Tracing Multilingual Knowledge Acquisition Dynamics in Domain Adaptation: A Case Study of Biomedical Adaptation

Xin Zhao[A,D], **Naoki Yoshinaga**[C,B,D], **Yuma Tsuta**[D*], **Akiko Aizawa**[B,D]
[A]The University of Tokyo [B]National Institute of Informatics
[C]Institute of Industrial Science, The University of Tokyo [D]NII, LLMC
xzhao@tkl.iis.u-tokyo.ac.jp ynaga@iis.u-tokyo.ac.jp aizawa@nii.ac.jp

Abstract

Multilingual domain adaptation (ML-DA) enables large language models (LLMs) to acquire domain knowledge across languages. Despite many methods, how domain knowledge is acquired within a language and transferred across languages remains, leading to suboptimal performance, particularly in low-resource settings. This work examines the learning dynamics of LLMs during ML-DA. Because prior ML-DA studies often train and evaluate on datasets with mismatched knowledge coverage, we propose **AdaXEval**, an adaptive evaluation method that constructs multiple-choice QA datasets from the same bilingual domain corpus used for training, thereby enabling direct analysis of multilingual knowledge acquisition. Through continual training of LLMs with diverse data recipes, we track how LLMs acquire domain facts and pinpoint the **loss shielding** mechanism behind the knowledge memorization and generalization in domain adaptation. Our experiments on multilingual LLMs reveal that cross-lingual transfer remains challenging. The code is released.

MLDA-Eval

1 Introduction

Large language models (LLMs) trained on general-domain corpora often struggle in specialized domains with limited data (Jang et al., 2022b,a). Domain adaptation addresses this by continually training LLMs on domain-specific data to acquire expertise (Jiang et al., 2024a; Laï-king and Paroubek, 2024; Xie et al., 2024b). Although data augmentation and cross-lingual transfer improve adaptation efficiency in low-resource settings, where data for the target domain and language is scarce (Fang et al., 2023; Gao et al., 2024), the mechanisms underlying effective domain knowledge acquisition and transfer remain insufficiently understood.

Understanding domain adaptation requires examining how LLMs acquire and internalize facts from domain data. Previous studies show that factual knowledge accumulates through repeated exposures, influenced by data frequency and model scale (Chang et al., 2024; Zhao et al., 2024a; Liu et al., 2025). In multilingual settings, additional complexity arises as equivalent facts may be encoded differently across languages (Zhao et al., 2024a; Mondal et al., 2025). However, most analyses focus on predefined relational probes rather than domain-specific facts with complex structures and specialized terminology. Moreover, the link between training data and acquired knowledge remains underexplored, which is crucial to optimizing domain adaptation strategies.

Our work investigates knowledge acquisition in domain adaptation from a mechanistic perspective. Specifically, we study how domain facts are *memorized* and *generalized* during the continual-training process on the domain corpus, across both *intralingual* (within a language) and *interlingual* (across languages) settings, and identify key factors that facilitate effective knowledge acquisition and transfer. To achieve the goal, we focus on three questions:

RQ1: How to effectively evaluate the domain knowledge acquisition from diverse aspects?

RQ2: What is the mechanism behind the transformation from training data to knowledge?

RQ3: What factors are critical to achieve cross-lingual transfer?

Current methods for evaluating knowledge acquisition rely mainly on public benchmarks (Singhal et al., 2022; Jiang et al., 2025) or training loss analysis (Zucchet et al., 2025; Liu et al., 2025). However, the knowledge attainable from the training data, especially in low-resouce settings, often mismatches the knowledge evaluated by benchmarks, while training loss does not solely reflect target domain knowledge. This train-test discrepancy complicates the assessment of acquired knowledge.

*Present address: Fixstars Corporation

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5739–5760

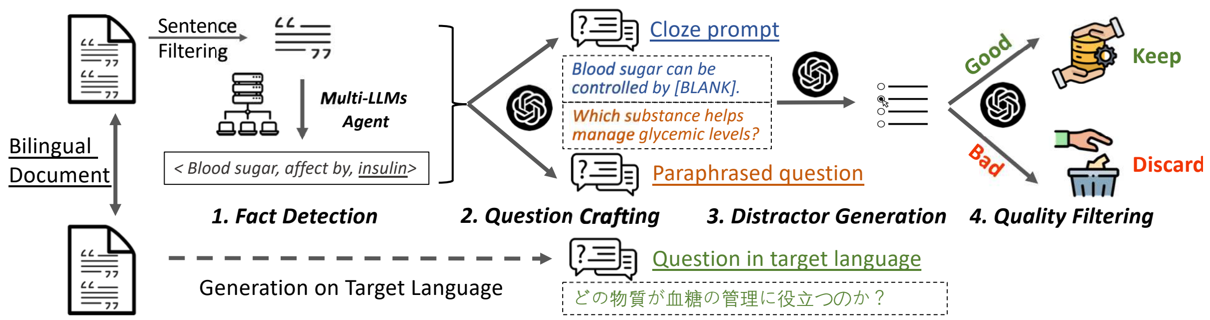


Figure 1: Overview of AdaXEval, a pipeline to adaptively generate domain knowledge evaluation datasets.

To address these gaps (**RQ1**), we propose **AdaXEval**, an adaptive domain knowledge evaluation pipeline (§ 2). AdaXEval automatically constructs multiple-choice datasets to evaluate knowledge *memorization*, intralingual generalization (*paraphrase*), and interlingual generalization (*cross-lingual transfer*). It operates on either a monolingual or a bilingual domain corpus. Human annotation from multiple perspectives confirms that AdaXEval provides reliable evaluation.

We next investigate how training data is transformed into LLMs' knowledge (**RQ2**). Specifically, we conduct a case study on biomedical domain adaptation using multiingual LLMs (LLM-jp, 2024) (§ 3). We begin with monolingual continual training on English and Japanese bilingual corpus constructed from J-STAGE.[1] By evaluating training checkpoints with AdaXEval-generated datasets, we observe a gradual knowledge acquisition for cloze queries and paraphrases. However, the LLMs struggle to achieve cross-lingual transfer. Further analysis reveals that knowledge is acquired as losses of correct options are shielded from rapid growth due to the model's overfitting to training data, a phenomenon we term **loss shielding**. This acquisition eventually plateaus as training leads to overfitting, causing loss across all options to increase. Examining losses on perturbed training data reveals that LLMs readily overfit to fixed token sequences in the training data, even under minor noise (§ 4).

Finally, we study factors affecting cross-lingual transfer of domain knowledge (**RQ3**) through **multilingual continual training** with diverse data recipes (§ 5). We examine translation and romanization as additional data sources. We find that cross-lingual token overlap in related domains is critical for effective knowledge transfer. However, cross-lingual transfer remains challenging even with strong alignment signals such as translation.

[1] https://www.jstage.jst.go.jp/browse/-char/ja

2 Knowledge Acquisition Evaluation

To effectively evaluate domain adaptation in low-resource scenarios, we propose AdaXEval, an adaptive pipeline for generating evaluation datasets. AdaXEval generates multiple-choice datasets for assessing knowledge memorization, generalization, and cross-lingual transfer using advanced LLMs.

2.1 AdaXEval Pipeline

AdaXEval is an adaptive evaluation pipeline that evaluates domain knowledge acquisition by generating evaluation datasets directly from the training corpus, ensuring evaluated facts stay aligned with the training data. The pipeline includes four steps: fact detection, question crafting, distractor generation and quality filtering, as illustrated in Figure 1.

Step 1: Fact Detection AdaXEval first detects sentences that may contain domain facts through a two-step strategy: named-entity-recognization (NER)-based sentence filtering and multi-agent fact detection. First, domain-specific NER tools and linguistic heuristics are used to identify sentences in the training corpora that contain multiple named entities. Next, we design Chain-of-Thought (CoT) instructions to detect sentences containing domain facts from the filtered sentences, and extract triples in the format ⟨*subject, relation, object*⟩ as the reference for question crafting. Specifically, AdaXEval employs a multi-LLM agent for fact detection and triple extraction, estimating the overall confidence of the outputs, and adapting the top-confident extraction result to improve evaluation reliability.

Step 2: Queries Crafting Given the factual sentence and referenced triple, we first prompt the LLM to generate reliable domain factual triples (*e.g.*, ⟨*blood sugar level, can be controlled by, insulin*⟩). As complex domain knowledge cannot be easily formalized into named entities or predefined relations, subjects and objects are preferably named entities, though descriptive phrases are acceptable. AdaXEval uses advanced LLMs to generate diverse question-answer pairs measuring three knowledge acquisition abilities.

1) Knowledge memorization uses a cloze prompt with [BLANK] as the placeholder for the object (*e.g.*, *Blood sugar level can be controlled by [BLANK].*). Given the original sentence and refined triple as reference, we prompt the LLM to generate a cloze question that closely matches the original sentence to assess memorization exclusively.

2) Intralingual generalization assesses LLMs' ability to acquire knowledge using linguistic expressions that vary from those in the training corpus. We design CoT instructions to let LLMs paraphrase the cloze queries into question-like style questions where different vocabulary is encouraged (*e.g.*, *Which substance helps manage glycemic levels in the body?*)

3) Interlingual generalization measures how learned facts can be transferred across languages. We consider two options: **1)** Adapting AdaXEval to a bilingual domain corpus containing languages X and Y, using the paraphrased dataset from language X to evaluate cross-lingual transfer capabilities in Y. **2)**. Do translation using strong LLMs.

Step 3: Distractor Generation AdaXEval then generates three plausible yet incorrect answer options that remain topically related but unambiguously wrong, while explicitly instructing the advanced LLM to avoid surface-level cues such as sequence length. The distractor generation instruction could be found in our codebase.

Step 4: Quality Filtering Finally, AdaXEval uses the LLM to filter low-quality multiple-choice QA instances that fail to meet the requirements in the above steps.

2.2 Evaluation metrics

For each evaluation dataset, we follow Gao et al. (2023) to compute the average cross-entropy loss over the target tokens of possible answers and select the one with the highest probability as the final answer. We use prediction accuracy as the metric for knowledge acquisition. See Appendix A.2 for the mathematical formulation of the metrics.

2.3 Experimental Setup

Domain corpus: Our study investigates biomedical domain adaptation in English–Japanese as a case study. Specifically, we utilize the J-STAGE, an English-Japanese bilingual biomedical corpus (see § 3.1), as the data source for both model training and AdaXEval generation.

Processing Step	English	Japanese
Sampled abstracts	10,000	10,000
Splitted sentences	81,770	71,661
Sentences with multiple entities	45,390	40,762
Triple extraction	4840	3926
Cloze queries generation	4840	3926
Paraphrases generation	4840	3926
After Quality Filtering	3236	2553

Table 1: Dataset statistics at each step of AdaXEval.

Details of generation: We randomly sampled 10,000 bilingual documents to generate the evaluation dataset. Abstracts were split into sentences, and sentences containing fewer than two biomedical entities were filtered out. For fact detection, we used three open-source LLMs[2] from different families to assess the confidence that the sentence is factual in its language. We summed the confidence scores from the three LLMs and retained sentences with combined scores greater than 2 (maximum 3). Finally, GPT-4.1 was used to generate cloze queries, paraphrases, and three distractors for each instance. The statistics of datasets is shown in Table 1. See Appendices A.1 and B for details.

Human evaluation: To assess the quality of our generated datasets, we conduct a comprehensive human evaluation across four key components of the knowledge extraction and question generation pipeline, including triple extraction quality evaluation, cloze prompt evaluation, paraphrased question evaluation, and distractor quality evaluation. Overall, the evaluation result indicates that AdaXEval is able to generate high-quality evaluation data, meeting the requirements for assessing diverse knowledge acquisition abilities. See Appendix B for evaluation results and dataset examples, and Appendix G for the human evaluation guideline.

3 Tracing Knowledge Acquisition

This section examines how domain adaptation unfolds during training and explores how knowledge is acquired from training data in a monolingual setting. We perform continual training on English and Japanese using a bilingual biomedical dataset.

[2]Open-source LLMs were used for local inference, as massive candidate sentences incur substantial computational costs.

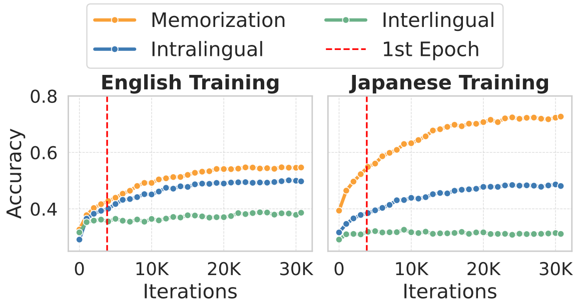


Figure 2: Dynamic knowledge acquisition evaluation during monolingual continual training.

3.1 Experimental Setup

Data preparation: We utilize a subset of the J-STAGE corpus[3] consisting of Japanese research papers with some abstracts translated into English. It contains 614,444 Japanese and 404,643 English biomedical paper abstracts, aligned as one-to-one bilingual pairs, which serve as the source data for AdaXEval. To enhance domain adaptation and support fine-grained analysis, we apply instruction pretraining as a data augmentation baseline (Jiang et al., 2024b). Biomedical instructions are generated using rule-based mining patterns (Cheng et al., 2024b) and LLM-based question-answer generation (Jiang et al., 2024b), and are combined with the original documents for continual training. Refer to Appendix E.1 for details of the dataset generation.

Training setup: We adopt llm-jp-3-13B (LLM-jp, 2024), a multilingual LLM, as the base model for pretraining, due to its strong performance in English as well as Japanese. From the constructed corpus, we sample 0.5B tokens per language and train llm-jp-3-13B for four epochs. Training hyperparameters are detailed in Appendix E.2.

3.2 Tracing Performance Dynamics

Figure 2 reports AdaXEval results for English and Japanese monolingual training. The results indicate that domain knowledge is gradually acquired during training in both languages.

Memorization: Accuracy increases from 36.1% to 59.6% (**+23.5%**) in English and from 45.7% to 75.9% (**+30.2%**) in Japanese. Note that direct cross-lingual comparison is not strictly fair due to differences in evaluation instances.

Intralingual generalization: On the paraphrased datasets, accuracy increases from 32.6% to 54.7% (**+22.1%**) in English and from 34.9% to 53.8% (**+18.9%**) in Japanese. The English gain matches the memorization gain, while the Japanese gain is about 10% lower, indicating language-dependent differences in intralingual generalization.

[3]Access to the dataset is restricted by the J-STAGE license.

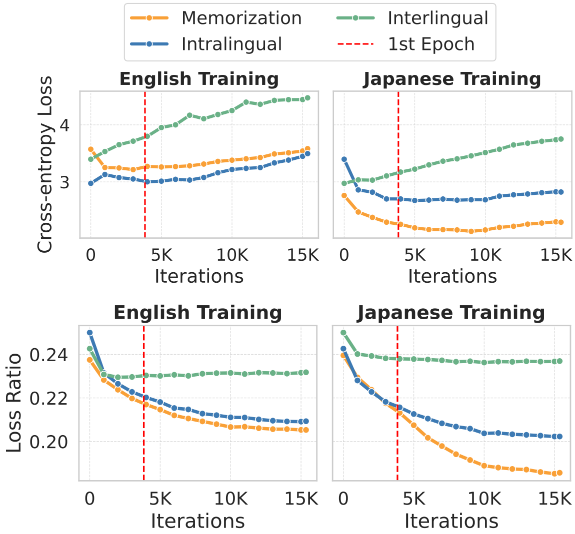


Figure 3: Loss (above) and loss ratio dynamics (below) during monolingual continual training.

Interlingual generalization: Figure 2 reveals that monolingual training results in limited cross-lingual knowledge transfer, yielding only **3.6%** improvement in English-to-Japanese and **3.1%** improvement in Japanese-to-English transfer.

3.3 Knowledge Acquisition via Loss Shielding

To elucidate how knowledge is acquired during training, we analyze sequence loss on the evaluation data. This loss directly drives predictions on our multiple-choice datasets (§ 2.2) and reflects the model's generation behavior, with lower-loss sequences being more likely to be generated.

(1) Training overfits to data, but *loss shielding* drives memorization. We compute sequence loss for queries paired with correct answers in the evaluation dataset obtained by AdaXEval. Figure 3 (above) shows the loss trajectory across training checkpoints for English and Japanese. On the cloze prompt dataset, the loss decreases early in training but rises later, indicating that model overfits the training corpus. However, Figure 2 reveals that memorization accuracy continues to improve until the 10K-th iteration. To investigate this divergence, we measure the ratio of the correct-sequence loss to the total loss across all four options. As shown in Figure 3 (below), this ratio mirrors the accuracy trend, suggesting that knowledge can still be mem-

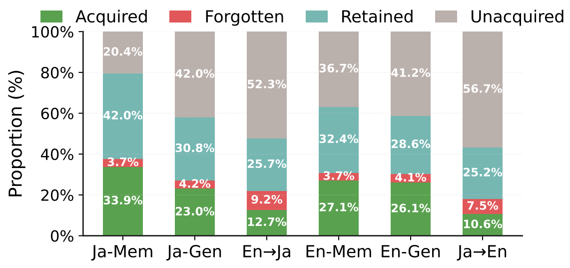


Figure 4: Instance state transitions after training.

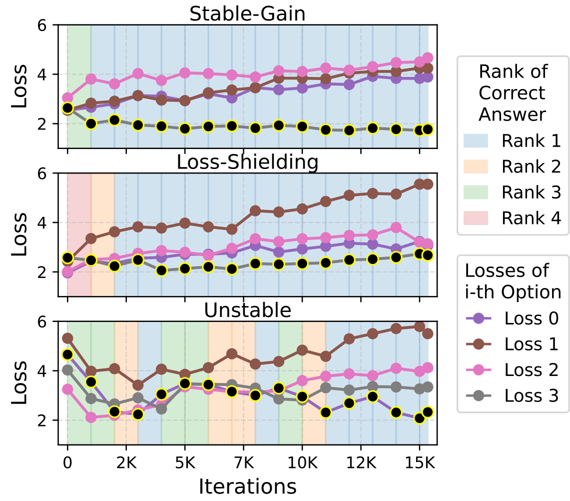


Figure 5: Acquired instances with different loss shapes. The line with bright circles indicates the correct answer.

orized under overfitting. Correct sequences are shielded from rapid loss growth, a phenomenon we term *loss shielding*. This ratio also helps explain the gap between cloze prompts and paraphrases, particularly in Japanese, where a large loss ratio gap corresponds to the accuracy shown in Figure 2.

(2) A trade-off exists between knowledge acquisition and forgetting. For each instance, we examine state transitions before and after training based on the loss. Figure 4 shows the proportions of instances that are retained, acquired, forgotten, or remain unacquired. Forgetting is limited in monolingual evaluations, including both memorization and intralingual generalization. In contrast, cross-lingual transfer exhibits a notable increase in forgotten cases, offsetting gains from acquired knowledge. This suggests that while training in one language can introduce transferable knowledge, it can degrade performance in others due to forgetting.

(3) Instance-level analysis reveals diverse knowledge acquisition patterns. We examine the loss trajectories of instances acquired after training, analyzing all four options. The correct answers exhibit three distinct patterns: they either decrease steadily (Stable-Gain), increase while remaining lower than incorrect options (Loss-Shielding), or exhibit unstable behavior. Figure 5 illustrates three representative examples of these patterns.

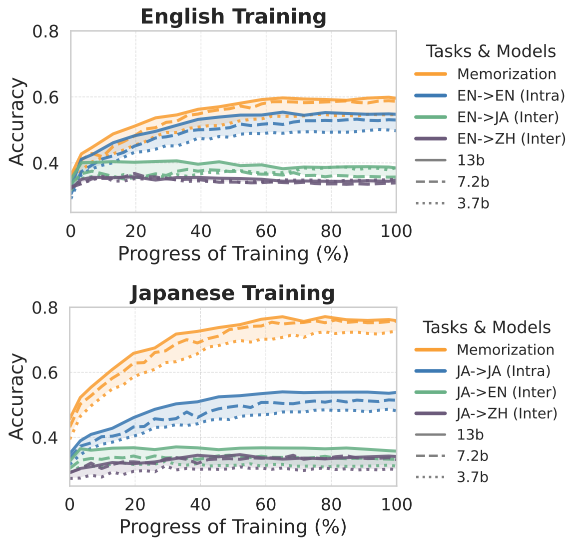


Figure 6: Dynamic knowledge acquisition evaluation on LLMs of various sizes in different target languages, where LLMs are trained on J-STAGE data for 4 epochs.

3.4 Validation Experiments

To further validate the generalization our finding of loss shielding phenomenon, we conduct validation experiments on different LLMs and language pairs.

Experiment setup: We extend English/Japanese domain training to two additional LLM-jp models, llm-jp-3-3.7B and llm-jp-3-7.2B, following the similar settings as the default. See Table 5 in Appendix E.2 for details. Training were conducted on J-STAGE data for 4 epochs. To verify whether the above observations apply to other language pairs, we created Chinese evaluation datasets by translating the intralingual English and Japanese data into Chinese using the GPT-4.1 API. These datasets were used to evaluate cross-lingual transfer from English/Japanese to Chinese.

Accuracy: Figure 6 reports the accuracy dynamics for both Japanese and English training. The results reveal consistent patterns of domain knowledge acquisition across models and language pairs: strong memorization, moderate intralingual, and weak interlingual generalization. Additionally, model size plays a critical role in acquiring new knowledge, with larger models achieving higher performance across all aspects of knowledge acquisition.

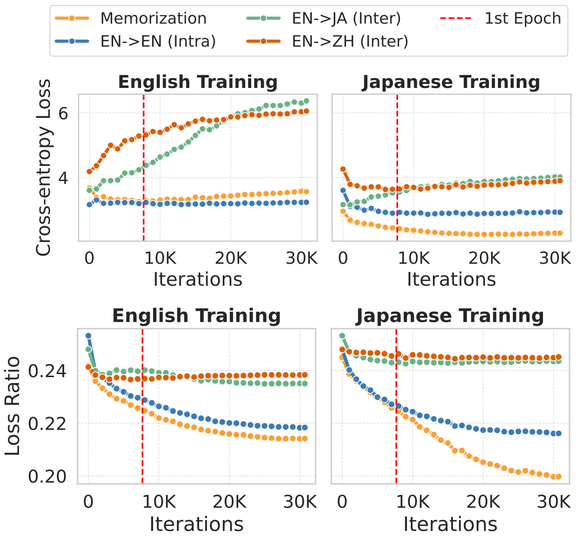


Figure 7: Loss (above) and loss ratio dynamics (below) during monolingual continual training with llm-jp-3-3.7B.

Loss shielding: Finally, we examine the loss dynamics to determine whether loss shielding occurs across diverse training settings. Result for the 3.7B model are shown in Figure 7, reconfirming the loss shielding phenomenon. Loss dynamics on other model sizes are provided in Appendix C.

4 From Training Data to Knowledge

This section examines how models acquire knowledge about unseen queries from training data. When domain knowledge appears in novel linguistic forms, the model's loss increases (Figure 3). Although loss shielding allows the model to continue acquiring knowledge even as the loss for such sequences rises, its effect diminishes once the loss grows too large. Understanding what sustains or breaks loss shielding is therefore essential for developing more robust training strategies.

To this end, we introduce controlled perturbations into the training data using various noise injection methods and track the resulting loss dynamics during training. Specifically, we randomly sample 2,000 sequences from Japanese monolingual training data and apply perturbations at both the token and sequence levels.

4.1 Token-level Perturbation

Token sequences are perturbed after tokenization with different methods and degrees,[4] as follows:

[4]We set $X = 2^1, \ldots, 2^5$ and $Y = 2^0, \ldots, 2^4$. Synonyms are collected from the Japanese WordNet 2.0 (Bond and Kuribayashi, 2023). See Appendix D for implementation details.

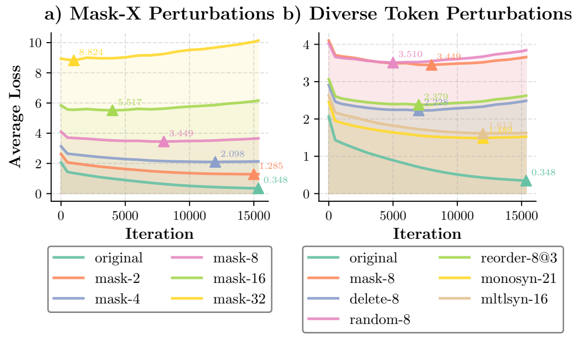


Figure 8: Loss dynamics of token-level perturbation during monolingual continual training.

mask-X replaces X% tokens with <unk> token.

random-X replaces X% tokens with randomly sampled tokens from the tokenizer vocabulary.

delete-X deletes X% tokens.

reorder-X@Y reorders tokens to achieve an edit distance equal to X% of the sequence length, with swaps restricted to a window of size Y.

monosyn-X replaces with Japanese synonyms.

mltlsyn-X replaces with English synonyms.

(1) More token edits cause larger loss harm. Taking *mask-X* as an example, Figure 8(a) illustrates the loss dynamics across edited sequences compared to their original counterparts. Triangles (▲) mark the *overfitting onset*, the point where minimum loss is attained before overfitting induces an upward trend. The results indicate that while the loss of original sequences decreases steadily, masking tokens cause a substantial initial loss increase and accelerate the onset of overfitting. Analyses of other perturbation patterns show the same trend. See Appendix D for additional results.

(2) Loss sensitivity varies with lexical and structural perturbations. Figure 8(b) illustrates the loss variations across sequences with 8% of tokens perturbed using diverse methods. Semantically aligned modifications (*monosyn* and *mltlsyn*), exhibit the least impact on loss, followed by structural alterations (*reorder* and *delete*) that maintain the original tokens. Perturbations introducing irrelevant tokens (*mask* and *random*) inflict the greatest harm, significantly increasing the initial loss and accelerating the onset of overfitting.

4.2 Sentence-level Perturbation

To simulate more realistic noise, we further perturb target sequences at the sentence level, as follows:

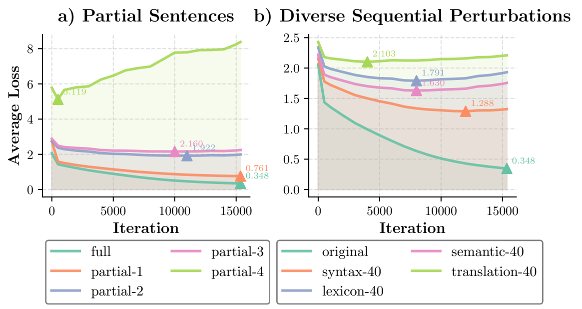


Figure 9: Loss dynamics of sequence-level perturbation during monolingual continual training.

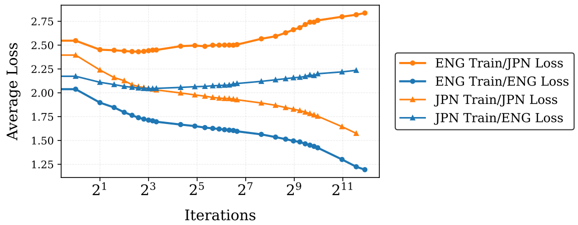


Figure 10: Losses on English and Japanese training data during monolingual continual training.

partial-a splits sentences in each document into four segments, then select the a-th segment.

syntax-X: rewrites $X\%$ sentences, modifying only syntax without changing vocabulary.

lexicon-X: rewrites $X\%$ sentences, modifying only vocabulary without changing syntax.

semantic-X rewrites $X\%$ sentences, allowing both syntactic and lexical changes.

translation-X translates $X\%$ sentences.

(1) Losses depend strongly on prior context. With the *partial-a* strategy using partial sentences, Figure 9(a) shows that later sentences evaluated in isolation incur higher loss. This suggests that losses depend heavily on prior context, limiting knowledge acquisition and calling for a more context-agnostic learning approach.

(2) Retaining vocabulary boosts loss robustness. Figure 9(b) shows loss dynamics for sentence-level perturbation methods applied to 40% of sentences, including paraphrasing and translation. Among these, syntax rewriting provides the strongest loss shielding, followed by semantic rewriting. This is because semantic paraphrasing requires fewer vocabulary replacements than lexicon paraphrases. These results highlight the need for lexicon-focused paraphrasing during training to enhance models' ability to generalize knowledge across diverse test inputs, consistent with the observations in § 4.1.

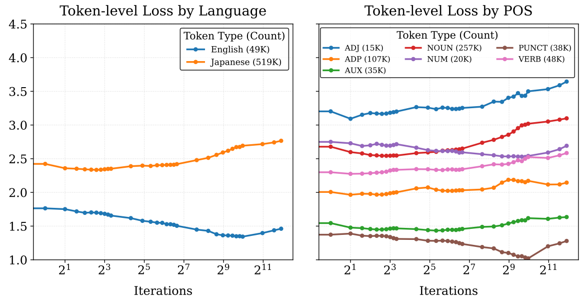


Figure 11: Japanese per-token loss by English training.

5 Bridging Gap across Languages

This section explores factors that enhance cross-lingual transfer in domain adaptation. We first analyze early loss dynamics in monolingual continual training (§ 5.1). We then perform multilingual domain adaptation to identify the factors (§ 5.2).

5.1 What Helps Cross-lingual Transfer?

We study how cross-lingual transfer occurs by tracking the loss on target-language documents in models trained on source-language data. As the AdaXEval interlingual loss rises (Figure 3), tracing loss on evaluation data becomes unreliable. Instead, we examine training document losses across languages as indirect evidence. Specifically, we sample 1,000 documents for each language and measure per-token loss under two monolingual settings, as shown in Figure 10. Since cross-lingual transfer converges rapidly within the early training stage, we analyze loss dynamics of the first epoch. We observe that in-language loss consistently decreases, whereas cross-language loss decreases only during the initial iterations, reflecting rapid but short-lived interlingual generalization.

To investigate which tokens contribute to the initial loss reduction, we measure per-token cross-entropy loss within the sequences. Specifically, we examine two linguistic characteristics: language and part-of-speech (POS). Tokens are classified by the language (Japanese or English) and grouped by their POS. Figure 11 shows token-level loss dynamics on Japanese documents during English training. Language-based analysis indicates that English tokens embedded in Japanese documents benefit from English training. POS-based analysis supports this findings, as only numerical (NUM) and punctuation (PUNCT) tokens, both present in the English corpus, exhibit notable loss reductions. These results suggest that tokens occurring in the training corpus experience greater learning gains.

5.2 Cross-lingual Transfer Enhancement

In this section, we explore the factors in the training corpus that enhance cross-lingual transfer during domain adaptation. Because monolingual training provides limited transfer across languages, we adopt multilingual domain adaptation. Specifically, we examine which types of multilingual corpora facilitate knowledge transfer in domain adaptation.

5.2.1 Multilingual Continual Training

To investigate the factors that facilitate cross-lingual transfer, we construct a series of multilingual corpora and evaluate their effectiveness. Each training corpus is composed of two parts: a **knowledge injection corpus** $\mathcal{C}_{\mathcal{K}}$ and a **cross-lingual transfer enhancement corpus** $\mathcal{C}_{\mathcal{T}}$. The $\mathcal{C}_{\mathcal{K}}$ contains the target knowledge expressed exclusively in the source language X, while $\mathcal{C}_{\mathcal{T}}$ does not introduce new knowledge in the target domain; it serves only to establish linguistic connections between source language X and target language Y. We assess cross-lingual knowledge transfer by measuring the model's ability to acquire knowledge in language Y, using the evaluation metrics introduced in § 2.2. For analysis in this section, we fix the source language X as English and the target language Y as Japanese. See Appendix F for the evaluation on reverse transfer direction.

Cross-lingual transfer enhancement corpora: We construct diverse corpora to enhance cross-lingual transfer using *translation* and *romanization*. These strategies build token-level connections between English and Japanese, which romanization requiring significantly less effort for data collection. As baselines, we consider an empty cross-lingual corpus ($\mathcal{C}_{\mathcal{T}} = \emptyset$, **Monolingual** and a strong domain-specific baseline using J-STAGE Japanese data, with documents related to the AdaXEval evaluation removed (**Medical-Japanese**).

To examine which translation data are most effective for domain adaptation, we prepare three types of bilingual corpora and use them to generate translation instructions as $\mathcal{C}_{\mathcal{T}}$:

JParaCrawl (Balanced-Translation) is an English–Japanese web-crawled corpus covering diverse domains (Morishita et al., 2022).

ASPEC (Science-Translation) is a multilingual corpus of academic paper abstracts across various scientific fields (Nakazawa et al., 2016); documents in the medical and chemical domains are excluded to distinguish them from the target medical domain.

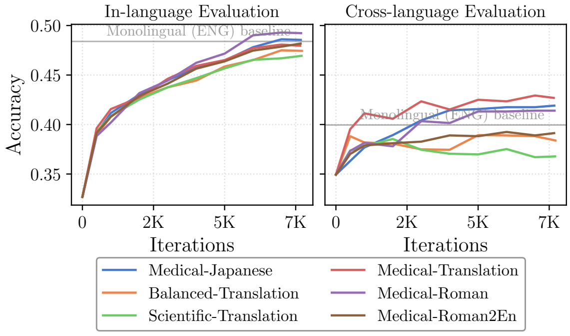


Figure 12: Enhancing cross-lingual transfer using diverse multilingual domain corpora.

J-STAGE (Medical-Translation) represents the closest domain to our target; documents included in the AdaXEval evaluation datasets are filtered out to avoid contamination.

To evaluate the romanization strategy, we construct a medical romanization dataset (**Medical-Roman**). We convert J-STAGE Japanese text to romaji using `cutlet`,[5] an open-source tool for romanization. We then generate romanization instructions to link the Latin script of English with the Japanese script (kanji, hiragana, etc.). Finally, we create translation instructions between romanized Japanese and English based on J-STAGE (**Medical-Roman2En**), which serves as a comparison group.

Details of training: Focusing on knowledge transfer from English to Japanese, we prepare two corpora, $\mathcal{C}_K$ and $\mathcal{C}_T$, each containing 0.5B tokens, except for the **Monolingual** baseline. The English J-STAGE training data serve as $\mathcal{C}_K$. For $\mathcal{C}_T$, we select each of the seven candidate datasets. We then combine the two 0.5B-token corpora into a single 1B-token corpus, shuffle it, and train llm-jp-3-13B for one epoch. See Appendix E for training details.

5.2.2 Cross-lingual Transfer Evaluation

(1) Target domain corpora enhance cross-lingual transfer. Figure 12 shows the accuracy dynamics of multilingual domain adaptation, comparing two baselines with three translation datasets and two romanization datasets. Among these datasets, only the corpora constructed from domain-specific datasets (**Medical-***) surpass the **Monolingual** baseline, with only **Medical-Translation** outperforming both baselines. These results indicate that effective cross-lingual transfer of domain knowledge relies on domain-specific signals, as general methods fail to yield comparable improvements.

[5] https://github.com/polm/cutlet

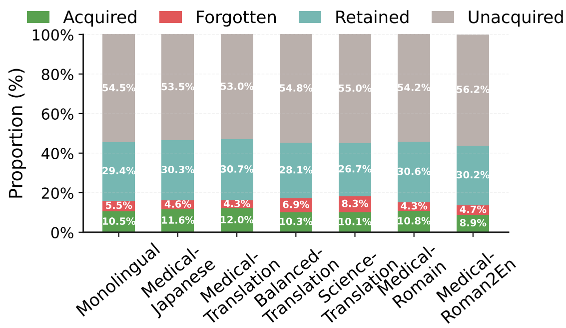


Figure 13: State transitions by cross-lingual transfer.

Figure 13 further illustrates the state transitions induced by cross-lingual transfer. Corpora that yield performance gains both increase acquired instances and reduce forgotten ones, whereas the other datasets perform worse on both aspects. Finally, stable performance across all recipes in the in-language evaluation shows that adding multilingual corpora unrelated to the target knowledge does not hinder in-language knowledge acquisition.

(2) Achieving effective cross-lingual transfer of domain knowledge is challenging. Although both **Medical-Translation** improves performance, the gains (around 3%) are limited given the doubled training cost and extra dataset construction effort. This underscores the need for more efficient methods of cross-lingual domain knowledge transfer.

6 Related Work

6.1 Domain Knowledge Evaluation

Domain knowledge is typically evaluated using public benchmarks (Xie et al., 2024a; Jiang et al., 2025), though such benchmarks are often unavailable for low-resource languages or specialized domains. General benchmarks also fail to fully assess intralingual or interlingual generalization, which requires paraphrase or translation datasets. Additionally, the misalignment between training data and benchmark knowledge coverage means evaluation does not always accurately reflect acquired knowledge. Some studies, in contrast, use loss reduction as evidence of knowledge acquisition (Chang et al., 2024; Zucchet et al., 2025; Liu et al., 2025), but loss does not necessarily correspond to domain-specific knowledge that users care about.

6.2 Mechanisms of Knowledge Acquisition

LLMs have been widely studied for their ability to store and retrieve factual knowledge (Petroni et al., 2019; Hernandez et al., 2024; Chang et al., 2024; Zhao et al., 2024b), sparking interest in understanding how such knowledge is encoded and accessed (Dai et al., 2022; Wang et al., 2022; Niu et al., 2024; Zhao et al., 2025). Recent work has explored the dynamics of knowledge acquisition by analyzing intermediate checkpoints during training. Chang et al. (2024) shows that LLMs accumulate factual knowledge through repeated exposures, with recall likelihood increasing gradually after each encounter. Key factors influencing acquisition include fact frequency, model scale, and batch size (Zhao et al., 2024a; Liu et al., 2025).

Although cross-lingual transfer has garnered significant attention for efficient knowledge acquisition across languages, transferring factual knowledge across languages remains a challenge (Zhao et al., 2024a; Liu et al., 2025). Facts expressed in different languages may be stored as distinct representations (Chen et al., 2024; Zhao et al., 2024a; Mondal et al., 2025), and only certain relational types transfer effectively (Zhao et al., 2024a; Liu et al., 2025). Moreover, existing studies focus on predefined relational facts (Zhao et al., 2024a; Liu et al., 2025), which inadequately capture complex domain-specific knowledge. The connection between training data and acquired knowledge remains underexplored, which is crucial for designing optimal training recipes.

7 Conclusions

In this paper, we explore how LLMs acquire and transfer domain knowledge across languages. We introduce **AdaXEval**, an adaptive evaluation pipeline that automatically generates datasets to assess domain knowledge across memorization, intralingual generalization, and interlingual generalization. Using AdaXEval and multilingual LLMs (LLM-jp, 2024), we conduct a case study focusing on multilingual biomedical domain adaptation. We analyze the training dynamics and find that knowledge acquisition is driven by **loss shielding**, wherein overfitting raises losses on irrelevant representations faster than on relevant ones. Through **perturbation analysis**, we reveal the sensitivity of LLMs to training data, offering practical insights for more robust training paradigms. We also identified key factors facilitating cross-lingual transfer through **controlled multilingual continual training**, emphasizing the importance of cross-lingual tokens in closely related domains.

8 Limitations

While this study provides novel insights into the mechanisms of bilingual domain adaptation, several limitations remain. First, our analysis is based on a single case study involving biomedical adaptation. Although we conduct generalization evaluation on other language pairs and LLMs with different model sizes, the generalizability of the findings to other domains, and model architectures remains to be validated. Future research should therefore extend the investigation to a broader range of large language models, including those trained under different configurations, as well as to diverse domain and linguistic contexts.

Second, our evaluation relies on loss-based quantitative metrics to characterize the dynamics of knowledge acquisition and forgetting. While these measures provide a consistent and interpretable framework for tracking learning behavior, they do not fully capture how such knowledge is manifested in model outputs. A complementary qualitative and text-level evaluation, examining generated outputs and factual correctness, will be necessary to bridge the gap between internal training dynamics and externally observable performance.

References

Francis Bond and Takayuki Kuribayashi. 2023. The Japanese Wordnet 2.0. In *Proceedings of the 12th Global Wordnet Conference*, pages 179–186, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? In *Advances in Neural Information Processing Systems*, volume 37, pages 60626–60668. Curran Associates, Inc.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024a. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, Miami, Florida, USA. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024b. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yihao Fang, Xianzhi Li, Stephen Thomas, and Xiaodan Zhu. 2023. ChatGPT as data augmentation for compositional generalization: A case study in open intent detection. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 13–33, Macao. -.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *Proceedings of the first Conference on Language Modeling*.

Ryosuke Ishigami. 2025. DeepSeek-R1-Distill-Qwen-32B-Japanese.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022a. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022b. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations*.

Junfeng Jiang, Fei Cheng, and Akiko Aizawa. 2024a. Improving referring ability for biomedical language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6444–6457, Miami, Florida, USA. Association for Computational Linguistics.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. JMedBench: A benchmark for evaluating Japanese biomedical large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935, Abu Dhabi, UAE. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024b. Improving domain adaptation through extended-text reading comprehension. *Preprint*, arXiv:2401.07284.

Mathieu Laï-king and Patrick Paroubek. 2024. Pretraining data selection for biomedical domain adaptation using journal impact metrics. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 363–369, Bangkok, Thailand. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. 2025. Tracing multilingual factual knowledge acquisition in pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2121–2146, Suzhou, China. Association for Computational Linguistics.

LLM-jp. 2024. LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. *Preprint*, arXiv:2407.03963.

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

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Social Computing Lab. 2023. Mednern-cr-ja (revision 13dbcb6).

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024a. FinBen: a holistic financial benchmark for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024b. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

Xin Zhao, Zehui Jiang, and Naoki Yoshinaga. 2025. Neuron empirical gradient: Discovering and quantifying neurons' global linear controllability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21446–21477, Vienna, Austria. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024a. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024b. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13186–13214, Miami, Florida, USA. Association for Computational Linguistics.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. 2025. How do language models learn facts? dynamics, curricula and hallucinations. In *Proceedings of the second Conference on Language Modeling*.

A AdaXEval Implementation Details

A.1 Generation Details

This section outlines the process used to generate the AdaXEval dataset from J-STAGE documents.

(1) Factual sentence filtering: We use the open-source NLP tool HanLP[6] for Japanese sentence segmentation, and scispaCy[7] (Neumann et al., 2019), which provides a full spaCy pipeline for scientific and biomedical texts, for English sentence segmentation. Subsequently, we perform biomedical named entity recognition (NER) on each sentence and filter out sentences containing fewer than two named entities. Specifically, for Japanese medical documents, we employ MedNERN-CR-JA[8] (Social Computing Lab, 2023) , a model specialized for the Japanese medical domain. For English texts, we use the "en_ner_bionlp13cg_md" model provided by scispaCy for biomedical entity extraction.

(2) Domain triple extraction: In the next step, we use a multi-LLM agent to first judge whether the given sentence contains biomedical facts and extract them if it does. We carefully design CoT instruction with few-shot examples to instruct each LLM in the agent to generate a structural output, containing three fields:

- `factuality`: answer `yes` or `no`. If `yes`, the model should output the `triple`; Otherwise, it outputs `None` to the `triple` field.
- `triple`: A nested JSON object with three fields: subject, relation, and object representing the extracted fact. Please note that if factuality is false, make sure this field is null
- `reason`: A brief explanation for why the sentence was or wasn't considered factual, referring to the criteria provided. This is included to improve the model's reasoning ability.

The models' confidence in judging the factuality is measured by the probability of `yes` token out by the `factuality` field.

For our experiments, we utilize three strong open-source LLMs for both English and Japanese, opting for open-source models to avoid the high computational cost of commercial APIs. For **English biomedical triple extraction**, we employ Qwen-32B (Qwen Team, 2025),[9] DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025),[10] and Llama-3.3-70B-Instruct (Grattafiori et al., 2024).[11] For **Japanese biomedical triple extraction**, we use Qwen-32B, llm-jp-3.1-8x13b-instruct4,[12] and Llama-3.3-Swallow-70B-Instruct-v0.4.[13] For each sentence, we aggregate the confidence scores from the three models and retain sentences with at least two models predicting a `yes` label. We then apply a heuristic method to select the final triple from the three candidates.

(3) Generation of queries and distractors: We then use the extracted triples and their corresponding context sentences as input to instruct the strong close-source LLM, GPT-4.1, for generating queries

[6] https://github.com/hankcs/HanLP

[7] https://github.com/allenai/scispacy

[8] https://github.com/sociocom/MedNERN-CR-JA

[9] https://huggingface.co/Qwen/Qwen3-32B

[10] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B

[11] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[12] https://huggingface.co/llm-jp/llm-jp-3.1-8x13b-instruct4

[13] https://huggingface.co/tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4

and distractors. We carefully design the prompts for both English and Japanese. The final generation is recorded and included in the final AdaXEval evaluation dataset.

(4) Quality filtering: Finally, we use LLM to conduct a two-step quality filtering. First, we introduce a rigorous annotation instruction for assessing the quality of automatically generated biomedical knowledge questions. Given a sentence, its knowledge triple, a fill-in-the-blank query, and a question-style paraphrase, we evaluate three aspects:

- The fidelity and clarity of the cloze prompt,
- The semantic equivalence and self-containment of the paraphrased question, and
- The correctness and plausibility of the answer and distractors.

Then, we craft the annotation instruction to let LLM decide whether the document in the target languages contains the knowledge matching the created instances in the source languages. The annotation is to confirm the quality of the AdaXEval dataset in providing interlingual generalization evaluation.

A.2 Evaluation Metrics

We follow (Gao et al., 2023) to compute the average cross-entropy loss over the target tokens of possible answers and select the one that has the highest generation possibility as the final answer. Specifically, for loss calculation of cloze queries, we use tokens before the `[BLANK]` as context and compute loss on the following tokens. For paraphrases, we treat the question as context and measure only the loss of answer tokens. We use prediction accuracy as the metric for knowledge acquisition. **Formulation:** Let the model be $p_\theta(\cdot \mid \cdot)$. Each dataset $\mathcal{D}$ contains pairs (q, a), where q is either a *cloze prompt* or a *paraphrase question*, and $a = (a_1, \ldots, a_m)$ is the tokenized answer sequence (*e.g.*, "insulin"). For cloze prompts, the prompt contains a special token `[BLANK]`, and for paraphrases, the question is a natural question. We denote by c the context tokens and by $s = (s_1, \ldots, s_n)$ the evaluation sequence whose loss we measure.

Cloze queries: For a cloze prompt such as:
`"[BLANK] can be used to control blood sugar level."`
the evaluation sequence s is the full completion after the `[BLANK]`, i.e.,

$$s = (a_1, \ldots, a_m, r_1, \ldots, r_k),$$

where $a_1, \ldots, a_m$ are answer tokens ("insulin"), and $r_1, \ldots, r_k$ are the remainder tokens (" can be used to control blood sugar level"). The average cross-entropy loss is defined as

$$\mathcal{L}_{\text{cloze}}(q, a) = -\frac{1}{n} \sum_{t=1}^{n} \log p_\theta(s_t \mid c, s_{<t}).$$

Paraphrase queries: For a paraphrase question such as
`"Which substance helps manage glycemic levels?"`
we take the entire question tokens as context c, and the evaluation sequence is only the answer tokens:

$$s = (a_1, \ldots, a_m).$$

The loss is computed as

$$\mathcal{L}_{\text{para}}(q, a) = -\frac{1}{m} \sum_{t=1}^{m} \log p_\theta(a_t \mid c, a_{<t}).$$

Prediction and accuracy: For multiple-choice answers $\mathcal{A} = \{a^{(1)}, \ldots, a^{(K)}\}$, we select the candidate with the lowest loss (equivalently, highest likelihood):

$$\hat{a} = \arg\min_{a \in \mathcal{A}} \mathcal{L}(q, a).$$

Finally, the knowledge acquisition metric is accuracy:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}|} \sum_{(q,a) \in \mathcal{D}} \mathbf{1}[\hat{a} = a].$$

B AdaXEval Dataset Details

We randomly sampled 10,000 parallel documents to generate the evaluation dataset. The number of instances after each step in AdaXEval generation is shown in Table 1.

B.1 Human Evaluation

To assess the quality of our generated datasets, we conduct a comprehensive human evaluation across four key components of the knowledge extraction and question generation pipeline. The annotation is conducted by the first author, who has a language background of both Japanese and English.

Table 2: Human evaluation results for knowledge acquisition datasets generated from the biomedical J-STAGE corpus.

Evaluation Metric	Japanese	English
Detected Sentence (Factuality)	2.72/3	2.84/3
Cloze prompt (Faithfulness)	2.84/3	2.89/3
Paraphrase (Fluency)	2.94/3	2.96/3
Paraphrase (Diversity)	2.48/3	2.62/3
Paraphrase (Factuality)	2.86/3	2.86/3
Paraphrase (Inter-Factuality)	1.68/2	1.76/2
Distractor (Plausibility)	2.35/3	2.16/3
Distractor (Incorrectness)	2.89/3	2.97/3

For the annotation that requires specific domain knowledge, the author uses advanced LLMs, such as ChatGPT or Claude, as an assistant for annotation. See the full human evaluation guideline documented in Appendix G.

(1) Cloze prompt evaluation checks the faithfulness of the generated prompt to the original sentence structure.

(2) Paraphrases evaluation is conducted on four dimensions: fluency and grammaticality, linguistic diversity in reformulation, factual correctness to the original sentence in the source language, and inter-lingual factual correctness using the corresponding documents in the target language.

(3) Distractor quality is measured through plausibility within the domain and apparent incorrectness relative to the original context.

Each metric employs structured scoring rubrics with scales ranging from 0-2 or 0-3, enabling systematic assessment of dataset quality across multiple linguistic and semantic dimensions. We randomly sample 50 instances for each language and conduct human evaluation following the guidance above. Our evaluation results in Table 2 show that AdaXEval is capable of generating high-quality evaluation data, meeting the requirements for assessing knowledge memorization as well as intralingual and interlingual generalization evaluation.

B.2 Examples

We randomly sample 10 examples from AdaXEval for both English and Japanese and display them in Table 3 (Japanese) and Table 4 (English).

C Validation Experiments

In this section, we report the accuracy dynamics and the loss dynamics under diverse training settings, including using three base models with different model sizes and different language pairs. See details of experiment setup in Section 3.4.

We report the loss dynamics and the loss ratio dynamics on the three models separately in Figure 7 (3.7B), Figure 14 (7.2B) and Figure 15 (13B). We observe loss shielding across all settings, confirming its generalizability as a mechanism for knowledge acquisition.

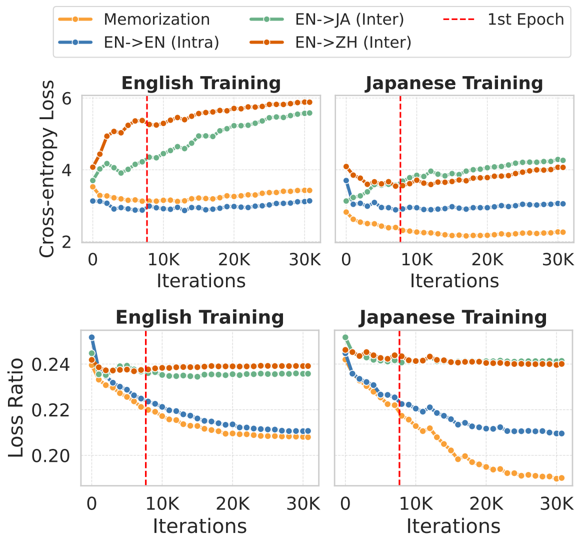


Figure 14: Loss dynamics and loss ratio dynamics with llm-jp-3-7.2B as base model.

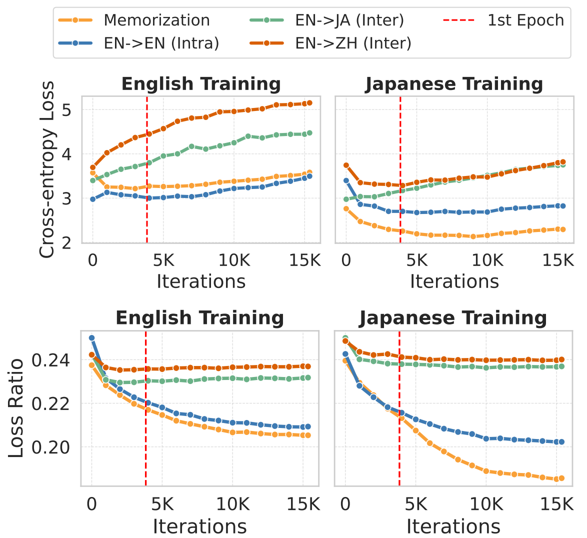


Figure 15: Loss dynamics and loss ratio dynamics with llm-jp-3-13B as base model.

D Sentence-level Perturbation Analysis

In this section, we introduce the detailed settings for sequence perturbation experiments and report the additional results.

Sentence	**cloze prompt**	**Paraphrase**	**Options** (Answer)
その結果, Ca-HApのカルシウムおよびリンの溶解は多くの有機酸水溶液中, pH4 6.5の範囲で下記の式に従うことがわかった (ただし, リンゴ酸, 酒石酸, クエン酸を除く)。	その結果、[BLANK]は多くの有機酸水溶液中、pH4～6.5の範囲で下記の式に従うことがわかった（ただし、リンゴ酸、酒石酸、クエン酸を除く）。	多くの有機酸水溶液中、pH4～6.5の範囲で下記の式に従うことがわかったのは何の溶解ですか？	A. Ca-HApのマグネシウムおよび鉄の溶解 B. Ca-HApの亜鉛および銅の溶解 C. Ca-HApのカルシウムおよびリンの溶解 D. Ca-HApのナトリウムおよびカリウムの溶解
それとともに, β-lactamase や amino-glycoside acetyltransferase (AAC)などによる抗菌薬の不活化や DNA gyrase の変化による抗菌薬親和性の減少, さらにbiofilm形成による抗菌薬の低浸透などが組み合わさり多剤耐性化する.	β-lactamaseやaminoglycoside acetyltransferase (AAC)による抗菌薬の不活化、DNA gyraseの変化による抗菌薬親和性の減少、さらに biofilm 形成による抗菌薬の低浸透などが組み合わさり[BLANK]する。	β-lactamase や AAC による抗菌薬の不活化、DNA gyraseの変化、biofilm形成などが組み合わさることで生じる現象は何ですか？	A. 単剤耐性化 B. 多剤耐性化 C. 抗菌薬感受性の増加 D. 抗菌薬の副作用増強
悪性腫瘍に伴う血液凝固能亢進状態により脳卒中をきたす病態はTrousseau症候群として知られている.	悪性腫瘍に伴う血液凝固能亢進状態により脳卒中をきたす病態は[BLANK]として知られている。	悪性腫瘍に伴う血液凝固能亢進状態が原因で脳卒中を発症する病態は何と呼ばれていますか？	A. Goodpasture症候群 B. Trousseau症候群 C. 抗リン脂質抗体症候群 D. Lambert-Eaton症候群
根部よりの吸収は化合物の疎水性 (logP) と負の相関性を示した.	[BLANK]は化合物の疎水性 (logP) と負の相関性を示した。	化合物の疎水性 (logP) と負の相関性を示すのはどのような吸収ですか？	A. 根毛からの吸収 B. 茎部よりの吸収 C. 葉部よりの吸収 D. 根部よりの吸収
このことからp53はH2AXリン酸化には関与しておらず、脱リン酸化やさらに下流の因子と関わってあると考えられる。	このことから[BLANK]は H2AX リン酸化には関与しておらず、脱リン酸化やさらに下流の因子と関わってあると考えられる。	H2AXリン酸化に関与していないと考えられるタンパク質は何ですか？	A. DNA-PK B. CHK2 C. p53 D. ATM
Solitary fibrous tumor(SFT)は,間葉系細胞由来の稀な腫瘍である.	[BLANK]は、間葉系細胞由来の稀な腫瘍である。	間葉系細胞由来の稀な腫瘍として知られているのは何ですか？	A. 神経膠腫 (glioma) B. リンパ腫 (lymphoma) C. Solitary fibrous tumor (SFT) D. 扁平上皮癌 (squamous cell carcinoma)
1990年代半ばに開発された脱窒菌法により,硝酸イオンのδ15N, δ18Oを微量で同時に測定できるようになった。	1990年代半ばに開発された[BLANK]により、硝酸イオンのδ15N, δ18Oを微量で同時に測定できるようになった。	硝酸イオンのδ15N およびδ18O を微量で同時に測定することを可能にした方法は何ですか？	A. 分光光度法 B. 脱窒菌法 C. イオンクロマトグラフィー法 D. ガスクロマトグラフィー質量分析法
遠隔転移は悪性腫瘍が全身化した状態で,治療の原則は薬物療法である。	遠隔転移は悪性腫瘍が全身化した状態で、治療の原則は[BLANK]である。	悪性腫瘍が全身化した状態である遠隔転移の治療の原則として用いられるのは何ですか？	A. 放射線療法 B. 免疫療法 C. 外科的切除 D. 薬物療法
門脈圧亢進症における消化管壁内粘膜下 A-Vanastomosis (A-VA) 開大増加に伴う循環亢進状態に関しては, 食道胃静脈瘤・門脈圧亢進症性胃症の発症に直接関連する病態として, 多くの検討がなされてきている.	門脈圧亢進症における消化管壁内粘膜下A-Vanastomosisの開大増加は[BLANK]に直接関連する病態として、多くの検討がなされてきている。	門脈圧亢進症における消化管壁内粘膜下A-Vanastomosisの開大増加が直接関連する病態として発症するのは何ですか？	A. 肝性脳症および肝腎症候群の発症 B. 胆道閉塞および胆石症の発症 C. 膵炎および十二指腸潰瘍の発症 D. 食道胃静脈瘤および門脈圧亢進症性胃症の発症
BLT1を介したシグナルはMyD88の遺伝子発現を誘導することで腸内細菌からの自然免疫シグナルを増強し,形質細胞の細胞増殖を促進することで経口ワクチン抗原に対する抗原特異的IgA産生を促進する作用があることがわかり,経口ワクチンの成立に必須の分子機構であることが明らかになった.	BLT1を介したシグナルはMyD88の遺伝子発現を誘導することで[BLANK]を増強し、形質細胞の細胞増殖を促進することで経口ワクチン抗原に対する抗原特異的IgA産生を促進する作用があることがわかった。	BLT1を介したシグナルがMyD88の遺伝子発現を誘導することで増強するのは何のシグナルですか？	A. 腸管上皮細胞の増殖シグナル B. 抗原提示細胞による炎症性サイトカインシグナル C. 腸内細菌からの自然免疫シグナル D. ウイルス感染による獲得免疫シグナル

Table 3: Samples of Japanese AdaXEval dataset.

D.1 Details of Perturbation

For *monosym@X* and *monosym@X* that require collecting synonyms from WordNet 2.0, we only conduct Japanese-to-English replacement. Specifically, we first tokenize the Japanese sequence by sudachipy[14], a Japanese morphological analyzer, and get the POS tags. Then we filter out tokens with stop words and the POS tags that are not "普通名詞", "固有名詞", "サ変接続", "形容動詞語幹", "動詞一般" to avoid introducing noisy words.

[14] https://github.com/WorksApplications/SudachiPy

Furthermore, the paraphrasing and translation are done by requesting GPT-4.1.

D.2 Perturbation Results

We show all the token-level perturbation results in Figure 16 and sentence-level perturbation result in Figure 17.

E Domain Adaptation Training Details

E.1 Training Data Generation

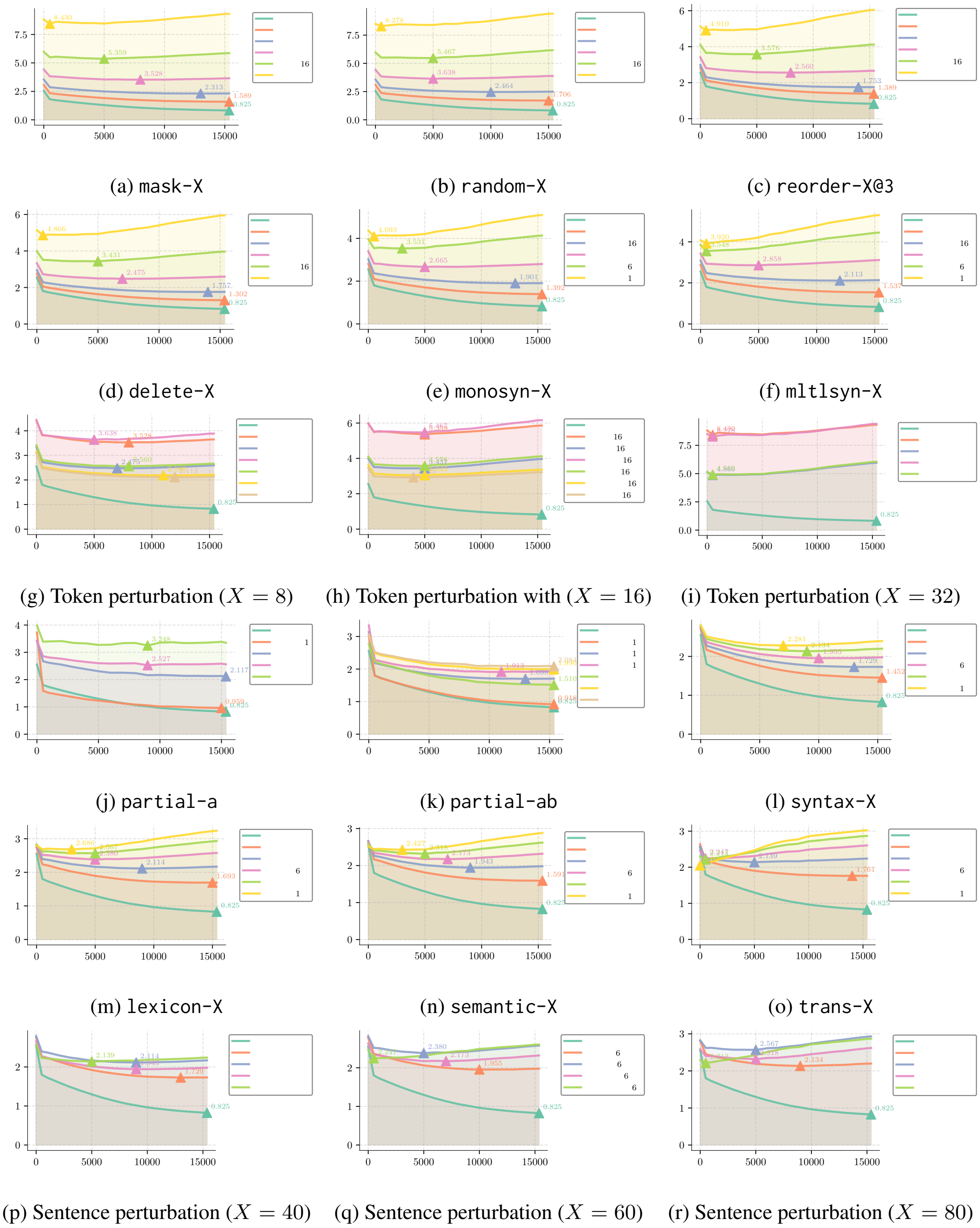
In this study, to address the scarcity of bilingual domain corpora and enhance domain understand-

Sentence	cloze prompt	Paraphrase	Options (Answer)
Fe2O3-SiO2 particles, which removes 112S and COS in hot coal gas, are prepared.	Fe2O3-SiO2 particles, which remove [BLANK] in hot coal gas, are prepared.	Which contaminants in hot coal gas are targeted for removal by Fe2O3-SiO2 particles?	A. NH3 and HCN B. SO2 and NOx C. H2S and COS D. CO2 and CH4
The antimicrobial activity of lomefloxacin against gram-positive bacteria was inferior to those of ofloxacin and gentamicin and comparable to that of chloramphenicol.	The antimicrobial activity of lomefloxacin against gram-positive bacteria was inferior to those of [BLANK] and comparable to that of chloramphenicol.	Against gram-positive bacteria, lomefloxacin shows lower antimicrobial activity than which other antibiotics?	A. ciprofloxacin and ampicillin B. ofloxacin and gentamicin C. vancomycin and clindamycin D. erythromycin and tetracycline
Basal cell adenoma is a rare type of salivary gland tumor.	[BLANK] is a rare type of salivary gland tumor.	Which rare type of tumor can occur in the salivary glands?	A. Adenoid cystic carcinoma B. Pleomorphic adenoma C. Basal cell adenoma D. Mucoepidermoid carcinoma
The common clinical signs were fever and hepatosplenomegaly.	The common clinical signs were [BLANK].	Which clinical signs are most frequently observed?	A. fever and hepatosplenomegaly B. jaundice and ascites C. rash and lymphadenopathy D. cough and chest pain
Phospholipids were found to be classified into three groups : (1) a lipid deactivating the glycolipid by strong hydrogen bond (phosphatidic acid analog), (2) a lipid likely to distribute the glycolipid rather homogeneously by weak hydrogen bond (phosphatidylglycerol and phosphatidylinositol analogs), and (3) a lipid enhancing the activity of a glycolipid by electrostatic effect (phosphatidylserine, phosphatidylcholine, and phosphatidylethanolamine analogs).	Phospholipids were found to be classified into three groups: (1) a lipid deactivating the glycolipid by strong hydrogen bond (phosphatidic acid analog), (2) a lipid likely to distribute the glycolipid rather homogeneously by weak hydrogen bond (phosphatidylglycerol and phosphatidylinositol analogs), and (3) [BLANK].	Which group of phospholipids enhances the activity of glycolipids by electrostatic effect?	A. lipids inhibiting glycolipid synthesis by covalent modification (ceramide analogs) B. lipids deactivating glycolipids by strong hydrogen bond (phosphatidic acid analog) C. lipids enhancing glycolipid activity by electrostatic effect (phosphatidylserine, phosphatidylcholine, and phosphatidylethanolamine analogs) D. lipids distributing glycolipids homogeneously by weak hydrogen bond (phosphatidylglycerol and phosphatidylinositol analogs)
Furthermore, the dried ARC, which was dehydrated in the presence of saccharides, can be recovered by dispersion of the powdered ARC in water.	The dried ARC, which was dehydrated in the presence of saccharides, can be recovered by [BLANK].	What process allows the recovery of dried ARC that was dehydrated with saccharides?	A. heating the powdered ARC in ethanol B. exposing the powdered ARC to ultraviolet light C. dispersion of the powdered ARC in water D. mixing the powdered ARC with organic solvents
About half of the reported cases of acanthosis nigricans are accompanied by various kinds of malignant neoplasms, mostly adenocarcinomas of the digestive system.	About half of the reported cases of acanthosis nigricans are accompanied by [BLANK].	What condition is present in about half of the reported cases of acanthosis nigricans?	A. various kinds of malignant neoplasms, mostly adenocarcinomas of the digestive system B. infectious diseases, such as tuberculosis and hepatitis C. autoimmune disorders, mainly lupus and rheumatoid arthritis D. benign skin tumors, primarily lipomas and fibromas
Suppressed cellular immunity, malignancy, diabetes mellitus and history of antibiotic usage are significant predisposing factors for the development of esophageal candidiasis.	[BLANK] are significant predisposing factors for the development of esophageal candidiasis.	Which conditions are considered important risk factors for developing esophageal candidiasis?	A. Chronic hypertension, obesity, hyperlipidemia, and smoking B. Suppressed cellular immunity, malignancy, diabetes mellitus, and history of antibiotic usage C. Asthma, seasonal allergies, eczema, and vitamin D deficiency D. Alcohol abuse, liver cirrhosis, renal failure, and hypothyroidism
Improved hydrogen sulfide removal is necessary for this apparatus to be applied to measurement of biogas produced by anaerobic digestion, since hydrogen sulfide influences catalysis.	Improved hydrogen sulfide removal is necessary for this apparatus to be applied to measurement of biogas produced by anaerobic digestion, since hydrogen sulfide influences [BLANK].	What process is affected by the presence of hydrogen sulfide in biogas measurement apparatus?	A. oxidation B. fermentation C. catalysis D. photosynthesis
The activity of ACC (1-aminocyclopropane-1-carboxylic acid) oxidase and the rate of ethylene production increased rapidly during fruit ripening at 20°C.	The activity of ACC (1-aminocyclopropane-1-carboxylic acid) oxidase and the rate of ethylene production increased rapidly during [BLANK].	During which process at 20°C do ACC oxidase activity and ethylene production rate increase rapidly?	A. fruit ripening at 20°C B. leaf senescence at 20°C C. flowering at 20°C D. seed germination at 20°C

Table 4: Samples of English AdaXEval dataset.

ing, we employ two data augmentation strategies: regex-based pattern mining and LLM-based QA generation. Both approaches yield instruction-like sequences, which we mix with raw corpora. Following prior work (Cheng et al., 2024a), we adopt an instruction-pretraining strategy for continual domain adaptation.

(1) Regex-based pattern mining: (Jiang et al., 2024b) verified that by transforming raw corpora into reading comprehension texts, continual training can consistently enhance performance across various tasks in different domains. We adopt a similar strategy by analyzing the training corpora and mining regex patterns to automatically create instruction-style data. Furthermore, to increase data diversity, we prepare ten instruction templates for each type of reading comprehension text. For



(a) mask-X (b) random-X (c) reorder-X@3

(d) delete-X (e) monosyn-X (f) mltlsyn-X

(g) Token perturbation ($X = 8$) (h) Token perturbation with ($X = 16$) (i) Token perturbation ($X = 32$)

(j) partial-a (k) partial-ab (l) syntax-X

(m) lexicon-X (n) semantic-X (o) trans-X

(p) Sentence perturbation ($X = 40$) (q) Sentence perturbation ($X = 60$) (r) Sentence perturbation ($X = 80$)

Figure 16: The loss dynamics over all perturbation patterns on Japanese sequences.

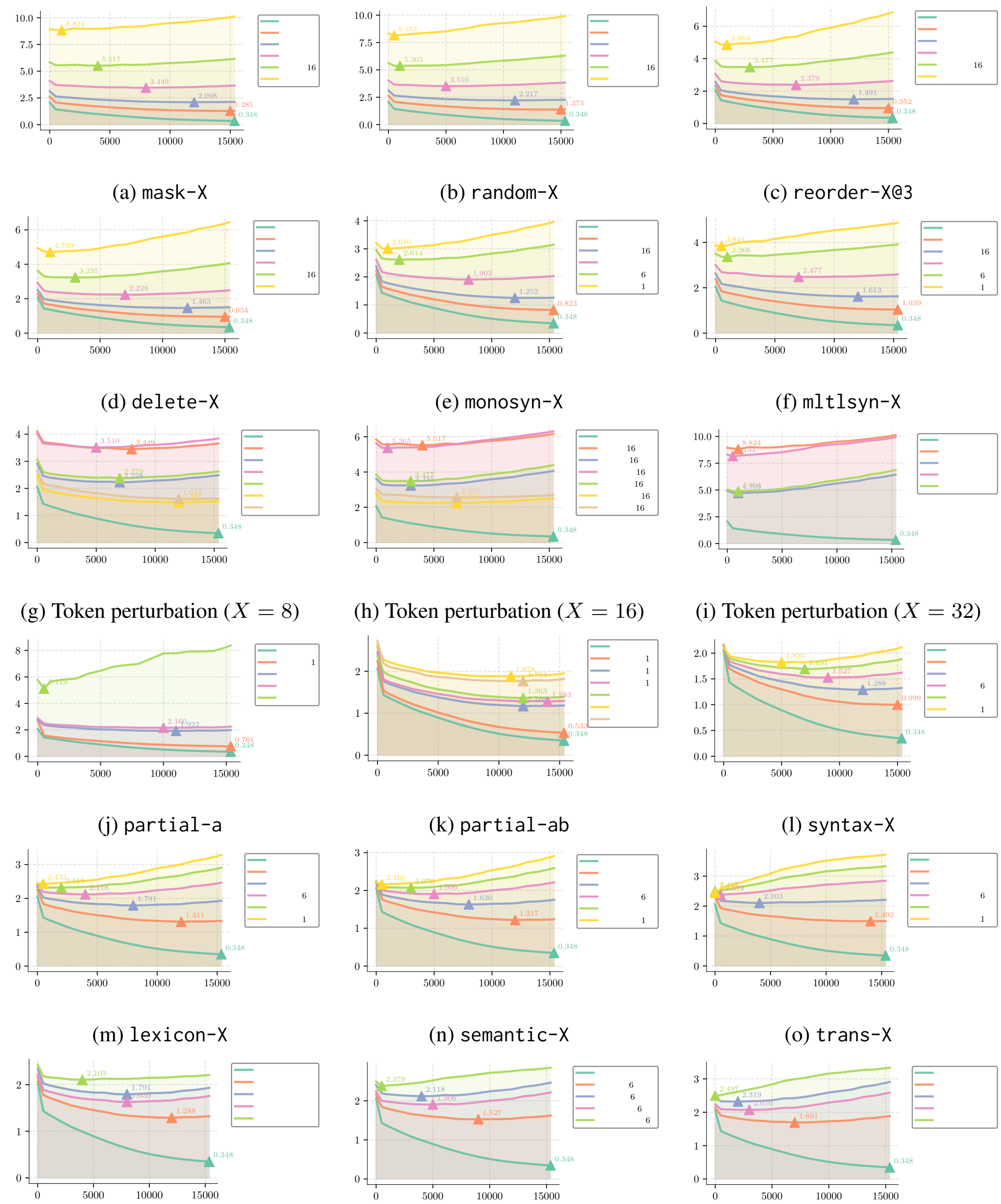
each document, however, we sample only one template per type.

Specifically, each document in J-STAGE contains multiple metadata fields, including:

- `title`: the title of the paper
- `abstract`: the paper abstract
- `keywords`: pre-defined keywords of the paper
- `fields`: research categories of the paper

Based on this information, we construct ten types of reading comprehension instructions as follows:

- **Summarization**: Summarize the context into one concise sentence, taking the abstract as input and the title as output.
- **Keyword Extraction**: Extract the keywords from the abstract, using the `keywords` field as the gold reference.
- **Field Identification**: Identify the research



(a) `mask-X` (b) `random-X` (c) `reorder-X@3`
(d) `delete-X` (e) `monosyn-X` (f) `mltlsyn-X`
(g) Token perturbation ($X = 8$) (h) Token perturbation ($X = 16$) (i) Token perturbation ($X = 32$)
(j) `partial-a` (k) `partial-ab` (l) `syntax-X`
(m) `lexicon-X` (n) `semantic-X` (o) `trans-X`
(p) Sentence perturbation ($X = 40$) (q) Sentence perturbation ($X = 60$) (r) Sentence perturbation ($X = 80$)

Figure 17: The loss dynamics over all perturbation patterns on English sequences.

field(s) of the paper, taking the abstract as input and the `fields` metadata as the expected output.

- **Translation**: Translate between English and Japanese, using bilingual metadata or parallel text segments as input-output pairs.
- **Text Completion**: Complete an incomplete abstract or title given the partial text, where the remainder of the text serves as the reference output.
- **Conclusion Derivation**: Derive the study's conclusion from its context, with the conclusion section as supervision.
- **Background Derivation**: Infer the background or motivation of the study from the provided abstract or introduction sentences.
- **Diagnosis**: Given a description of symptoms (extracted from biomedical corpora), predict

the corresponding diagnosis, using annotated datasets where available.

- **Reordering**: Reorder shuffled sentences into their natural sequence, ensuring coherence with the original abstract or section structure.

- **Goal–Method–Result–Conclusion (GMRC)**: Derive one missing component (*e.g.*, goal, method, result, or conclusion) based on the other three, enabling comprehension of scientific discourse structures.

(2) LLM-base QA generation: To enhance the data diversity, we further generate five question-answer pairs for each document. Specifically, we use DeepSeek-R1-Distill-Llama-70B for English QA-pair generation and use DeepSeek-R1-Distill-Qwen-JP-32B (Ishigami, 2025)[15] for Japanese QA generation. Noted that not all documents can successfully generate five QA pairs.

E.2 Continual training Settings

We perform continual training with **Megatron-LM** based on the llm-jp-3-13B model, and extend the same experimental protocol to two additional models: llm-jp-3-3.7B and llm-jp-3-7.2B. All models are trained under an identical distributed configuration to ensure fair evaluation of continual learning behavior across architectures.

Taking the llm-jp-3-13B model as an example, the training setup follows a distributed configuration with 4 compute nodes, each equipped with 8 A100 GPUs. We apply a tensor parallel size of 2 and a pipeline parallel size of 4, enabling efficient large-scale training with a sequence length of 4096. The optimizer is configured with a learning rate of 2×10^{-5}, weight decay of 0.1, and gradient clipping of 1.0, with a minimum learning rate of 2×10^{-6}. We adopt a micro-batch size of 1 and a global batch size of 32 to stabilize training. Under this configuration, training one epoch on 0.5B tokens requires approximately 7 hours, demonstrating the computational feasibility of continual training while maintaining efficiency on large-scale biomedical and cross-lingual corpora. We introduce the training hyperparameters of all three models as listed in Table 5.

[15] https://huggingface.co/cyberagent/DeepSeek-R1-Distill-Qwen-32B-Japanese

Hyperparameter	3.7B	7.2B	13B
Nodes	2	2	4
GPUs / Node	8	8	8
Tensor Parallel (TP)	2	2	2
Pipeline Parallel (PP)	2	2	4
Sequence Length	4096	4096	4096
Batch Size	16	16	32
Learning Rate	2e-5	2e-5	2e-5
Minimum LR	2e-6	2e-6	2e-6
Weight Decay	0.1	0.1	0.1
Gradient Clipping	1	1	1

Table 5: Training configuration of llm-jp-3.7B, llm-jp-7.2B and llm-jp-13B.

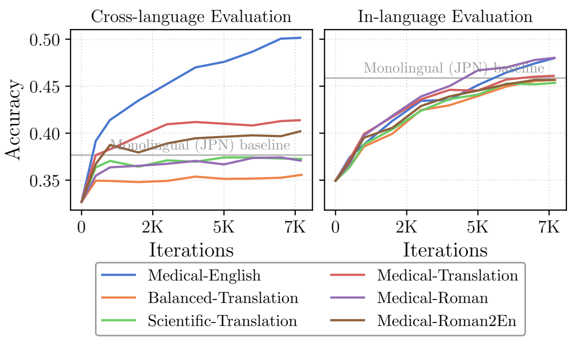


Figure 18: Japanese-to-English transfer evaluation with diverse strategies.

F Transfer From Japanese to English

Here, we report the results of the Japanese-to-English cross-lingual transfer evaluation using different recipes, as a supplement to the analysis in § 5.2.2. Specifically, in this setting, $\mathcal{C}_{\mathcal{K}}$ is composed of the Japanese monolingual training corpora, while we vary the data source of $\mathcal{C}_{\mathcal{T}}$ to investigate efficient cross-lingual transfer. Evaluation results are shown in Figure 18.

The evaluation results shown in Figure 18 exhibit trends that differ from those in Figure 12, as the strong baseline using multilingual corpora (a mixture of Japanese and English monolingual training corpora) largely outperforms other translation corpora. This suggests that generalization within English knowledge is an easier task than transferring knowledge from Japanese to English. Nevertheless, we still observe the strongest cross-lingual transfer in the closely related domain compared to the other three domains.

Score	Description
3	**Biomedical factual knowledge is clearly expressed**; the sentence states medical, clinical, physiological, genetic, or molecular biology facts. A precise biomedical factual triple can be extracted.
2	**Non-biomedical factual knowledge is expressed**; the sentence contains general domain facts or scientific/technical facts outside biomedical fields. A factual triple is extractable, but the content is not biomedical.
1	**Implied or incomplete factual knowledge**; the sentence suggests a fact but lacks sufficient context to extract a full, explicit triple. Content is vague as context might be contained in previous sentences.
0	**No factual knowledge**; the sentence expresses opinions, commands, emotions, or is too incomplete to yield any factual triple. No meaningful fact can be extracted.

Table 6: Annotation guide for *Detected Sentence (Factuality)* in Table 2.

G AdaXEval Human Evaluation Guideline

G.1 Fact Detection Quality Evaluation

Given the sentences that are detected as factual, please assign a score based on whether the sentence expresses factual knowledge and what type of factual knowledge it encodes.

G.2 Cloze prompt Quality Evaluation

Objective:

Evaluate whether fill-in-the-blank prompts generated from biomedical academic sentences and subject–relation–object triples are clear, faithful to the original sentence, and free from factual distortion. Each prompt removes the object from the triple and replaces it with a blank.

Input: You will be shown the following for each item:

- Original Sentence: The full academic sentence from which the triple is extracted.
- Triple: A `<subject, relation, object>` triple derived from that sentence.
- Generated Prompt: A fill-in-the-blank sentence where the object is replaced with `[BLANK]`.

Faithfulness Criteria

Score	Description
3	The prompt closely mirrors the original sentence's structure and meaning. Minor surface-level changes (*e.g.*, auxiliary verbs, punctuation, or sentence breaks) are acceptable.
2	The core meaning is preserved, but there are moderate changes in wording, omission that does not change the semantic meaning, or noticeable rephrasing.
1	The prompt differs significantly in structure or phrasing, or some key information is missed
0	The prompt is not clearly based on the original sentence or appears unrelated in meaning or form.

Table 7: Annotation guide for *Cloze prompt (Faithfulness)* in Table 2.

When the `[BLANK]` is replaced with the object, does the prompt preserve the structure and meaning of the original sentence, including key contextual information?

- **What it checks:** Structural and contextual similarity between the prompt and the original sentence
- **Focuses on:** Wording, phrasing, sentence structure, presence of supporting context.

G.3 Paraphrased Question Quality Evaluation

Objective: Evaluate whether **question-style prompts**, automatically generated from fill-in-the-blank biomedical prompts, are:

- Semantically faithful to the original prompt (i.e., same question being asked)
- Grammatically correct and fluent
- Natural as questions a human would realistically ask

Input:

- Original Fill-in-the-Blank Prompt (*e.g.*, "EGFR is highly expressed in [BLANK].")
- Paraphrased Question-style Prompt (*e.g.*, "In which condition is EGFR highly expressed?")

Example:

- Original Sentence: "EGFR is highly expressed in non-small cell lung carcinoma."

Score	Description
3	Fully natural and fluent; well-formed question
2	Mostly fluent; minor grammatical issues or slight awkwardness
1	Understandable but ungrammatical or clearly unnatural
0	Ungrammatical, confusing, or not a valid question

Table 8: Annotation guide for *Paraphrase (Fluency)* in Table 2.

Score	Description
3	The question fully matches the meaning of the original facts; no distortion or loss of critical details; context is complete and the answer remains uniquely correct.
2	The question is generally accurate but has minor factual ambiguity or slightly softer interpretation that could cause mild uncertainty while still pointing to the same answer.
1	Some key context or factual precision is missing or altered; relationships are weakened; the answer is still inferable but not strictly unique or reliable.
0	The question introduces clear factual errors, distorts the original meaning or relationships, or misleads so the answer could be wrong or invalid.

Table 9: Annotation guide for *Paraphrase (Factuality)* in Table 2.

- Triple: *(EGFR, is highly expressed in, non-small cell lung carcinoma)*
- Prompt: "EGFR is highly expressed in [BLANK]."

Evaluation Criteria

1) Fluency and Grammaticality: Is the question grammatically correct, fluent, and natural-sounding in English?

- Focuses on syntax, awkward phrasing, unnatural interrogative forms

2) Factual Consistency: When given the answer, is the question factually consistent with the original sentence and triple (no distortion of meaning or relationships)?

- **Semantic Accuracy:** Does the question preserve the intended meaning of the original sentence and triple?
- **Relationship Integrity:** Are the logical relations (*e.g.*, cause/effect, association, identity) between subject, relation, and object kept intact?
- **Context Preservation:** Does the question retain essential context needed to uniquely identify the correct answer (*e.g.*, disease type, anatomical site, conditions, time point, numerical thresholds)? Missing such context should lower the score.

Score	Description
3	Excellent reformulation; uses different vocabulary and structure while maintaining meaning
2	Good variation; some different wording but follows similar structure or only changing the structure without introducing new vocabulary
1	Minimal variation; mostly replaces the blank with a question word
0	No meaningful reformulation; essentially the same as the original with a question mark

Table 10: Annotation guide for *Paraphrase (Diversity)* in Table 2.

Score	Description
2	Direct match – same triple is clearly expressed in one sentence or consecutive sentences in the abstract in the target language.
1	Inferable without extra biomedical knowledge out of the given content – not in one sentence, but can be reasonably inferred from the paragraph as a whole.
0	Not supported or requires extra knowledge – the triple cannot be inferred from the abstract, or unrelated.

Table 11: Annotation guide for *Paraphrase (Factuality)* in Table 2.

3) Linguistic Diversity: How well does the paraphrased question use different wording and structure from the original prompt?

- **What it checks:** Lexical and syntactic variation between the original and paraphrased versions
- **Focuses on:** Synonym usage, sentence structure changes, reformulation techniques

4) Cross-Lang Factual Consistency: Can the evidence supporting the same fact be found or inferred from the document in the target language?

Score	Description
3	Highly plausible: Very convincing as an answer; can confuse even experts; fits subject, relation, domain well.
2	Moderately plausible: Makes sense in general; fits domain and context somewhat; can be ruled out by basic domain knowledge.
1	Barely plausible: Awkward or uncommon; easily ruled out by surface cues or common sense without any domain knowledge.
0	Implausible: Irrelevant, nonsensical, or grammatically incorrect; not a valid answer option.

Table 12: Annotation guide for *Distractor (Plausibility)* in Table 2.

Score	Description
3	Definitely wrong: contradicts or is not supported by the original sentence.
2	Likely wrong: but could be ambiguous or partially true given the original sentence.
1	Borderline: Possibly true or partially correct; ambiguous given the sentence.
0	Incorrectly labeled – This distractor is actually correct or the original answer given the original sentence.

Table 13: Annotation guide for *Distractor (Incorrectness)* in Table 2.

G.4 Distractor Quality Evaluation

Objective: Evaluate the quality of three distractor options (incorrect candidates) accompanying the correct answer (object) in a multiple-choice setting derived from biomedical fill-in-the-blank prompts. Each distractor should be:

- Plausible given the question
- Incorrect (not the original object)
- Relevant in context and domain

Input:

- Original Sentence
- Triple
- Fill-in-the-Blank Prompt
- Answer options

Example:

- Original Sentence: `EGFR is highly expressed in non-small cell lung carcinoma.`
- Triple: *(EGFR, is highly expressed in, non-small cell lung carcinoma)*
- Prompt: `EGFR is highly expressed in [BLANK].`

Evaluation Criteria: Need evaluations for both the cloze prompt and the paraphrased question.

1) Plausibility in Context: Is the distractor believable given the prompt and domain knowledge (biomedical)?

- **What it checks:** subject, relation, and expected answer type (should be the object)
- **Focuses on:** Be cautious of meaning shifts, incorrect substitutions, or role reversals.

2) Incorrectness: Is the distractor clearly incorrect given the original sentence and triple?