statistics.

A.5 Additional Qualitative Evidence

A.5.1 Token-level Structural Repair Visualization

We visualize token-level significance scores induced by MorphKA (i.e., adapter/gating weights) on representative examples from two domains: PubMedQA (medical) and CaseHOLD (legal) as illustrated in Fig.6. Across both cases, MorphKA exhibits *selective intervention*: it assigns consistently higher scores to morpheme-rich, semantically salient spans (e.g., medical terms such as *antibiotic-refractory* and *pouchitis*, and legal terms such as *trade secret(s)*, *customer lists*, *confidential*, *competitor*), while down-weighting less informative fragments and function tokens. This behavior supports our claim that MorphKA mitigates fragmentation-induced structural knowledge collapse by re-centering representation mass on structurally coherent units.

A.5.2 Cross-domain Case Studies

We present representative case studies on PubMedQA and CaseHOLD in Table 11 and Table 6, respectively. Together with the token-level visualization in Figure 6, these cases suggest that selective token reweighting translates into more evidence-grounded outputs.

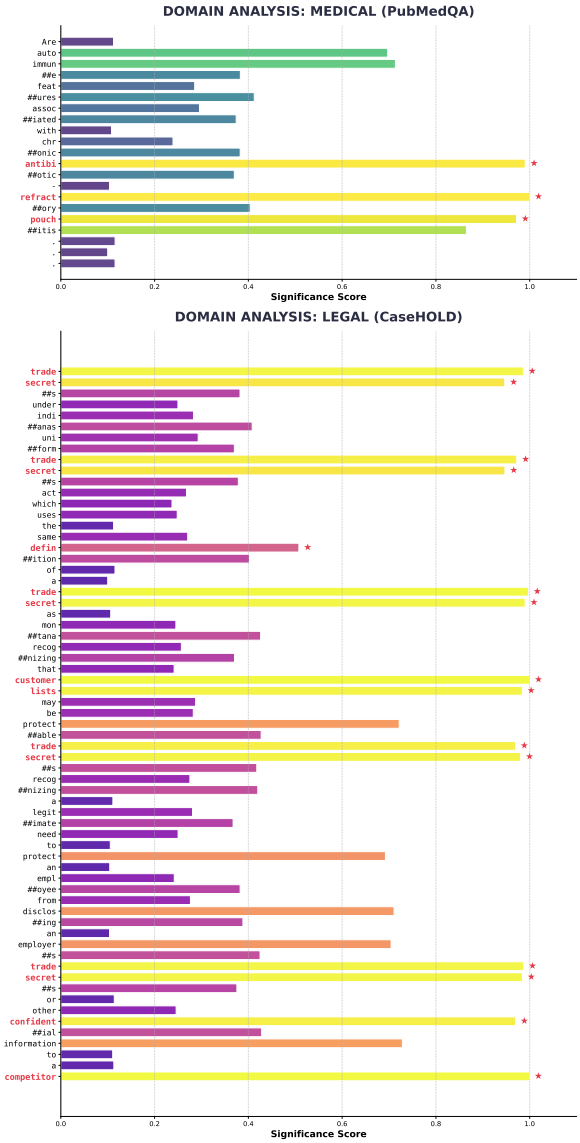

Figure 6: Token-level significance scores induced by MorphKA on PubMedQA (medical, top) and CaseHOLD (legal, bottom). Morpheme-rich domain spans receive higher scores, indicating selective intervention against fragmentation.

A.6 High/Low Activation Terms

To provide intuition for MorphKA's *selective intervention*, we list representative terms associated with high versus low gate activation (γ) across domains. High-γ terms are typically domain-specific and morpheme-rich (and thus often more prone to subword fragmentation), where MorphKA tends to allocate more adapter/gating capacity. In contrast, low-γ terms are mostly frequent function words or structurally stable tokens that carry limited domain-specific semantics; MorphKA largely bypasses intervention on these tokens to avoid unnecessary perturbation. Examples are shown in Table 12.

Dataset	PubMedQA (PMID: 25437817)
Question	Are autoimmune features associated with chronic antibiotic-refractory pouchitis?
Key evidence	(1) Elevated **microsomal antibody** positivity (OR 6.8); (2) increased **IgG4-expressing** plasma cell infiltration (OR 9.6); both are reported as **independent** risk factors for CARP.
Baseline output	**Yes.** The study suggests autoimmune comorbidities are more frequent in CARP patients, but it does not clearly specify which markers independently predict CARP.
MorphKA output	**Yes.** CARP is associated with autoimmune features: **microsomal antibody** positivity (OR 6.8) and **IgG4-positive plasma cell infiltration** (OR 9.6) are identified as **independent** risk factors, supporting a strong autoimmune-associated profile.

Table 11: PubMedQA case study. Compared with the baseline, MorphKA produces a more evidence-grounded answer by explicitly integrating morpheme-rich biomedical concepts (highlighted), consistent with mitigating fragmentation-induced structural knowledge collapse.

A.7 Prompt Examples and Case Analysis

This section provides representative prompt examples and a qualitative analysis of the MorphKA adapter's behavior. To ensure consistency during evaluation, all models are prompted using an instruction-following format with specific control tokens, as detailed in Table 14.

A.7.1 Medical Domain

In medical reasoning, critical diagnostic terms are often fragmented into nonsensical subwords (e.g., `pouchitis` → `pouch` + `itis`). Standard fine-tuning yields diffuse representations across these fragments, leading to weaker reasoning about overlapping immune features.

MorphKA addresses this by aggregating K/V vectors across these spans. As shown in Table 13, the `-itis` (inflammation) and `immuno-` prefixes receive substantially higher weights. This enables the model to correctly reason that a "pleuroperitoneal membrane defect" implies a diaphragmatic hernia, leading to the correct prediction of "Gastric fundus in the thorax" in MedQA.

A.7.2 Legal Domain

Legal terminology relies heavily on compound nouns and Latin roots that BPE frequently splits (e.g., `proprietary` → `propri` + `etary`). Baseline models often fail to link "customer lists" to the

Activation	Examples
High-γ	**PubMedQA**: antibiotic-refractory; ileal pouch-anal anastomosis; primary sclerosing cholangitis; microsomal antibody; IgG4-expressing plasma cells. **CaseHOLD**: trade secrets; confidential and proprietary business information; customer lists; uniform trade secrets act; disclosing confidential information.
Low-γ	Frequent function words and structurally stable tokens (e.g., of, the, and, is), where MorphKA is designed to bypass intervention.

Table 12: Representative terms with high vs. low CSU gate activation γ. MorphKA tends to assign higher activation to domain-specific, morpheme-rich terminology, while largely bypassing frequent and structurally stable tokens, supporting targeted mitigation of structural knowledge collapse (SKC).

Highest Adapter Weights (Significant Morphemes)	**Lowest Adapter Weights (Common/Stopwords)**
pouchitis, refractory, immunoglobulin, sclerosing, cholangitis	the, and, of, in, a
autoimmune, hyperplasia, dysplasia, immunohistochemistry	is, to, with, for, on
trade secret, proprietary, confidential, disclosure, mercatoria	it, was, be, by, as
lex, mercator, holding, jurisdiction, precedent	., „, (,), “

Table 13: Tokens receiving the highest and lowest average output weights from the MorphKA adapter. MorphKA effectively assigns higher significance to domain-specific morphemes that are typically fragmented by BPE tokenizers.

broader concept of “trade secrets” when the tokens are processed in isolation.

Through IMA and AMRF, MorphKA restores the compositional semantics of these terms. By dynamically increasing the attention weight on terms like `mercatoria` or `disclosure`, the model maintains a coherent representation of the legal context, resulting in higher accuracy for holding identification in CaseHOLD.

Dataset	Prompt (Instruction-Tuning Format)	Response
PubMedQA	<lim_startl>system You are a helpful assistant.<lim_endl> <lim_startl>user Is delayed duodenal stump blow-out following total gastrectomy for cancer ... the key to a successful duodenal stump disruption management? Context: Duodenal stump disruption remains one of the most dreadful postgastrectomy complications... <lim_endl><lim_startl>assistant	yes <lim_endl>
MedQA	<lim_startl>system You are a helpful assistant.<lim_endl> <lim_startl>user A 3900-g male infant is delivered at 39 weeks' gestation... a prenatal ultrasound showed a defect in the pleuroperitoneal membrane. Further evaluation is most likely to show which of the following findings? Options: (A) Gastric fundus in the thorax (B) Pancreatic ring (C) Hypertrophy (D) Large bowel <lim_endl><lim_startl>assistant	A <lim_endl>
CaseHOLD	<lim_startl>system You are a helpful assistant.<lim_endl> <lim_startl>user Context: ...A request to consent to search does not constitute an interrogation. See United States v. Burns, 33 M.J. 316, 320 (<HOLDING>). Which holding best follows? Options: (0) holding that a consent to search... (1) holding that persons knowledge... (2) holding that because consent is not a statement... <lim_endl><lim_startl>assistant	2 <lim_endl>
BillSum	<lim_startl>system You are a helpful assistant.<lim_endl> <lim_startl>user Summarize the following bill: SECTION 1. SHORT TITLE. This Act may be cited as the "Medical Laboratory Personnel Shortage Act of 2001"... <lim_endl><lim_startl>assistant	Medical Laboratory Personnel Shortage Act of 2001 - Amends the... <lim_endl>

Table 14: Prompt examples across domain-specific datasets. The **Prompt** column illustrates the input structure including system instructions and context, while the **Response** column shows the expected target output.