Revealing the Dynamics of Multilingual Knowledge Acquisition in Domain Adaptation

Xin Zhao[1,2,a)] Naoki Yoshinaga[3,b)] Yuma Tsuta[2,4] Akiko Aizawa[2,c)]

Abstract: Multilingual domain adaptation is a common approach for learning new domain knowledge and transferring it across languages. Although various methods have been proposed to enhance domain adaptation, the process of multilingual knowledge acquisition, specifically how monolingual domain knowledge is learned and transferred across languages, remains underexplored. This often results in suboptimal performance, which can be particularly detrimental in low-resource domains and languages. This work examines the learning dynamics of large language models (LLMs) during adaptation. To directly investigate how multilingual knowledge acquisition is achieved, we introduce AdaXEval, an adaptive evaluation method that constructs multiple-choice QA datasets using the bilingual domain corpus. Through continual pre-training with diverse data recipes, we track how LLMs acquire domain facts and pinpoint the mechanism behind the transformation process from domain training data to knowledge. Experiments on a 13B bilingual LLM reveal that cross-lingual transfer remains challenging despite a high-quality bilingual corpus, underscoring the need for more effective domain knowledge transfer methods.

Keywords: Domain Adaptation, Cross-lingual Transfer, Knowledge Acquisition, Biomedical NLP

1. Introduction

Domain adaptation is a practical approach for adapting large language models (LLMs) trained on general-domain corpora to new knowledge [1], [2] or specialized domains such as biomedicine and finance [3], [4], [5]. This process typically involves continual pretraining on domain-specific corpora to infuse LLMs with relevant knowledge [6], [7], [8]. However, acquiring such corpora is often challenging, particularly for low-resource languages where domain-specific data is scarce [9].

To achieve effective domain adaptation under low-resource settings, a common strategy is data augmentation, which expands limited domain corpora by generating synthetic or derived data. Techniques such as regex-based pattern mining [10], paraphrasing [11], and leveraging generative models [12] can increase data diversity, thereby enhancing model exposure to domain-specific terminology and usage patterns. Another line of research investigates cross-lingual transfer, which leverages domain knowledge from high-resource languages to benefit low-resource ones. By augmenting training corpora with multilingual or bilingual corpora [13], [14], [15] or creating code-switching data [16], [17], it is possible to transfer knowledge across languages. While these data augmentation methods are shown to facilitate domain knowledge acquisition in LLMs, prior work predominantly focuses on empirical solutions rather than understanding the underlying mechanisms, leaving the critical factors influencing domain adaptation unclear.

Our work aims to investigate the process of factual knowledge acquisition in domain adaptation from a mechanistic perspective. Specifically, we seek to understand, during the continual-training process on the domain corpus, how domain facts are **memorized** and **generalized** across different linguistic contexts, including both **intralingual** (within a language) and **interlingual** (across languages) variations, and to identify the key factors that facilitate effective knowledge acquisition and transfer. To achieve the goal, we focus on three research questions:

RQ1: How to effectively evaluate the factual knowledge memorization and generalization during domain adaptation?

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

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

Existing approaches to evaluating domain knowledge largely rely on public benchmarks [18], [19], [20] or training loss analysis [21], [22], [23]. However, these methods offer limited coverage for low-resource domains and fail to capture generalization across paraphrases and translations adequately. Moreover, prior analyses have been largely restricted to relational facts involving named entities and predefined relations [23], [24], which fail to capture the complexity of domain-specific knowledge. To address these gaps, we propose **AdaXEval**, an adaptive multilingual evaluation pipeline for domain adaptation. AdaXEval uses the bilingual corpus to automatically generate multiple-choice datasets for knowledge **memorization**, intralingual generalization (**paraphrase**), and interlingual generalization (**cross-lingual transfer**), ensuring applicability to new domains and lan-

[1] The University of Tokyo
[2] National Institute of Informatics
[3] Institute of Industrial Science, The University of Tokyo
[4] Present address: Fixstars Corporation
[a)] xzhao@tkl.iis.u-tokyo.ac.jp
[b)] ynaga@iis.u-tokyo.ac.jp
[c)] aizawa@nii.ac.jp

guages. Human annotation from multiple perspectives confirms that AdaXEval provides reliable and informative evaluation results, demonstrating its effectiveness in evaluation.

We next investigate how training data is dynamically transformed into knowledge that LLMs can store to deepen our understanding of models' knowledge acquisition ability. Specifically, we study biomedicine domain adaptation for Japanese using a 13B English/Japanese bilingual model [25], with English serving as a comparison and source for knowledge transfer. We begin with monolingual continual pretraining on both English and Japanese using the J-STAGE corpus containing biomedical abstracts for both languages. By evaluating training checkpoints with the AdaXEval-generated dataset (hereafter, AdaXEval), we observe a gradual knowledge acquisition process for both cloze queries and paraphrases; however, the model struggles to achieve cross-lingual transfer. Further analysis reveals that knowledge is acquired as losses for correct options are **shielded** from rapid growth due to overfitting, a phenomenon we term **loss shielding**. This acquisition eventually plateaus as continued training causes the model to overfit, leading to increasing loss across all options. Examining losses on noisy training data confirms that the LLM overfits readily to fixed token sequences, even when only partial tokens are perturbed. This overfitting explains the limited cross-lingual transfer, as there is minimal token overlap between training and evaluation data across languages.

Finally, we investigate factors critical for cross-lingual transfer of domain knowledge. Since monolingual training yields only limited cross-lingual transfer, we employ multilingual continual pretraining using various data recipes. We specifically examine translation and romanization strategies to facilitate transfer. Our experiments confirm that the presence of cross-lingual tokens in closely related domains is essential for successful knowledge transfer. Nonetheless, even with high-quality alignment signals such as translations, cross-lingual transfer remains challenging, underscoring the need for more effective methods.

The contributions of this paper are:

(1) We propose AdaXEval, an adaptive multilingual evaluation pipeline that can automatically generate datasets to evaluate domain knowledge across memorization, intralingual and interlingual generalization.

(2) We analyze the training dynamics of domain adaptation, showing that knowledge acquisition is driven by differential loss allocation but eventually plateaus due to overfitting.

(3) We reveal the mechanism of the transformation from training corpora to knowledge and investigate the crucial factors to achieve cross-lingual transfer during domain adaptation.

2. Related Work

2.1 Domain Knowledge Acquisition Evaluation

The evaluation of domain knowledge is typically performed using public benchmarks [18], [19], [20]. However, such benchmarks are not always for low-resource languages or specialized domains. Also, evaluating knowledge acquisition usually requires paraphrase datasets for intralingual generalization and translation datasets for interlingual generalization, which can not all be captured by standard benchmarks. Moreover, the knowledge contained in the training data and that covered by the benchmarks may differ substantially, making the evaluation results an imperfect reflection of the actual knowledge acquired during training. On the other side, some studies observe acquisition progress with loss reduction as evidence [21], [22], [23]. However, the loss does not necessarily reflect the domain-specific knowledge that end users care about. Finally, most existing knowledge analyses rely on predefined relational facts [23], [24], which are insufficient for representing the richer and more complex knowledge structures characteristic of specialized domains.

2.2 Mechanisms of Knowledge Acquisition in LLMs

Large language models (LLMs) have been widely studied for their capacity to store and retrieve factual knowledge [22], [26], [27], [28]. This has spurred growing interest in understanding the underlying mechanisms by which such knowledge is encoded and accessed [29], [30], [31], [32]. Recent work has also investigated the dynamics of knowledge acquisition by analyzing intermediate model checkpoints during training. [22] shows that LLMs accumulate factual knowledge through repeated exposures, gradually increasing the likelihood of correct recall with each encounter. Key factors influencing this acquisition include fact frequency, model scale, and batch size [22], [23], [24].

Parallel to monolingual knowledge studies, the cross-lingual transfer capabilities of LLMs have garnered considerable attention, particularly for enabling efficient knowledge transfer from high-resource to low-resource languages. Unlike general linguistic understanding, the transfer of factual knowledge across languages remains notably challenging [23], [24]. Studies suggest that facts expressed in different languages may be stored as distinct representations within the model [24], [33], [34], and only certain relational types are effectively transferable across languages [23], [24]. Despite these advances, existing studies primarily focus on pretraining dynamics and structurally relational facts. The connection between specific training data and the knowledge it represents remains insufficiently explored, highlighting the need for a deeper investigation into how LLMs acquire and organize complex, domain-specific knowledge.

2.3 Cross-lingual Knowledge Transfer

Cross-lingual transfer in domain adaptation has been studied through various strategies that aim to bridge language gaps in knowledge transfer. Translation data are widely used in training LLMs to enhance cross-lingual knowledge transfer [35], [36], [37], supporting applications such as machine translation [36] and instruction following [38], [39]. However, since in-domain translation data are often scarce, it remains unclear whether translation can effectively transfer complex and sparse domain knowledge across languages. Furthermore, the romanization strategy reduces script barriers by converting text into romanized forms, thereby aligning linguistic representations with English [40], [41], [42]. Finally, code-switching has been shown to be effective in aligning token semantics across languages by interleaving tokens from multiple languages within the same context [16], [17]. However, it is still uncertain whether these alignment signals are sufficient for transferring complex domain knowledge across languages.

3. Domain Knowledge Acquisition Evaluation

To evaluate knowledge acquisition in domain adaptation under realistic scenarios, we propose AdaXEval, an adaptive multilingual evaluation pipeline. This pipeline generates multiple-choice datasets for assessing knowledge memorization, generalization, and cross-lingual transfer using advanced LLMs. We validate AdaXEval's effectiveness through human annotation.

3.1 AdaXEval

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

3.1.1 Fact Extraction

Extracting facts from realistic domain corpora is challenging, as complex domain knowledge cannot be easily formalized into named entities or predefined relations. AdaXEval addresses this challenge through a two-step strategy: named-entity-recognization (NER)-based sentence filtering and multi-agent triple extraction. First, domain-specific NER tools and shallow linguistic heuristics are employed to identify sentences in the training corpora that contain multiple named entities. Next, we design Chain-of-Thought (CoT) instructions to guide the extraction of factual triples from the filtered sentences, with each triple represented in the format <subject, relation, object>. The subject and object are preferably named entities, though longer descriptive phrases are also acceptable to fit in realistic fact representation. The implementation details of these instructions can be found in the open-source code. Specifically, AdaXEval employs a multi-LLM agent for triple extraction, estimating the overall confidence of the outputs, and adapting the top-confident extraction result to improve evaluation reliability.

3.1.2 Question Crafting

AdaXEval utilizes advanced LLMs to generate multiple-choice datasets to measure three key knowledge acquisition abilities.

1) Knowledge memorization aFor each recognized fact (e.g., ⟨*blood sugar level, can be controlled by, insulin*⟩), we generate a fill-in-the-blank query with [BLANK] as the placeholder for the object (e.g., *Blood sugar level can be controlled by [BLANK].*) The query should closely match the original sentence from the training corpus to assess memorization ability exclusively, using the original sentence as a reference.

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. While we consider translation as a strong candidate, translation of domain knowledge is challenging due to specialized named entities, technical terminology, and domain-specific concepts that may not have direct equivalents across languages. To address this challenge, we adapt 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 language Y. [*1]

Table 1: Dataset statistics at each step of the AdaXEval pipeline.

Step	English	Japanese
Sampled abstracts	10,000	10,000
Splitted sentences	81,770	71,661
Sentences after entity filtering (≥2 entities)	45,390	40,762
Triple extraction	4840	3926
Cloze queries generation	4840	3926
Paraphrases generation	4840	3926

3.1.3 Distractor Generation

Finally, AdaXEval generates three plausible yet incorrect answer options that remain topically related but unambiguously wrong, while explicitly instructing the model to avoid surface-level cues such as sequence length.

3.2 Experimental Setup

Domain corpus: Our study investigates biomedical domain adaptation in English–Japanese language pairs. Specifically, we utilize the J-STAGE biomedical corpus, which consists of academic paper abstracts, as the training data for both model training and evaluation (see § 4.1 for details).

Details of Generation: We randomly sampled 10,000 parallel documents for the evaluation dataset generation. The NER tools we used for sentence filtering are () for English and () for Japanese. We split abstracts into sentences and filter out sentences with fewer than two entities, retaining the remainder for triple extraction. For triple extraction in each sentence, we use three open-source LLMs [*2] from different families to assess the confidence of whether the sentence is factual. We sum the confidence scores across the three LLMs and retain sentences with combined confidence scores greater than 2 (maximum 3), using their generated triples as the final sentence set. Finally, we use GPT-4.1 to generate cloze queries, paraphrases, and generate three distractors for each question instance. The number of instances after each step is shown in Table 1. See Appendix A.1 for details of the LLMs used, instructions, and confidence calculation methods.

3.3 Evaluation metric

For each dataset, we follow [43] 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

[*1] Note that filtered and extracted triples may differ across languages. Thus, facts evaluated for interlingual generalization differ from those used in the other two. While this prevents direct comparison between three abilities, our primary focus on accuracy changes during training makes this a reasonable trade-off for evaluation quality.

[*2] We employ open-source LLMs for local inference, as the large number of candidate sentences would otherwise incur substantial computational costs.

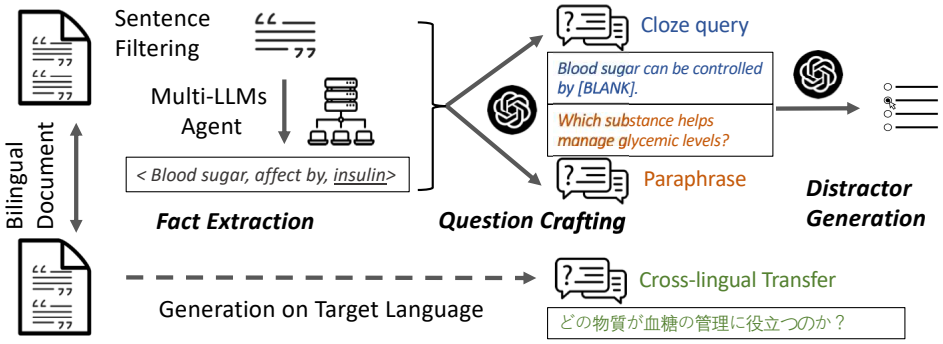


Fig. 1: Overview of AdaXEval pipeline.

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. Finally, we take the accuracy of prediction as the knowledge acquisition metric.

Formulation: Let the model be $p_\theta(\cdot \mid \cdot)$. Each dataset $\mathcal{D}$ contains pairs (q, a), where q is either a *cloze query* or a *paraphrase query*, and $a = (a_1, \ldots, a_m)$ is the tokenized answer sequence (e.g., "insulin"). For cloze queries, the query string contains a special token [BLANK], and for paraphrases, the query 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 query 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 query 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:

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

3.4 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. **1) Triple extraction quality evaluation** examines both in-language correctness and cross-lingual correctness, which indicates whether the same knowledge can be found in another language corpus. **2) Cloze query evaluation** checks the faithfulness of the generated prompt to the original sentence structure and factual consistency. **3) Paraphrases evaluation** is conducted on three dimensions: fluency and grammaticality, semantic equivalence to original prompts, and linguistic diversity in reformulation. **4) Distractor quality** is measured through plausibility within the domain and apparent incorrectness relative to the original context. See the full human evaluation guideline documented in Appendix A.6.

Each metric employs structured scoring rubrics with scales ranging from 0-2 for triple evaluation to 0-3 for other components, enabling systematic assessment of dataset quality across multiple linguistic and semantic dimensions. We randomly sample 100 instances for each language and conduct human evaluation following the guidance above. As shown in Table 2, in-language triple correctness is high, whereas cross-lingual correctness is slightly lower. This difference is attributable to the fact that bilingual abstracts in J-STAGE represent document-level alignments rather than strict translations. The lower faithfulness score reflects a tendency of the generations to introduce moderate rephrasings or lexical substitutions while retaining the core meaning. The high quality of paraphrase and distractor generation can be attributed to the use of a closed-source LLM (GPT-4.1). Overall, our evaluation results indicate that AdaXEval is able to generate high-quality evaluation data, meeting the requirements for assessing knowledge memorization as well as intralingual and interlingual generation.

4. Tracing Knowledge Acqusition

In this section, we examine the training dynamics of domain adaptation and explore the mechanism underlying the transformation from training data to knowledge. We perform monolingual continual pretraining on English and Japanese using a biomedical domain corpus as a case study.

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

Evaluation Metric	Japanese	English
Triple (In-lang)	1.78/2	1.76/2
Triple (cross-lang)	1.54/2	1.62/2
Cloze query (Faithfulness)	2.18/3	2.52/3
Cloze query (Factuality)	2.72/3	2.58/3
Paraphrase (Fluency)	2.97/3	2.96/3
Paraphrase (Equivalence)	2.67/3	2.84/3
Paraphrase (Diversity)	2.83/3	3.0/3
Distractor (Plausibility)	2.24/3	2.32/3
Distractor (Incorrectness)	2.93/3	2.89/3

4.1 Experimental Setup

Data preparation: For bilingual data, we employ a subset of the J-STAGE corpus, a collection of Japanese research papers with some abstracts translated into English. From this corpus, we extract biomedical bilingual papers, comprising 614,444 Japanese documents and 404,643 English documents, where each English document has a corresponding Japanese counterpart. These bilingual pairs serve as candidates for AdaXEval. To enhance domain adaptation performance and enable more fine-grained analysis, we adopt instruction pretraining as a strong data augmentation baseline for constructing the training corpus. Specifically, we generate biomedical instructions from raw text using both rule-based mining patterns [44] and LLM-based question–answer generation [45]. The raw documents are then combined with the generated instructions for continual pretraining. Details of the training dataset construction are provided in Appendix A.4.1.

Training setup: For each training run, we sample 0.5B tokens from the constructed corpus. We adopt llm-jp-3-13B, a strong Japanese–English bilingual LLM, as the base model for pretraining, owing to its superior language understanding ability in both languages, particularly Japanese. Each model is trained for four epochs using 32 A100 GPUs. Further implementation details, dataset specifications, and additional pretraining hyperparameters are provided in Appendix A.4.

4.2 Dynamic Evaluation on Knowledge Acquisition

Figure 2 reports AdaXEval results for English and Japanese monolingual training. The results show that domain knowledge is gradually memorized during training in both languages. In addition, both languages exhibit strong intralingual generalization on the paraphrased datasets close to memorization performance. Interlingual evaluation further reveals that monolingual training can induce cross-lingual knowledge transfer, although the extent of transferable knowledge remains limited.

Despite these benefits, several issues emerge as follows. Addressing these questions could provide deeper insights into the mechanisms underlying domain knowledge acquisition.

(1) Why does memorization accuracy converge at around 50% rather than approaching 100%?

(2) What accounts for the superiority of paraphrasing over memorization in English training?

(3) Why does cross-lingual transfer happen in monolingual training and only happen in initial training?

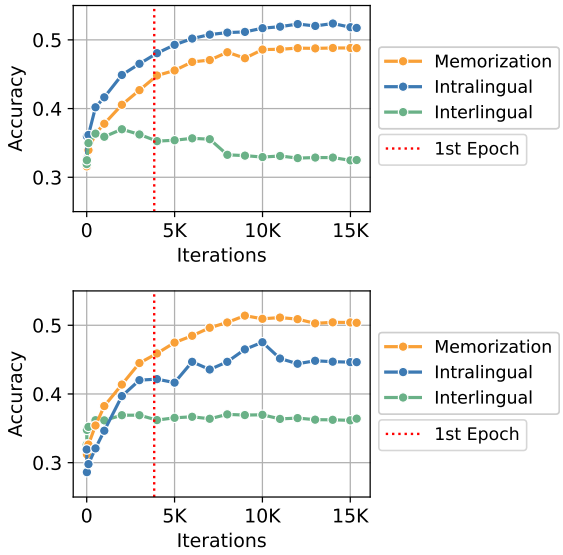


Fig. 2: AdaXEval result during 0.5B continual pretraining in **English (above)** and **Japanese (below)**.

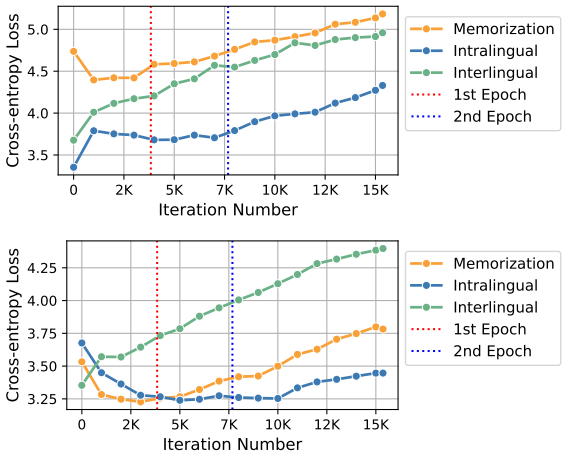


Fig. 3: Loss dynamics of question-answer sequences in AdaXEval, during **English (above)** and **Japanese (below)** training.

4.3 Tracing Knowledge through Loss Dynamics

This section analyzes the sequence loss on both training and evaluation data to address the questions raised above and to elucidate the mechanisms of knowledge acquisition during training. We focus on loss because it directly drives predictions on our multiple-choice datasets (see § 3.3) and reflects the model's token generation behavior, where sequences with lower loss are more likely to be generated by the model.

(1) Training overfits to data, but the *loss shielding* drives knowledge memorization. We calculate the loss of sequences corresponding to correct answers in AdaXEval. Figure 3 illustrates the loss changes across training checkpoints for both English and Japanese training. On the cloze query dataset, the loss initially decreases during early training but begins to increase after the first epoch, indicating that the model is overfitting to the

training corpus. In contrast, the accuracy dynamics in Figure 2 indicate that memorization continues until about 10K iterations. To understand this divergence, we measure the ratio of the sequence loss for correct answers to the total loss across the four candidate options. Figure 4 shows that the loss ratio aligns with the accuracy dynamics, suggesting that knowledge can still be memorized even as the model is overfitted to training sequences, since correct sequences are shielded from rapid loss growth.

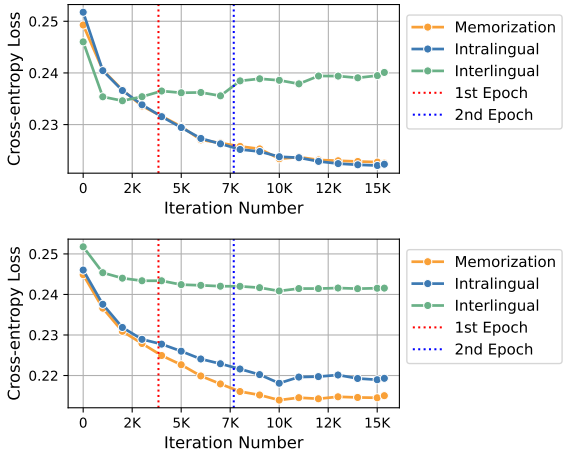


Fig. 4: Sequence loss ratio for correct answers versus all candidates during **English (above)** and **Japanese (below)** training.

(2) Training memorizes knowledge first and then generalizes. By comparing the loss changes between the cloze query and paraphrase datasets in Figure 3, we observe that while the paraphrase loss decreases more slowly, or even increases during the initial stage, its reduction persists longer than that of the cloze query. This suggests that the training process first memorizes surface patterns and then generalizes them to different linguistic styles. Moreover, Figure 4 shows that sequence loss increases more rapidly on paraphrase than on cloze query datasets in Japanese, leading to better memorization performance as evidenced in Figure 2. Finally, the lower memorization accuracy compared to intralingual generalization in English can be explained by loss volume: although both exhibit synchronous loss trends, intralingual generalization consistently maintains a lower overall loss. This is likely attributable to the built-in biomedical knowledge in English in llm-jp-3-13B and the additional contextual information provided in the paraphrase queries (see AdaXEval examples in Appendix A.2).

(3) Sequences in interlingual datasets also exhibit loss shielding. Although the absolute loss of the interlingual datasets continues to increase (Figure 3), the loss ratio dynamics in Figure 4 show that interlingual generalization still benefits from loss shielding during the first 10K iterations. However, this effect diminishes rapidly, likely due to the substantial linguistic divergence between the training data and the interlingual sequences.

5. From Training Data to Knowledge

In this section, we investigate how training data are transformed into knowledge. § 4 illustrates the role of loss shielding in knowledge acquisition. However, this effect is short-lived, as it is eventually overridden by severe overfitting. Examining how the loss behaves on sequences that did not appear in training can provide deeper insights into the mechanisms by which knowledge is derived from training data.

5.1 Linguistic Divergence Harms Loss Shielding

Although prior analysis highlights the role of loss shielding in enhancing knowledge acquisition, the relationship between the degree of loss shielding and the linguistic divergence between test and training sequences remains unclear. To address this, we systematically introduce controlled perturbations into the training data by injecting noise according to different rules and observe their loss dynamics over the training process. Specifically, we randomly sample 2,000 sequences from Japanese monolingual training data and apply perturbations at both the token and sequence levels.

5.1.1 Token-level Perturbation

Token sequences are perturbed after tokenization using the methods described below. To quantify the magnitude of perturbation, we compute the token edit distance between the original and perturbed sequences.

- **mask-X**: Replace X% tokens with <unk> token.
- **random-X**: Replace X% tokens with randomly sampled tokens from the tokenizer vocabulary.
- **delete-X**: Delete X% tokens.
- **reorder-X@Y**: Reorder tokens to achieve an edit distance equal to X% of the sequence length, with swaps restricted to a window of size Y.
- **monosyn-X**: Replace X% tokens with Japanese synonyms.
- **mltlsyn-X**: Replace X% tokens with English synonyms.

Specifically, we set $X = 2^{1,\ldots,5}$, $Y = 2^{0,\ldots,4}$. Random replacements are sampled from the Japanese-specific llm-jp tokenizer. For synonym substitutions, we use the Japanese WordNet 2.0 [46] to retrieve both Japanese and English synonyms. Additional implementation details are provided in Appendix A.3.

(1) More token editing causes a larger harm in loss. Taking *mask-X* as an example, Figure 5 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 introduce a substantial initial loss increase and accelerate the onset of overfitting. The analysis of other perturbation patterns leads to the same conclusion. See Appendix A.3 for all token perturbation results.

(2) Loss sensitivity varies with vocabulary and structural perturbations. Figure 6 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*, *delete*) that avoid introducing new vocabulary. Perturbations introducing irrelevant tokens (*mask*, *random*) inflict the greatest harm, significantly increasing the initial loss and hastening the overfitting onset. These findings underscore the sensitiv-

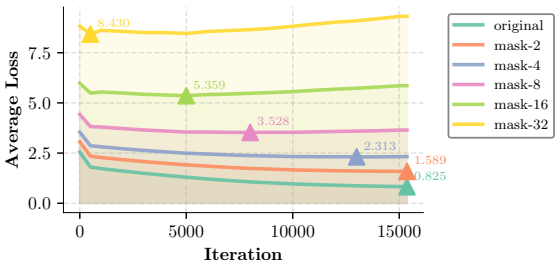


Fig. 5: The loss dynamics of mask-X sequences.

ity of loss dynamics to vocabulary perturbations, highlighting the importance of enhancing data diversity during training.

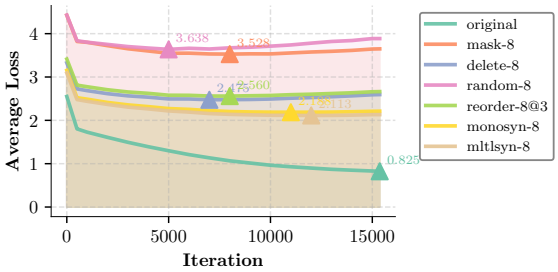


Fig. 6: The loss dynamics of token perturbation with 8% edit distances.

(3) Loss is robust to token order. Figure 7 presents the loss dynamics for the *reorder-4@Y* perturbation, where the window size Y governing the reordering is varied. The analysis reveals that increasing Y has a minimal impact on loss, suggesting that the extent of positional shuffling within larger windows does not significantly alter the model's loss landscape. This indicates that the primary influence on loss originates from the vocabulary change rather than the token's spatial information, underscoring the model's robustness to token order.

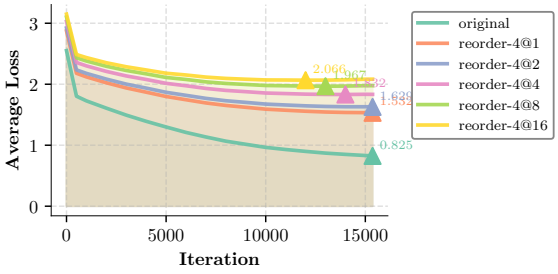


Fig. 7: The loss dynamics of reorder perturbation with varied window sizes.

5.1.2 Sentence-level Perturbation

We further perturb target sequences at the sentence level, rather than at the token level, to simulate more realistic noise patterns, considering the following perturbation strategies.

- **partial-a**: Split each training sequence into four segments, then select the *a*-th segment.
- **syntax-X**: Rewrite *X*% sentences, modifying only syntax without changing vocabulary.
- **lexicon-X**: Rewrite *X*% sentences, modifying only vocabulary without changing syntax.
- **semantic-X**: Rewrite *X*% sentences, allowing both syntactic and lexical changes.
- **translation-X**: Translate *X*% sentences into target language.

(1) Later sentences in a document exhibit stronger dependence on prior context. Using partial sentences following the *partial-a* strategy, Figure 8 shows that sentences appearing later in a training document incur higher loss when evaluated in isolation. This indicates that these sentences rely more heavily on their prior context. It highlights that LLMs learn rich, long-range dependencies. Meanwhile, it limits knowledge acquisition ability due to heavy dependence on context, underscoring the importance of a more balanced context-free learning paradigm.

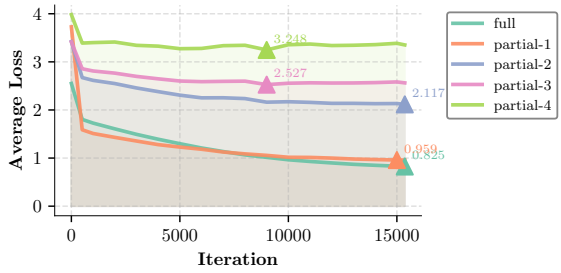


Fig. 8: The loss dynamics of partial sentences.

(2) Preserving vocabulary improves loss robustness. Figure 9 presents the loss dynamics for all rewriting patterns applied to 40% of sentences, including both paraphrasing and translation. Among these patterns, syntax rewriting shows the most substantial loss shielding effect, followed by semantic rewriting. This is because semantic paraphrasing does not require extensive vocabulary replacement, resulting in fewer word substitutions compared to the lexicon paraphrases. These results highlight the necessity of incorporating lexicon-focused paraphrasing during training to improve models' ability to generalize knowledge across diverse test inputs, consistent with the observations in § 5.1.1.

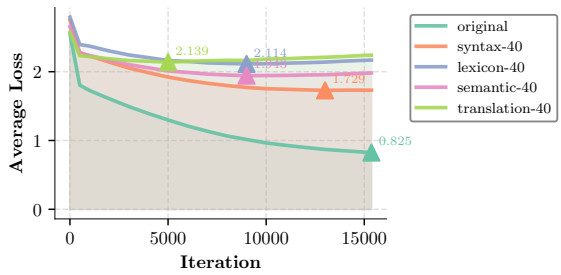


Fig. 9: The loss dynamics of sentence rewriting.

5.2 Cross-lingual Tokens Help Transfer

We then explore how cross-lingual transfer is achieved. Figure 2 shows that, compared to significant in-language knowledge acquisition when the training and evaluation languages align, monolingual training can achieve cross-lingual transfer but only

with limited performance improvement. Specifically, we observe a stable improvement in Japanese-to-English transfer, whereas English training yields little to no improvement in Japanese performance, as shown in Figure 2.

As the AdaXEval interlingual loss continues to increase, making it inappropriate for analysis, we instead observe the loss dynamics of the training sequence in different languages. This can serve as indirect evidence, as the evaluation instance is basically the noisy version of the training sequences. Specifically, we sample 1,000 training examples for each language and measure per-token loss dynamics under two monolingual training settings. Since cross-lingual transfer converges rapidly within the first 1,000 iterations, we analyze the loss dynamics at a finer granularity during the first epoch. The results are shown in Figure 10. We observe that in-language loss consistently decreases, whereas cross-language loss decreases only during the initial iterations, reflecting the rapid but short-lived interlingual generalization, with no subsequent improvement and even degradation, as also illustrated in Figure 2.

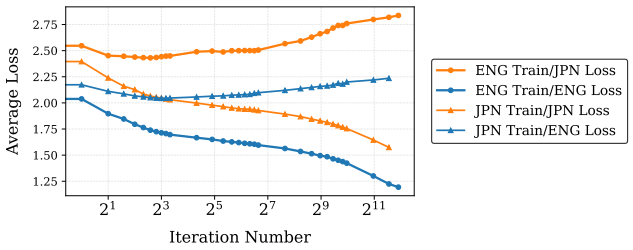


Fig. 10: English/Japanese sequence loss on English/Japanese training.

To investigate which tokens contribute to the initial loss reduction, we further measure the per-token cross-entropy loss within the given sequences. Specifically, we examine the impact of two linguistic characteristics on loss changes: part-of-speech (POS) and language. For POS, tokens are grouped by their syntactic category, and for language, tokens are classified by the language in which they occur, either Japanese or English. Figure 11 shows token-level loss dynamics on Japanese sequences during English monolingual training. The results indicate that English tokens embedded in Japanese sequences benefit from English training. Among different POS tags, only numerical (NUM) and punctuation (PUNCT) tokens, both present in the English corpus, exhibit notable loss reduction. This finding suggests that tokens occurring in the training corpus experience greater learning gains. It also explains why Japanese-to-English transfer is easier than English-to-Japanese: Japanese academic texts often include English terminology, whereas English texts rarely contain Japanese vocabulary.

6. Bridging Languages in Domain Adaptation

Although prior analysis emphasizes the critical role of cross-lingual tokens in enabling interlingual generalization, the factors driving improvements in cross-lingual transfer remain poorly understood. This uncertainty stems from the limited cross-lingual transfer observed in monolingual training settings. To bridge this

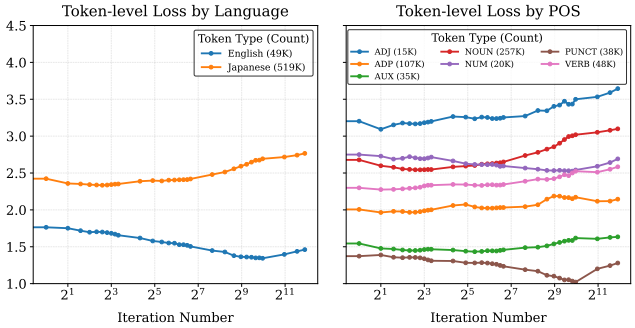


Fig. 11: Per-token loss dynamics by language and POS.

gap, this section explores the key factors that enhance effective cross-lingual transfer. Specifically, we investigate: what types of cross-lingual corpora can optimize knowledge transfer during domain adaptation?

6.1 Experimental Setup

(1) Multilingual continual pretraining: To examine the factors that facilitate cross-lingual transfer, we construct a series of cross-lingual corpora and evaluate their effectiveness. To ensure a fair comparison, each training corpus is composed of two parts: a **knowledge injection corpus** $C_{\mathcal{K}}$ and a **cross-lingual transfer enhancement corpus** $C_{\mathcal{T}}$. The $C_{\mathcal{K}}$ contains the target knowledge expressed exclusively in the source language X, while $C_{\mathcal{T}}$, by contrast, does not provide any new knowledge; instead, it serves only to establish linguistic connections between language X and the target language Y. We carefully ensure that the knowledge to be evaluated in language Y is absent from $C_{\mathcal{T}}$. Finally, we evaluate cross-lingual knowledge transfer by measuring the model's ability to express the acquired knowledge in language Y, using the same evaluation metrics introduced in § 3.3. We fix the source language X as English and the target language Y as Japanese to facilitate analysis in this section. See Appendix A.5 for the evaluation on reverse transfer direction.

(2) Cross-lingual transfer enhancement corpora: We construct diverse corpora to enhance cross-lingual transfer using two primary strategies: *translation* and *romanization*. As baselines, we consider an empty cross-lingual corpus ($C_{\mathcal{T}} = \emptyset$) (**Monolingual**) and a strong domain-specific baseline using J-STAGE Japanese data, with documents related to the AdaXEval evaluation filtered out (**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 that provide the model with cross-lingual transfer capabilities:

- **JParaCrawl (Balanced domain)**: An English–Japanese web-crawled corpus covering diverse domains [47].
- **ASPEC (Science domain)**: A multilingual corpus in English, Japanese, and Chinese, containing academic paper abstracts across various scientific fields [48]. Documents in the medical and chemical domains are excluded to distinguish this corpus from the target medical domain.
- **J-STAGE (Medical domain)**: The J-STAGE corpus represents the closest domain to our target. We filter out any doc-

uments included in the AdaXEval evaluation dataset to avoid contamination.

To evaluate the romanization strategy, we construct a medical romanization dataset (**Medical-Roman**). We first convert J-STAE Japanese text to romaji using `cutlet`[*3], 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.

(3) Details of training: Focusing on knowledge transfer from English to Japanese, we prepare two corpora, C_K and C_T, each containing 0.5 billion tokens, except for the **Monolingual** baseline. The English monolingual pretraining data serve as C_K. For C_T, we select seven candidate datasets as described in Sec. 6.1 (2). We then combine the two 0.5B-token corpora into a single 1B-token corpus, shuffle it, and train llm-jp-3-13B on it for one epoch. See Appendix A.4 for more training information.

6.2 Cross-lingual Transfer Evaluation

(1) Only domain-specific translation data enhances cross-lingual transfer. Figure 12 presents the accuracy dynamics for multilingual domain adaptation, comparing two baselines with three translation-based cross-lingual transfer datasets. After one epoch of training, only the **Medical-Translation** corpus leads to a clear improvement in transferring domain-specific knowledge to Japanese. Interestingly, the **Monolingual** baseline outperforms the **Medical-Japanese** corpus, suggesting that, unlike general linguistic capabilities (e.g., POS tagging), the sparsity and specificity of domain adaptation knowledge prevent effective in-language generalization. Finally, the stable knowledge acquisition performance across all recipes in the in-language evaluation indicates that using additional corpora unrelated to the target knowledge does not impair in-language knowledge acquisition.

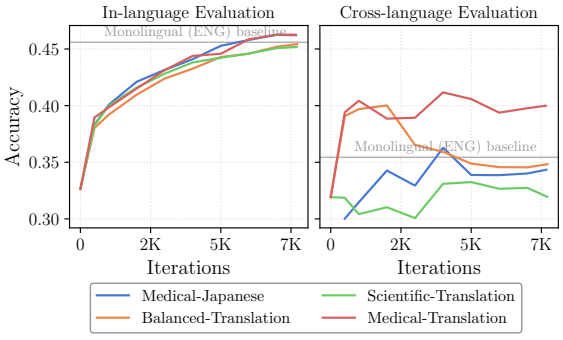


Fig. 12: Multilingual evaluation using translation datasets.

(2) Romanization is an effective strategy for transferring domain knowledge across languages. We report the evaluation results using C_T constructed from different transfer strategies. As shown in Figure 13, **Medical-Roman**, despite not relying on high-quality translation data, achieves cross-lingual performance comparable to **Medical-Translation**. This demonstrates that romanization provides a cost-efficient methodology for cross-lingual knowledge acquisition, especially in low-resource settings.

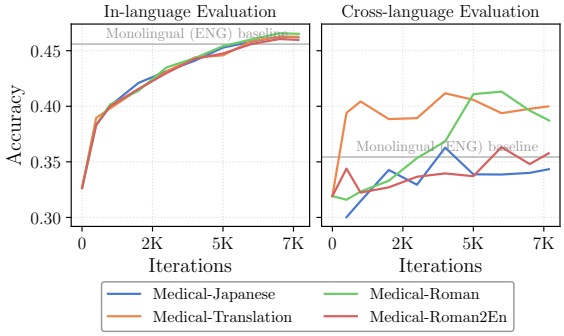


Fig. 13: Multilingual evaluation with diverse cross-lingual transfer strategies.

(3) Achieving cross-lingual transfer of domain knowledge remains challenging. Although both **Medical-Translation** and **Medical-Roman** yield improvements, the gains are still modest (around 5%) relative to the doubled training cost and additional dataset construction effort. This highlights the need for developing more efficient methods for cross-lingual domain knowledge transfer.

7. Conclusion

In this paper, we studied how LLMs acquire domain knowledge and transfer it across languages. We proposed **AdaXEval**, an adaptive multilingual evaluation pipeline that automatically generates datasets to assess domain knowledge across memorization, paraphrase, and cross-lingual transfer. Using AdaXEval, we analyzed the training dynamics of domain adaptation in biomedical corpora, revealing that knowledge acquisition is driven by **loss shielding**, where overfitting causes losses on irrelevant representations to increase faster than on relevant ones. Our **sequence perturbation analysis** further demonstrates the sensitivity of LLMs to training data, providing practical insights for designing more robust training paradigms. We also identified key factors that facilitate cross-lingual transfer, showing that the presence of cross-lingual tokens in closely related domains is crucial, while romanization offers a cost-effective alternative when high-quality translations are unavailable.

As future work, we aim to develop a more training-robust and efficient pretraining paradigm to achieve domain knowledge acquisition in a more effective and scalable manner.

[*3] `https://github.com/polm/cutlet`

References

[1] Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J. and Seo, M.: Towards Continual Knowledge Learning of Language Models (2022).

[2] Jang, J., Ye, S., Lee, C., Yang, S., Shin, J., Han, J., Kim, G. and Seo, M.: TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y., Kozareva, Z. and Zhang, Y., eds.), Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 6237–6250 (online), DOI: 10.18653/v1/2022.emnlp-main.418 (2022).

[3] Jiang, J., Cheng, F. and Aizawa, A.: Improving Referring Ability for Biomedical Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2024* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 6444–6457 (online), DOI: 10.18653/v1/2024.findings-emnlp.375 (2024).

[4] Laï-king, M. and Paroubek, P.: Pre-training data selection for biomedical domain adaptation using journal impact metrics (2024).

[5] Xie, Y., Aggarwal, K. and Ahmad, A.: Efficient Continual Pre-training for Building Domain Specific Large Language Models, *Findings of the Association for Computational Linguistics: ACL 2024* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 10184–10201 (online), DOI: 10.18653/v1/2024.findings-acl.606 (2024).

[6] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. and Smith, N. A.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., eds.), Online, Association for Computational Linguistics, pp. 8342–8360 (online), DOI: 10.18653/v1/2020.acl-main.740 (2020).

[7] Biesialska, M., Biesialska, K. and Costa-jussà, M. R.: Continual Lifelong Learning in Natural Language Processing: A Survey, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, (online), DOI: 10.18653/v1/2020.coling-main.574 (2020).

[8] Ke, Z., Shao, Y., Lin, H., Xu, H., Shu, L. and Liu, B.: Adapting a Language Model While Preserving its General Knowledge, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y., Kozareva, Z. and Zhang, Y., eds.), Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 10177–10188 (online), DOI: 10.18653/v1/2022.emnlp-main.693 (2022).

[9] Marashian, A., Rice, E., Gessler, L., Palmer, A. and von der Wense, K.: From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation, *Proceedings of the 31st International Conference on Computational Linguistics* (Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D. and Schockaert, S., eds.), Abu Dhabi, UAE, Association for Computational Linguistics, pp. 7087–7098 (online), available from ⟨https://aclanthology.org/2025.coling-main.472/⟩ (2025).

[10] Cheng, D., Huang, S. and Wei, F.: Adapting Large Language Models to Domains via Reading Comprehension (2024).

[11] Fang, Y., Li, X., Thomas, S. and Zhu, X.: ChatGPT as Data Augmentation for Compositional Generalization: A Case Study in Open Intent Detection, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting* (Chen, C.-C., Takamura, H., Mathur, P., Sawhney, R., Huang, H.-H. and Chen, H.-H., eds.), Macao, -, pp. 13–33 (online), available from ⟨https://aclanthology.org/2023.finnlp-1.2/⟩ (2023).

[12] Jiang, T., Huang, S., Luo, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., Zhang, Q., Wang, D. and Zhuang, F.: Improving Domain Adaptation through Extended-Text Reading Comprehension (2024).

[13] Fujinuma, Y., Boyd-Graber, J. and Kann, K.: Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 1500–1512 (online), DOI: 10.18653/v1/2022.acl-long.106 (2022).

[14] Gao, C., Hu, H., Hu, P., Chen, J., Li, J. and Huang, S.: Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Duh, K., Gomez, H. and Bethard, S., eds.), Mexico City, Mexico, Association for Computational Linguistics, pp. 6101–6117 (online), DOI: 10.18653/v1/2024.naacl-long.339 (2024).

[15] Zhang, Z., Lee, D.-H., Fang, Y., Yu, W., Jia, M., Jiang, M. and Barbieri, F.: PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 7025–7046 (online), DOI: 10.18653/v1/2024.acl-long.379 (2024).

[16] Yamada, I. and Ri, R.: LEIA: Facilitating Cross-lingual Knowledge Transfer in Language Models with Entity-based Data Augmentation, *Findings of the Association for Computational Linguistics: ACL 2024* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 7029–7039 (online), DOI: 10.18653/v1/2024.findings-acl.419 (2024).

[17] Hong, S., Lee, S., Moon, H. and Lim, H.: MIGRATE: Cross-Lingual Adaptation of Domain-Specific LLMs through Code-Switching and Embedding Transfer, *Proceedings of the 31st International Conference on Computational Linguistics* (Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D. and Schockaert, S., eds.), Abu Dhabi, UAE, Association for Computational Linguistics, pp. 9184–9193 (online), available from ⟨https://aclanthology.org/2025.coling-main.617/⟩ (2025).

[18] Jiang, J., Huang, J. and Aizawa, A.: JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models, *Proceedings of the 31st International Conference on Computational Linguistics* (Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D. and Schockaert, S., eds.), Abu Dhabi, UAE, Association for Computational Linguistics, pp. 5918–5935 (online), available from ⟨https://aclanthology.org/2025.coling-main.395/⟩ (2025).

[19] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A. and Natarajan, V.: Large Language Models Encode Clinical Knowledge (2022).

[20] Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., Deng, Z., Jiang, Y., Yao, Z., Li, H., Yu, Y., Hu, G., Huang, J., Liu, X.-Y., Lopez-Lira, A., Wang, B., Lai, Y., Wang, H., Peng, M., Ananiadou, S. and Huang, J.: FinBen: A Holistic Financial Benchmark for Large Language Models (2024).

[21] Zucchet, N., Bornschein, J., Chan, S., Lampinen, A., Pascanu, R. and De, S.: How do language models learn facts? Dynamics, curricula and hallucinations (2025).

[22] Chang, H., Park, J., Ye, S., Yang, S., Seo, Y., Chang, D.-S. and Seo, M.: How Do Large Language Models Acquire Factual Knowledge During Pretraining? (2024).

[23] Liu, Y., Wang, M., Kargaran, A. H., Körner, F., Nie, E., Plank, B., Yvon, F. and Schütze, H.: Tracing Multilingual Factual Knowledge Acquisition in Pretraining (2025).

[24] Zhao, X., Yoshinaga, N. and Oba, D.: Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (Graham, Y. and Purver, M., eds.), St. Julian's, Malta, Association for Computational Linguistics, pp. 2088–2102 (online), DOI: 10.18653/v1/2024.eacl-long.127 (2024).

[25] LLM-jp, :, Aizawa, A., Aramaki, E., Chen, B., Cheng, F., Deguchi, H., Enomoto, R., Fujii, K., Fukumoto, K., Fukushima, T., Han, N. et al.: LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs (2024).

[26] Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H. and Riedel, S.: Language models as knowledge bases?, *arXiv preprint arXiv:1909.01066* (2019).

[27] Hernandez, E., Li, B. Z. and Andreas, J.: Measuring and manipulating knowledge representations in language models, *arXiv preprint arXiv:2304.00740* (2023).

[28] Zhao, X., Yoshinaga, N. and Oba, D.: What Matters in Memorizing and Recalling Facts? Multifaceted Benchmarks for Knowledge Probing in Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2024* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 13186–13214 (online), DOI: 10.18653/v1/2024.findings-emnlp.771 (2024).

[29] Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B. and Wei, F.: Knowledge Neurons in Pretrained Transformers, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 8493–8502 (online), DOI: 10.18653/v1/2022.acl-long.581 (2022).

[30] Wang, X., Wen, K., Zhang, Z., Hou, L., Liu, Z. and Li, J.: Finding Skill Neurons in Pre-trained Transformer-based Language Models, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y., Kozareva, Z. and Zhang, Y., eds.), Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 11132–11152 (online), DOI: 10.18653/v1/2022.emnlp-main.765 (2022).

[31] Niu, J., Liu, A., Zhu, Z. and Penn, G.: What does the Knowledge Neuron Thesis Have to do with Knowledge?, *The Twelfth International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=2HJRwwbV3G⟩ (2024).

[32] Zhao, X., Jiang, Z. and Yoshinaga, N.: Neuron Empirical Gradient: Discovering and Quantifying Neurons' Global Linear Controllability, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Che, W., Nabende,

J., Shutova, E. and Pilehvar, M. T., eds.), Vienna, Austria, Association for Computational Linguistics, pp. 21446–21477 (online), DOI: 10.18653/v1/2025.acl-long.1041 (2025).

[33] Chen, Y., Cao, P., Chen, Y., Liu, K. and Zhao, J.: Journey to the Center of the Knowledge Neurons: Discoveries of Language-Independent Knowledge Neurons and Degenerate Knowledge Neurons (2023).

[34] Mondal, S. K., Sen, S., Singhania, A. and Jyothi, P.: Language-Specific Neurons Do Not Facilitate Cross-Lingual Transfer, *The Sixth Workshop on Insights from Negative Results in NLP* (Drozd, A., Sedoc, J., Tafreshi, S., Akula, A. and Shu, R., eds.), Albuquerque, New Mexico, Association for Computational Linguistics, pp. 46–62 (online), DOI: 10.18653/v1/2025.insights-1.6 (2025).

[35] Lin, P., Martins, A. and Schuetze, H.: A Recipe of Parallel Corpora Exploitation for Multilingual Large Language Models, *Findings of the Association for Computational Linguistics: NAACL 2025* (Chiruzzo, L., Ritter, A. and Wang, L., eds.), Albuquerque, New Mexico, Association for Computational Linguistics, pp. 4038–4050 (online), DOI: 10.18653/v1/2025.findings-naacl.225 (2025).

[36] Zhao, J., Zhang, Z., Gao, L., Zhang, Q., Gui, T. and Huang, X.: LLaMA Beyond English: An Empirical Study on Language Capability Transfer (2024).

[37] Gao, C., Hu, H., Hu, P., Chen, J., Li, J. and Huang, S.: Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Duh, K., Gomez, H. and Bethard, S., eds.), Mexico City, Mexico, Association for Computational Linguistics, pp. 6101–6117 (online), DOI: 10.18653/v1/2024.naacl-long.339 (2024).

[38] Shaham, U., Herzig, J., Aharoni, R., Szpektor, I., Tsarfaty, R. and Eyal, M.: Multilingual Instruction Tuning With Just a Pinch of Multilinguality, *Findings of the Association for Computational Linguistics: ACL 2024* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 2304–2317 (online), DOI: 10.18653/v1/2024.findings-acl.136 (2024).

[39] Zhang, Z., Lee, D.-H., Fang, Y., Yu, W., Jia, M., Jiang, M. and Barbieri, F.: PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 7025–7046 (online), DOI: 10.18653/v1/2024.acl-long.379 (2024).

[40] J, J., Dabre, R., M, A., Gala, J., Jayakumar, T., Puduppully, R. and Kunchukuttan, A.: RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 15593–15615 (online), DOI: 10.18653/v1/2024.acl-long.833 (2024).

[41] Purkayastha, S., Ruder, S., Pfeiffer, J., Gurevych, I. and Vulić, I.: Romanization-based Large-scale Adaptation of Multilingual Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 7996–8005 (online), DOI: 10.18653/v1/2023.findings-emnlp.538 (2023).

[42] Xhelili, O., Liu, Y. and Schuetze, H.: Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment, *Findings of the Association for Computational Linguistics: EMNLP 2024* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 11283–11296 (online), DOI: 10.18653/v1/2024.findings-emnlp.659 (2024).

[43] Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K. and Zou, A.: A framework for few-shot language model evaluation (2023).

[44] Cheng, D., Huang, S. and Wei, F.: Adapting Large Language Models via Reading Comprehension, *The Twelfth International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=y886UXPEZ0⟩ (2024).

[45] Jiang, T., Huang, S., Luo, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., Zhang, Q., Wang, D. and Zhuang, F.: Improving Domain Adaptation through Extended-Text Reading Comprehension (2024).

[46] Bond, F. and Kuribayashi, T.: The Japanese Wordnet 2.0, *Proceedings of the 12th Global Wordnet Conference* (Rigau, G., Bond, F. and Rademaker, A., eds.), University of the Basque Country, Donostia - San Sebastian, Basque Country, Global Wordnet Association, pp. 179–186 (online), available from ⟨https://aclanthology.org/2023.gwc-1.22/⟩ (2023).

[47] Morishita, M., Chousa, K., Suzuki, J. and Nagata, M.: JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, pp. 6704–6710 (online), available from ⟨https://aclanthology.org/2022.lrec-1.721⟩ (2022).

[48] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S., eds.), Portorož, Slovenia, European Language Resources Association (ELRA), pp. 2204–2208 (online), available from ⟨https://aclanthology.org/L16-1350/⟩ (2016).

[49] Neumann, M., King, D., Beltagy, I. and Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy, Association for Computational Linguistics, pp. 319–327 (online), DOI: 10.18653/v1/W19-5034 (2019).

[50] Social Computing Lab: MedNERN-CR-JA (Revision 13dbcb6) (2023).

[51] Team, Q.: Qwen3 Technical Report (2025).

[52] DeepSeek-AI: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (2025).

[53] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A. and Others: The Llama 3 Herd of Models (2024).

[54] Cheng, D., Gu, Y., Huang, S., Bi, J., Huang, M. and Wei, F.: Instruction Pre-Training: Language Models are Supervised Multitask Learners, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 2529–2550 (online), DOI: 10.18653/v1/2024.emnlp-main.148 (2024).

[55] Ishigami, R.: DeepSeek-R1-Distill-Qwen-32B-Japanese (2025).

Appendix

A.1 AdaXEval Generation Details

In this section, we provide details on the process used to generate the AdaXEval dataset from J-STAGE documents.

(1) Factual sentence filtering: We use the open-source NLP tool HanLP[*4] for Japanese sentence segmentation, and scispaCy[*5] [49] , 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[*6] [50] , a model specialized for NER in 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

[*4] https://github.com/hankcs/HanLP

[*5] https://github.com/allenai/scispacy

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

- `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 [51][*7], DeepSeek-R1-Distill-Llama-70B [52][*8], and Llama-3.3-70B-Instruct [53][*9]. For **Japanese biomedical triple extraction**, we use Qwen-32B, llm-jp-3.1-8x13b-instruct4[*10], and Llama-3.3-Swallow-70B-Instruct-v0.4[*11]. 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 finally use the extracted triples and their corresponding context sentences as input to instruct the strong close-source LLM, GPT-4.1, for generating queries 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.

A.2 Samples of AdaXEval Dataset

We randomly sample 10 examples from AdaXEval for both English and Japanese and display them in Table A·2 (English) and Table A·1 (Japanese).

A.3 Sequence Perturbation Analysis

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

A.3.1 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[*12], 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. Furthermore, the paraphrasing and translation are done by requesting GPT-4.1.

A.3.2 Perturbation Results

A.4 Domain Adaptation Training Details

A.4.1 Training Data Generation

In this study, to address the scarcity of bilingual domain corpora and enhance domain understanding, 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 [54], we adopt an instruction-pretraining strategy for continual domain adaptation.

(1) Regex-based pattern mining: [45] verified that by transforming raw corpora into reading comprehension texts, continual pretraining 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 each document, however, we sample only one template per type.

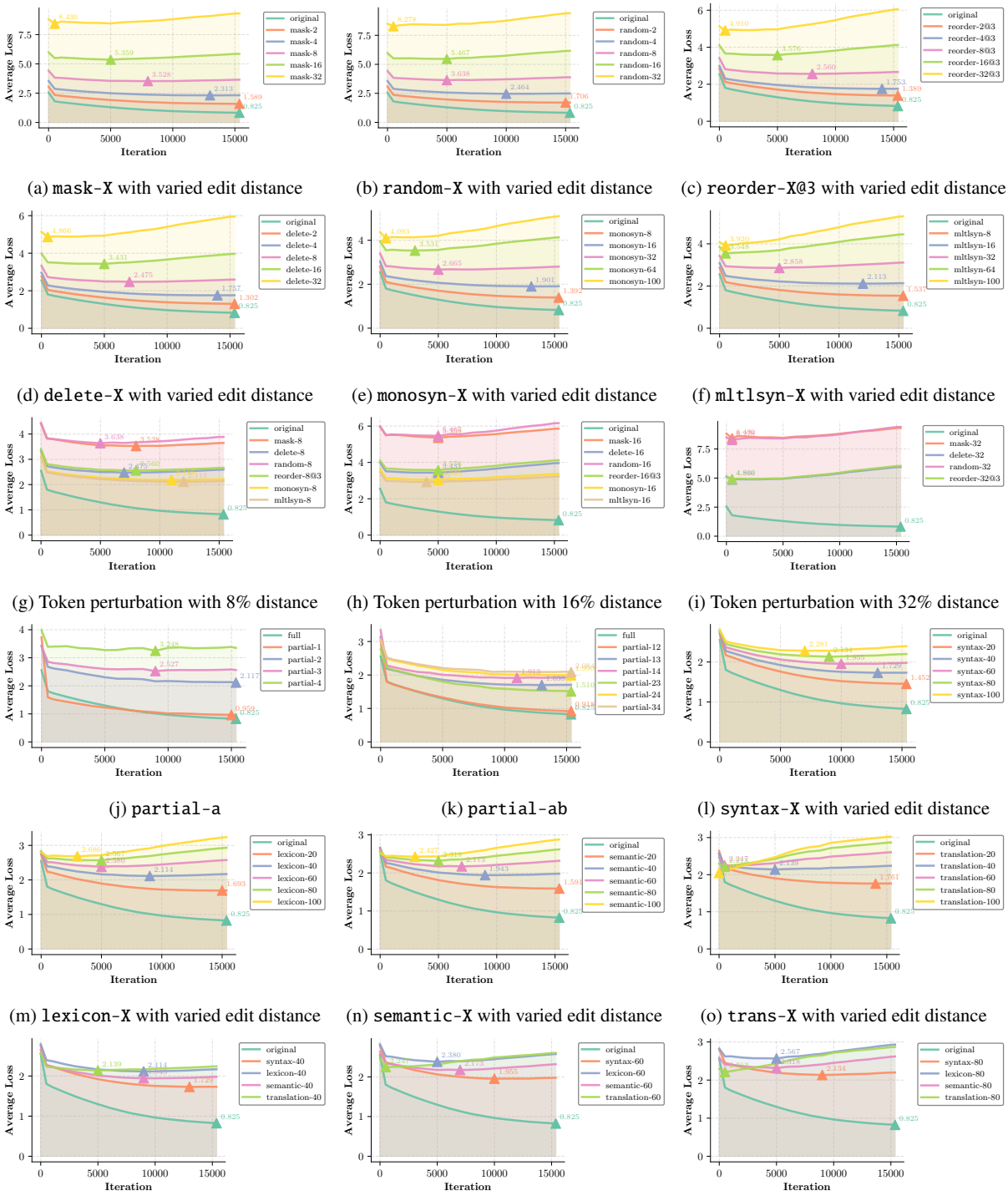
Specifically, each document in J-STAGE contains multiple metadata fields, including:

- `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 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.

[*7] https://huggingface.co/Qwen/Qwen3-32B
[*8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B
[*9] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[*10] https://huggingface.co/llm-jp/llm-jp-3.1-8x13b-instruct4
[*11] https://huggingface.co/tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4
[*12] https://github.com/WorksApplications/SudachiPy



(a) `mask-X` with varied edit distance (b) `random-X` with varied edit distance (c) `reorder-X@3` with varied edit distance

(d) `delete-X` with varied edit distance (e) `monosyn-X` with varied edit distance (f) `mltlsyn-X` with varied edit distance

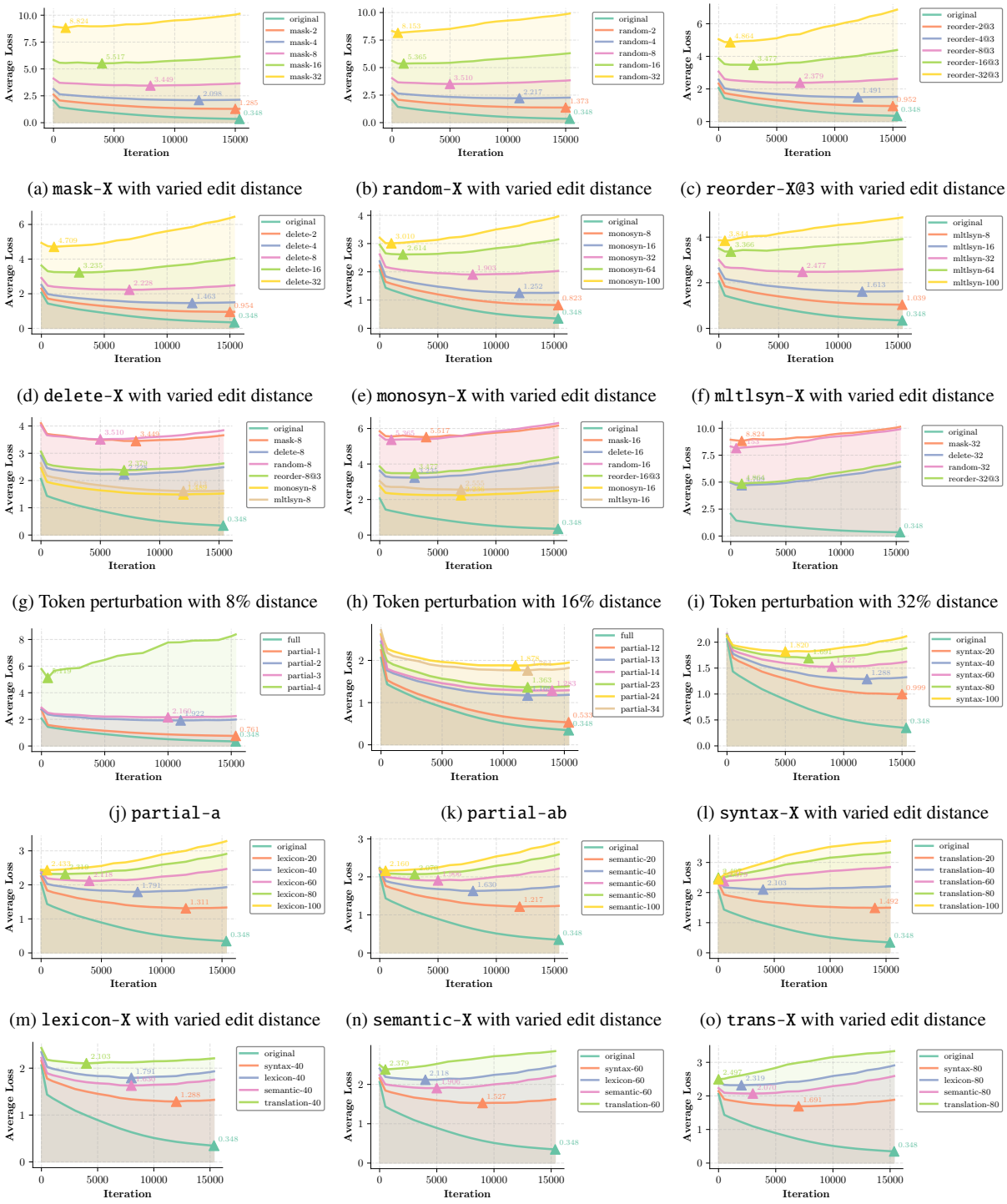
(g) Token perturbation with 8% distance (h) Token perturbation with 16% distance (i) Token perturbation with 32% distance

(j) `partial-a` (k) `partial-ab` (l) `syntax-X` with varied edit distance

(m) `lexicon-X` with varied edit distance (n) `semantic-X` with varied edit distance (o) `trans-X` with varied edit distance

(p) Sentence perturbation with 40% edit distance (q) Sentence perturbation with 60% edit distance (r) Sentence perturbation with 80% edit distance

Fig. A·1: The loss dynamics over all perturbation patterns on Japanese sequences.



(a) `mask-X` with varied edit distance (b) `random-X` with varied edit distance (c) `reorder-X@3` with varied edit distance

(d) `delete-X` with varied edit distance (e) `monosyn-X` with varied edit distance (f) `mltlsyn-X` with varied edit distance

(g) Token perturbation with 8% distance (h) Token perturbation with 16% distance (i) Token perturbation with 32% distance

(j) `partial-a` (k) `partial-ab` (l) `syntax-X` with varied edit distance

(m) `lexicon-X` with varied edit distance (n) `semantic-X` with varied edit distance (o) `trans-X` with varied edit distance

(p) Sentence perturbation with 40% edit distance (q) Sentence perturbation with 60% edit distance (r) Sentence perturbation with 80% edit distance

Fig. A·2: The loss dynamics over all perturbation patterns on English sequences.

(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 [55][*13] for Japanese QA generation. Noted that not all documents can successfully generate five QA pairs.

A.4.2 Continual Pretraining Settings

We conduct continual pretraining using **Megatron-LM** on the `llm-jp-3-13B` model. 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 pretraining while maintaining efficiency on large-scale biomedical and cross-lingual corpora.

A.5 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 § 6.2. Specifically, in this setting, $C_{\mathcal{K}}$ is composed of the Japanese monolingual training corpora, while we vary the data source of $C_{\mathcal{T}}$ to investigate efficient cross-lingual transfer. Evaluation results are shown in Figure A·3, A·4.

The evaluation results shown in Figure A·3 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.

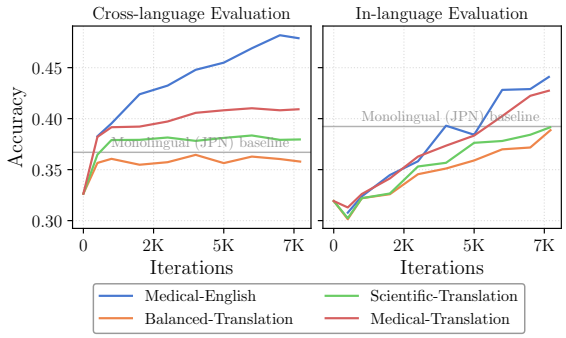


Fig. A·3: Japanese-to-English transfer evaluation using translation datasets.

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

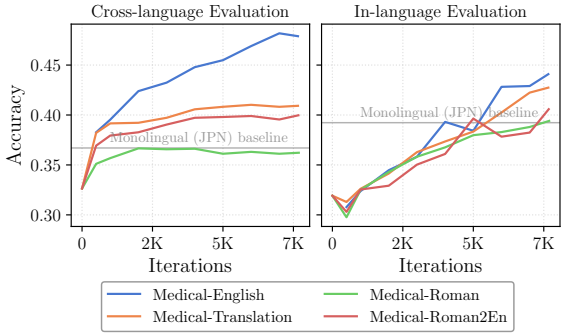


Fig. A·4: Japanese-to-English transfer evaluation with diverse cross-lingual transfer strategies.

A.6 AdaXEval Human Evaluation Guideline

A.6.1 Triple Quality Evaluation

Objective: Evaluate whether a triple containing `<subject, relation, object>` is:

- Correctly extracted from the original biomedical sentence (in-language correctness)
- Supported by a corresponding abstract in another language (cross-lingual correctness)

Input: For each item, you will be shown:

- Original Sentence in source language (e.g., English)
- Extracted Triple
- Related Abstract in target language (e.g., Japanese). This abstract is assumed to discuss the same or closely related content as the original sentence.

Evaluation Criteria

1) In-language Correctness: Does the triple accurately reflect the original English sentence?

Score	Description
2	The triple is fully correct – subject, relation, and object are all faithful to the sentence and do not introduce extra information.
1	Partially correct – partial elements are slightly incorrect, ambiguous, or underspecified.
0	Incorrect or unrelated – the triple misrepresents the sentence or expresses a different fact.

2) Cross-lingual Correctness: Can the same knowledge (triple) be found or inferred from the non-English abstract?

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.

A.6.2 Cloze Query Quality Evaluation

Objective: Evaluate whether the cloze queries generated from biomedical academic sentences and 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]`.

Evaluation Criteria

1) Faithfulness: 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

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.

2) Factual Consistency: When replacing `[BLANK]` with the object, does the prompt remain factually consistent with the original sentence and triple (no distortion of meaning or relationships)?

- **What it checks:** Whether the prompt introduces or implies factual distortion compared to the sentence/triple
- **Focuses on:** Semantic accuracy – does the prompt reflect the true meaning and factual relationship of the original sentence/triple

Score	Description
3	Fully consistent with the original facts.
2	Mostly consistent; minor factual ambiguity or soft misinterpretation.
1	Partially inconsistent; some facts or relationships are altered or key contextual information is missing.
0	Factually incorrect or misleading.

A.6.3 Paraphrase 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."
- 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

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

2) Semantic Equivalence: Does the question ask for the same information as the original sentence with the target triple?

- **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.

Score	Description
3	Fully equivalent: asks for the same object with no change in meaning.
2	Mostly equivalent: minor differences, but still asks essentially the same thing.
1	Partially equivalent: some meaning lost or changed; answer may differ.
0	Not equivalent: asks a different question or changes the relation/logic.

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

Score	Description
3	Excellent reformulation; uses different vocabulary and structure while maintaining meaning
2	Good variation; some different wording but follows similar structure
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

A.6.4 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.

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.

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

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 A·1: Samples of Japanese AdaXEval Dataset.

Sentence	Cloze Query	Paraphrase	Options	Answer ID
日本産のLepraria lobificansも同様の成分を持つが地衣体縁部が明瞭で裂片を持つので区別できる.	日本産のLepraria lobificansを区別するための特徴としては、[BLANK]が挙げられる。	日本産のLepraria lobificansを他の種と見分ける際に重要となる特徴は何ですか？	A. 地衣体の色の濃淡 B. 生育する基質の種類 C. 成長速度の違い D. 地衣体縁部の明瞭さと裂片の存在	D
胆道癌は肝門部領域胆管癌,遠位胆管癌,十二指腸乳頭部癌,胆嚢癌などに分類されるが,特に肝門部領域胆管癌は手術難易度が高い.	胆道癌は [BLANK] などに分類される。	胆道癌はどのような種類に分類されますか？	A. 肝細胞癌, 膵頭部癌, 胃癌, 食道癌 B. 腎細胞癌, 膀胱癌, 前立腺癌, 膵体尾部癌 C. 肺腺癌, 大腸癌, 直腸癌, 膵管癌 D. 肝門部領域胆管癌, 遠位胆管癌, 十二指腸乳頭部癌, 胆嚢癌	D
血中薬物濃度は循環血への薬物の流入速度と循環血からの消失速度によつて決まる値である.	血中薬物濃度は [BLANK] によって決まる。	血中薬物濃度はどのような要因によって決定されますか？	A. 循環血への薬物の流入速度と循環血からの消失速度 B. 肝臓での薬物の代謝速度 C. 薬物の投与経路の種類 D. 腎臓での薬物の再吸収率	A
このことよりSPAによる感染防御は体液性抗体のみでなく, 他の因子も関与していることが判る.	SPAによる感染防御は体液性抗体のみでなく、[BLANK]も関与していることが判る。	SPAによる感染防御には体液性抗体以外にどのような要素が関与していると考えられますか？	A. 補体 B. マクロファージ C. T細胞 D. 他の因子	D
大腸癌の進展様式として腸間膜静脈内に腫瘍塞栓を形成することは稀である.	大腸癌の進展様式として[BLANK]は稀である。	大腸癌の進展様式のうち、稀にみられる現象は何ですか？	A. 腸間膜静脈内に腫瘍塞栓を形成すること B. 腸間膜動脈への直接浸潤 C. リンパ節転移の発生 D. 腹膜播種の形成	A
これとは逆に、LDL受容体を欠く場合、例えばホモ接合型の家族性高コレステロール血症あるいは腎癌の場合に、LDLはLDL-受容体非依存性経路 (“スカベンジャ-経路”) により細胞内に入り、GalT-2を促進的に調節しラクトシルセラミドの細胞内レベルを増加させる。	LDL受容体の欠損により、[BLANK]。	LDL受容体が欠損している場合、細胞内でどのような現象が起こりますか？	A. LDLがスカベンジャ経路で細胞内に入り、GalT-2を促進的に調節しラクトシルセラミドの細胞内レベルを増加させる B. LDLが細胞外に蓄積し、GalT-2の活性が低下する C. LDLが通常の受容体経路で細胞内に入り、ラクトシルセラミドのレベルが減少する D. LDL受容体の発現が増加し、細胞内コレステロールが減少する	A
超音波ガイド下穿刺を困難にする要因の一つに,穿刺針や超音波プローブが静脈を潰すことがある.	穿刺針や超音波プローブが静脈を潰すことは、[BLANK]である。	穿刺針や超音波プローブが静脈を潰してしまうことは、どのような要因の一つと考えられていますか？	A. 穿刺部位の感染リスクを減少させる要因の一つ B. 超音波ガイド下穿刺の成功率を高める要因の一つ C. 超音波ガイド下穿刺を困難にする要因の一つ D. 静脈の可視化を容易にする要因の一つ	C
また, 新生児では炎症性サイトカインの産生能も未熟であるが, TCRを介した反応性, type 1, type 2サイトカインの産生能は成人T細胞と同等の機能を有する.	新生児のTCRを介した反応性およびtype 1, type 2サイトカインの産生能は[BLANK]。	新生児のTCRを介した反応性やtype 1, type 2サイトカインの産生能は、成人T細胞と比較してとのような機能を持っていますか？	A. 成人T細胞と同等の機能を有する B. 成人T細胞とは異なる特異的な機能を有する C. 成人T細胞よりも著しく低い機能を示す D. 成人T細胞よりも高い活性を持つ	A
この際に,複数のデフォーカス面におけるスポット像を用いることでスポット像の空間分解能を補足する.	複数のデフォーカス面におけるスポット像を用いることで、[BLANK]を補足することができる。	複数のデフォーカス面で得られたスポット像を利用することで、どのような特性を補うことができますか？	A. スポット像の色再現性 B. スポット像の明るさ C. スポット像の時間分解能 D. スポット像の空間分解能	D
血管周皮細胞腫は血管の周皮細胞から発生する血管系腫瘍のひとつで,病理組織学的に悪性所見が乏しくても,再発や転移の頻度が高い.	血管周皮細胞腫は [BLANK] から発生する腫瘍である。	血管周皮細胞腫はどの細胞由来の腫瘍ですか？	A. 線維芽細胞 B. 平滑筋細胞 C. 血管内皮細胞 D. 血管の周皮細胞	D

Table A·2: Samples of English AdaXEval Dataset.

Sentence	Cloze Query	Paraphrase	Options	Answer ID
Hyporesponsiveness in TRH test and abnormal pattern in clomiphene test suggest the hypothalamic-pituitary axis dysfunction as well as atrophy of testis.	Hyporesponsiveness in TRH test and abnormal pattern in clomiphene test suggest [BLANK].	What conditions are indicated by both a reduced response in the TRH test and an abnormal result in the clomiphene test?	A. adrenal insufficiency and thyroid gland enlargement B. pituitary adenoma and hyperplasia of testis C. gonadal hyperfunction and normal hypothalamic activity D. hypothalamic-pituitary axis dysfunction and atrophy of testis	D
3) Adrenocorticotropic hormone (ACTH)-producing adenoma (Cushing's disease) TSS is the first-choice treatment for Cushing's disease.	TSS is the first-choice treatment for [BLANK].	For which medical condition is transsphenoidal surgery (TSS) considered the preferred initial therapy?	A. Cushing's disease B. nonfunctioning pituitary adenoma C. prolactinoma D. acromegaly	A
Clinical research is classified into patient-oriented research, in which subjects are patients, and disease-oriented research, in which patient-derived tissue, cell and blood samples are investigated.	Clinical research is classified into [BLANK].	Into which two main categories is clinical research typically divided?	A. population-based research and experimental research B. patient-oriented research and disease-oriented research C. basic science research and translational research D. laboratory-based research and epidemiological research	B
The rest of the gastric wall around the tumor is dissected with ultrasonic coagulating shears and the tumor is placed in a collection bag.	The gastric wall around the tumor is dissected with [BLANK].	Which surgical instrument is used to dissect the gastric wall surrounding a tumor during this procedure?	A. ultrasonic coagulating shears B. conventional surgical scissors C. laser ablation devices D. electrosurgical forceps	A
Spine calcium imaging, which records synaptic inputs as calcium transients at individual spines using calcium ion-sensitive fluorophores, is a unique method for studying the spatiotemporal patterns of synaptic input.	Spine calcium imaging is a method for studying [BLANK].	What aspect of neural activity does spine calcium imaging specifically help researchers investigate?	A. molecular mechanisms of neurotransmitter release B. long-term potentiation in neural circuits C. structural changes in dendritic spines D. spatiotemporal patterns of synaptic input	D
Amylopectin, the major branched fraction is itself a highly organized molecule displaying a succession of clusters of glucans packed in crystal arrays.	Amylopectin is a highly organized molecule displaying [BLANK].	What structural feature does amylopectin exhibit as a highly organized molecule?	A. multiple helices of amylose intertwined with protein complexes B. a series of repeating disaccharide units forming amorphous regions C. a network of linear glucose chains arranged in parallel layers D. a succession of clusters of glucans packed in crystal arrays	D
At the tip of the germ-tubes an appressorium is formed, and an infection hypha is sent, which penetrates the epidermis at the anticlinal wall of the epidermal cells.	At the tip of the germ-tubes, an [BLANK] is formed.	What specialized structure is produced at the end of germ-tubes?	A. sporangiophore B. haustorium C. appressorium D. conidium	C
The results of multivariate Cox hazard regression analyses also showed that age (HR 1.05, per 1-year increase), presence of hypertension and diabetes (HR 2.99), ... (omitted)	Age is an independent predictor of [BLANK].	According to the analysis, what outcome is independently predicted by a patient's age?	A. all-cause mortality B. hospital readmission C. functional disability D. stroke recurrence	A
It is difficult for large molecular size organic liquids having a single hydrogen bond-forming group to swell wood.	It is difficult for large molecular size organic liquids having a single hydrogen bond-forming group to swell [BLANK].	Which material is challenging to swell when using large organic liquids with only one hydrogen bond-forming group?	A. plant stems B. wood C. cellulose fibers D. cotton fabric	B
The prevalence of teeth with NCCLs was greater in the maxilla than mandible.	Teeth with NCCLs have higher prevalence in [BLANK].	In which location are teeth affected by NCCLs found more frequently when comparing the upper and lower jaws?	A. maxilla compared to mandible B. mandible compared to maxilla C. anterior region compared to posterior region D. premolars compared to molars	A