

Exemplar Choice Matters for Morphological Reasoning: Model Confidence vs. Exemplar Similarity

Tsedeniya Kinfe Temesgen^{1,2} and Marion Di Marco^{1,2} and Alexander Fraser^{1,2}

¹School of Computation, Information and Technology,

Technische Universität München (TUM)

²Munich Center for Machine Learning

{tsedeniya.temesgen|marion.dimarco|alexander.fraser}@tum.de

Abstract

LLMs have shown impressive linguistic competencies in a wide range of NLP tasks, yet morphological generalization remains challenging. Models can benefit from in-context learning, where the choice of few-shot examples significantly influences the outcome. We probe LLMs’ morphological capabilities by assessing how two targeted in-context exemplar selection strategies, *Iterative Confidence-Based Exemplar Selection (i-CES)* and *Similarity-Based Exemplar Selection (SES)*, influence the model’s ability for derivational and inflectional modeling across multiple languages. Our experiments with LLaMA 3.3-70B-instruct and Qwen2.5-32B-Instruct models indicate that *SES* is more effective for accessing latent linguistic knowledge, showing a substantial performance advantage over *i-CES* in both derivational and inflectional tasks. In contrast, *i-CES* offers no clear advantage over random sampling.

1 Introduction

Large language models (LLMs) have demonstrated remarkable competence across a variety of NLP benchmarks. Despite these advances, a growing body of work shows that this performance does not extend to robust linguistic generalization (Ismayilzada et al., 2025; Wilson et al., 2023; Weissweiler et al., 2023; Goldman et al., 2022), showing the limitations in understanding core structural properties of language.

Morphological competence is essential for interpreting and generating novel word forms, especially in morphologically rich languages where grammatical meaning is expressed through inflection. Prior work demonstrates that models struggle to handle neologisms and unseen morphological variants (for example, Zheng et al. (2024)), highlighting a gap between surface fluency and structural understanding.

	Derivation	Inflection
exemplar	approval → approve (-al)	leaf + PL → leaves
query	retrieval → retrieve (-al)	knife + PL → knives

Table 1: Derivational segmentation and inflectional generation transformations that exhibit similar patterns between the example and the query.

However, explicitly providing morphological information can improve performance for different NLP tasks (Zhang et al., 2024; Yang et al., 2025) and partially compensate for model limitations. One effective way to implement this at inference time is through **In-Context Learning (ICL)**. With carefully selected exemplars, predictions can be guided without retraining (Ismayilzada et al., 2025; Blevins et al., 2023; Brown et al., 2020), which is particularly valuable in low-resource settings.

Muradoglu et al. (2024) and Muradoglu and Hulden (2022) explore ICL through confidence-based exemplar selection for inflectional morphology. Similarly, Margatina et al. (2023) demonstrate that selecting semantically similar examples can improve performance on classification and multi-choice tasks on the Crossfit benchmark (Ye et al., 2021). Building on these findings, we study a confidence- and similarity-driven retrieval approach for two morphological tasks: obtaining a derivational segmentation and generating an inflected form from a lemma and inflectional features. Our experiments focus on a diverse set of low-resource languages (Amharic, Galician, Irish, Mongolian, and Zulu), while high-resource languages (English, French, Hungarian, Russian, and Spanish) are included as baselines for comparison. While prior studies largely focus on inflection, derivational morphology remains understudied, particularly in low-resource languages where derivational datasets and benchmarks are limited.

Consequently, it remains unclear how well LLMs can generalize derivational patterns under limited supervision.

In this work, we introduce **Iterative Confidence-Based Exemplar Selection (i-CES)** and **Similarity-Based Exemplar Selection (SES)** as mechanisms for enabling LLMs to better access and leverage morphological knowledge for both **inflectional** and **derivational** analyses in multiple languages. The two strategies target complementary aspects of exemplar selection: i-CES leverages model confidence to adjust internal morphological representations, while SES retrieves surface-aligned examples relevant to the input query.

For **i-CES**, we iteratively select in-context examples using the model’s posterior probabilities as a confidence signal. Specifically, we consider two strategies: (i) selecting instances where the model makes **high-confidence (HC)** but **incorrect** predictions. We include these instances with their correct labels as demonstrations; and (ii) selecting instances where the model shows **low confidence (LC)**, and using them as demonstrations to strengthen weakly represented morphological patterns. **SES** aims at selecting optimal exemplars based on similarity to each specific test query: we consider **Semantic Similarity (SM)** and surface-form similarity measured by the **Longest Common Substring (LCS)** and **Lemma-Feature match (LF)**¹ where we consider examples that share lemmas and inflectional features with the input query. Table 1 shows that exemplars aligned with the query word can provide a useful template for handling novel inputs.

Our findings on two instruction-tuned models, LLaMA 3.3-70B and Qwen2.5-32B, for both inflectional and derivational morphology, show that exemplar selection plays a crucial role in eliciting morphological knowledge from LLMs. Query-specific SES consistently outperforms i-CES and random baselines across languages and in both models. For the inflection task, we observe a significant performance gain with the LF method. For the derivational task, the LCS achieved the best scores, with SM following closely behind.

¹This method is only applied for generating inflected forms where morpho-syntactic features are part of the input.

2 Related work

2.1 Morphology and LLMs

Recent research finds that LLMs possess substantial linguistic knowledge across morphology, syntax, and semantics, with evidence that models like GPT-3 exhibit syntactic and semantic competence beyond simple memorization (Blevins et al. (2023)). However, other studies show their limitations.

He et al. (2024) and Waldis et al. (2024) show that LLMs continue to struggle at the syntax–semantics interface, with semantic reasoning and discourse-level understanding, and with several morphological phenomena. For morphologically rich languages, models such as GPT-4-turbo fail to handle complex morphology reliably; for example, they struggle to provide accurate analyses of rare and complex Finnish word forms (Moisio et al. (2024)).

Another research strand shows that while LLMs often can handle frequent and regular morphological patterns, they struggle with novel, complex, or nonce word forms, particularly in low-resource languages (Anh et al. (2024); Di Marco and Fraser (2024); Weissweiler et al. (2023); Goldman et al. (2022)). Several studies argue that models rely on surface-level heuristics rather than linguistic rules (Weller-Di Marco and Fraser (2024)), and overall morphological generalization remains a major limitation of current LLMs (Ismayilzada et al. (2025); Kodner et al. (2023); Temesgen et al. (2025)).

2.2 ICL exemplars selection

LLMs’ linguistic capabilities are highly sensitive to inference-time design choices such as prompt formatting and the number of in-context examples. Prior work has shown that providing additional demonstrations can improve performance on out-of-distribution data (Ismayilzada et al., 2025) across various NLP tasks (Blevins et al., 2023). However, the effectiveness of this approach depends on the task and the morphological complexity of the language (Brown et al., 2020).

A substantial body of research explores ways to select effective exemplars. Zhang et al. (2022) use reinforcement learning to iteratively select and annotate informative examples; and Chen et al. (2023) show that pseudo-inputs and pseudo-labels generated in the zero-shot settings can serve as useful demonstrations. Similarly, Lee et al. (2025) evaluate rationales through follow-up questions and train

models on these rationales, while Li et al. (2024) iteratively reuse high-confidence translation pairs as exemplars, yielding substantial improvements over zero-shot.

Confidence-based selection methods have proven particularly effective. Muradoglu et al. (2024) investigate confidence-driven exemplar selection, particularly low-probability, for multilingual inflectional analysis using encoder-decoder models. Similarly, Muradoglu and Hulden (2022) examine high-confidence instances alongside entropy-based and random selection strategies. Together, these findings show that carefully selected confidence-based examples can guide models more effectively than randomly sampled demonstrations.

Another line of work highlights semantic similarity and diversity-based selection. Margatina et al. (2023) identify uncertainty (perplexity-based) sampling as a robust baseline; in addition, selecting semantically similar examples provides significant performance improvements. Mirbostani et al. (2023) use entropy and clustering to identify informative demonstrations, reporting that random selection leads to passive learning, whereas structured selection achieves peak performance with fewer examples.

Baldrige and Palmer (2009) compare entropy-based, random, and sequential sampling strategies for annotation, showing that example selection strongly influences learning outcomes. Related work on gloss prediction and low-resource language learning (Ginn et al., 2024; Yang et al., 2024; Elsnar and Liu, 2025) similarly demonstrate that carefully chosen examples and structured guidance improve model performance. In-context learning is explored for clause-level morphology and morphosyntactic phenomena (Su et al., 2024; Haley, 2024), as well as for downstream tasks such as machine translation (Zebaze et al., 2025; Ramos et al., 2025) and grammatical error correction (Sorokin and Nasyrova, 2025), highlighting that ICL is a promising approach for eliciting linguistic knowledge.

Building on prior work, our study examines similarity-based and confidence (logit) based retrieval for selecting exemplars for morphological analysis. Whereas earlier research has focused primarily on inflectional paradigms, derivational morphology has received comparatively little attention. Furthermore, because annotated datasets for derivational morphology are often unavailable for low-

resource languages, we investigate whether LLMs can be effectively prompted using a limited set of curated, labeled examples. Consequently, we apply these selection strategies to both derivational and inflectional tasks in an in-context learning setting.

3 Tasks and Datasets

ICL has shown to improve model performance; this is particularly relevant in settings where there is limited annotated data available. Existing morphological benchmarks such as SIGMORPHON² and UNIMORPH³ cover many languages. However, their coverage for low-resource languages is sparse and primarily restricted to inflection, often without morpheme segmentation. Derivational morphology is even less represented, with available resources concentrated on high-resource languages. We center our study on low-resource languages and include high-resource languages for comparison.

Languages and Models We study a total of 10 typologically different languages.

Low-resource languages cover Galician-gal (Romance), Irish-gle (Celtic), Mongolian-mon (Mongolic), Amharic-amh (Semitic), and Zulu-zul (Bantu). High-resource languages include French-fra and Spanish-spa (Romance), Russian-rus (Slavic), English-eng (Germanic), Hungarian-hun (Uralic).

We evaluate two instruction-tuned models, LLaMA-3.3-70B and Qwen2.5-32B, using greedy decoding.

Dataset We use publicly available datasets, sourced from the SIGMORPHON 2022 SHARED TASK⁴ (Girrbach, 2022), UNIMORPH⁵ (Batsuren et al., 2022), and MORPHYNET⁶ (Batsuren et al., 2021), restricting our scope to languages with both derivational and inflectional annotations. For high-resource languages, we extract derivational segmentation data from SIGMORPHON. We select instances based on their annotated morphological categories (e.g., *Inflection only*, *Derivation only*, *Root*; Cf. Appendix B) and retain instances labeled “*Derivation only*”. For low-resource languages, we sourced data from UNIMORPH, while Mongolian data is retrieved from MORPHYNET, Amharic from

²<https://sigmorphon.github.io/sharedtasks/>

³<https://unimorph.github.io/>

⁴<https://github.com/sigmorphon/2022SegmentationST>

⁵<https://unimorph.github.io/>

⁶<https://github.com/kbatsuren/MorphyNet>

Temesgen et al. (2025), and Zulu from Gaustad and Puttkammer (2022). For inflectional tasks, we use data from UNIMORPH for all languages except Mongolian, which we source from MORPHYNET.

Across all datasets, we ensure that all surface forms are unique and exclude lexical ambiguities where a word can map to multiple parts of speech⁷. To provide equal conditions for all languages, with a low-resource setting with limited resources in mind, we retain 1K examples for training. The test set is capped at 1K examples. If splits are unavailable, we split the data into training and test sets of up to 1K examples each (Cf. Appendix B).

Tasks (i) *Derivation task*: models are prompted to provide the derivational segmentation of a given input word. (ii) *Inflection task*: models are asked to generate an inflected form from a lemma and morphosyntactic features, as illustrated in Table 1. The models are expected to provide the canonical segmentation, resolving orthographic or phonological changes rather than literal surface substrings. (Cf. Appendix D for prompt instructions).

Evaluation We use **Accuracy** (exact match) to evaluate the generation and segmentation tasks.

4 Exemplar Selection Strategies

The selection strategies aim to retrieve optimal demonstration examples for in-context learning during inference. The prompt is constructed incrementally by extending the set of exemplars by one in each iteration until the maximum of $K = 10$ is reached.

4.1 Iterative Confidence-Based Exemplar Selection

We use prediction confidence as a selection criterion. Confidence is the model’s assigned probability (or normalized log-probability) of the predicted output sequence. This aims at strengthening the model’s general comprehension of the task and the underlying morphological structures. We investigate two different approaches.

High confidence, but wrong (HC) To identify systematic weaknesses in the model’s predictions, we select examples for which the model expresses strong confidence despite being incorrect. These instances reveal cases where the model has learned misleading morphological patterns and therefore

Query	Model output	Prob.	Gold answer
exhilaratingly	exhilarate ly	0.99	exhilarate ing ly
nonconformably	non conform able ly	0.99	non conform able ly
...	...	...	...
haptenic	hapten ic	4.5e-3	hapten ic
Nerdly	nerd ly	3.4e-3	nerd ly

Table 2: Demonstrations ranked by model confidence using LLaMA in the zero-shot setting. HC selects high-confidence examples (e.g., *exhilaratingly*), while LC selects low-confidence examples (e.g., *nerdly*). Selected examples are removed from the training pool and used as few-shot demonstrations. The process iterates until $K = 10$.

demonstration examples can provide informative corrective signals. The model is first prompted on the training pool, and all predictions are ranked by their posterior confidence scores. At each iteration, we select one instance with an incorrect prediction but high confidence ($p > 0.9$) and add it to the demonstration set paired with its correct label (See the example in Table 2). The updated demonstration set is then used to prompt the remaining training pool and the test set; predictions on the training pool from this iteration are used to select the next example. One example is added per iteration.

Low confidence (LC) To reinforce morphological patterns for which the model shows uncertainty, we select instances assigned low posterior confidence. These examples indicate a weakly represented morphological pattern and may therefore benefit from additional guidance through demonstrations. From the predictions on the training pool, we select instances with low confidence and add these instances, paired with their correct label, to the demonstration set (See the example in Table 2). As in HC, one example per iteration is selected and subsequently removed from the training pool.

4.2 Similarity-Based Exemplar Selection

In contrast to the confidence-based strategies, which build a single global example set shared across all test items, similarity-based selection provides examples relevant to each test query, constructing a query-specific set of examples.

Longest common substring (LCS) Being interested mainly in the construction of words

⁷For example, *judge* can be either a noun or a verb.

at the surface level, we base the selection of few-shot examples on string similarity. As a result, we expect words that share either prefixes/suffixes or the stem with the query word.⁸ For instance, for the word *astroseismicity* LCS retrieves the following top-4 exemplars: hypoendemicity, metapolycentricity, thermoelectricity, astrophotometrical.

Semantics (SM) Following previous work, e.g. Li et al. (2024), we base the selection of few-shot instances on semantic similarity using fasttext embeddings⁹ (Bojanowski et al., 2017), where similarity is measured via cosine distance. While this does not primarily look at similarity at the level of word construction, it aims at finding words that are related, but not accessible via string match. Similar words in an embedding space also tend to share morpho-syntactic features such as word class. As fasttext also relies on subword statistics, substring similarity might play an indirect role, especially when considering that many of the words in our dataset are uncommon and infrequent.

For instance, given a query word *incompatibility* (noun), the most semantically similar words retrieved from the training set include obsolescence, uninterest, and reliability, all of which are nouns.

Lemma-Feature Match (LF) for the task of generating inflected forms: To retrieve demonstrations that share morphological properties with the query, we rank training examples by the number of overlaps computed jointly over the lemma and the inflectional feature set, giving equal weight to each. Examples with the highest overlap scores are used as demonstrations. Examples with the *same lemma* but different features are expected and intended to illustrate general inflectional properties of that word, such as stem changes. Examples with shared (or identical) morphological features, but a *different lemma*, are to demonstrate the inflections corresponding to the feature set in the query. The exact test instance, i.e., the same lemma and set of features, is never present in the training set. LF is applied only to the inflection task where we have access to the lemma and the inflectional features, which is not the case in the segmentation

task. This approach aims at selecting demonstrations that share similar inflectional patterns with the query, allowing us to evaluate the model’s ability to generate the correct inflected form given the full set of morphological features.

The example (French verbal inflection) below illustrates how an identical set of features can help the model to derive the correct inflection suffix:

Example

Lemma: grincer (*gloss: squeak*)
 Feature(s): Indicative, Past, Perfective, Second person, Plural
 Answer: grinçâtes
 (*gloss: (you) squeaked*)

Query

Lemma: constater (*gloss: notice*)
 Feature(s): Indicative, Past, Perfective, Second person, Plural
 Answer:
 Gold answer: constatâtes
 (*gloss: (you) noticed*)

Random (R) We additionally include random sampling as a baseline for comparison.

5 Results

Derivational task: Shown in Figure 1, SES demonstrates substantial improvements over both i-CES and the random baseline for both LLaMA and Qwen models. The second best approach is SM. In many cases, SM performs close to LCS, and for some languages, particularly for Irish and French, SM outperforms LCS by a small margin in LLaMA. Note that SM was not evaluated for Zulu, which is currently not supported by fastText.

A significant performance gap is observed between SES and i-CES. Both LCS and SM clearly outperform the confidence-based methods except for English, where the differences between the retrieval approaches are relatively small compared to other languages. The other exception is Amharic, where SM performs on par with random (R); in contrast, for other languages, SM generally achieves performance close to the best method, LCS.

Moreover, i-CES does not provide a clear advantage over R in many cases, as their performance is often comparable. However, for the LLaMA model,

⁸We also evaluated Levenshtein distance as an alternative similarity metric and observed comparable average performance; we therefore chose LCS (See Appendix C).

⁹<https://fasttext.cc/docs/en/crawl-vectors.html>

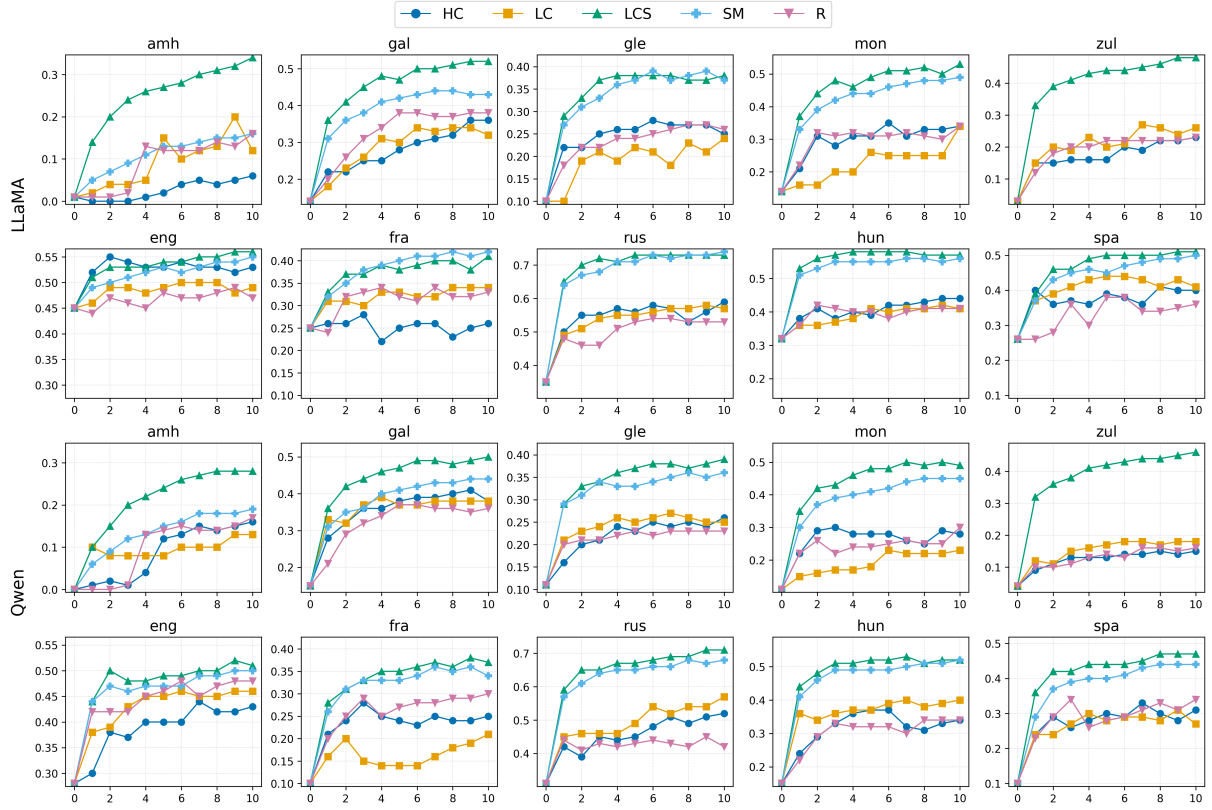


Figure 1: **Derivational task: Accuracy** of different retrieval strategies under varying number of shots. HC: high-confidence; LC: low-confidence; LCS: Longest Common Substring; SM: Semantic; R: Random

HC yields the lowest scores for Amharic, Galician, and French, while LC performs worst for Irish and Mongolian. Similarly, for the Qwen model, HC is lowest for English, and LC is the lowest for French and Mongolian. Additionally, at the highest shot, the performance of HC and LC tends to converge. This convergence is observed in languages such as Galician, Irish, and Spanish in both models. Zulu and Mongolian also converge in LLaMA.

Inflectional task: Shown in Figure 2, LF achieves the best performance across all languages except English, where the performance differences between the selection strategies are minimal. However, LCS, SM, and i-CES have a limited impact on performance on both models. For low-resource languages such as Amharic, Galician, and Zulu, the gains remain minimal, and improvements are also limited for high-resource languages such as Hungarian and Spanish.

Within SES strategies, the SM method achieves the best results for Irish, outperforming the random baseline in LLaMA and LCS in Qwen. A larger divergence appears in Mongolian with LLaMA, where HC performs the worst among the methods. In contrast, Qwen shows divergence at higher shots,

with LCS achieving the strongest results.

Overall, performance remains broadly similar across example selection strategies for both models. English and French achieve accuracy scores exceeding 80%, possibly due to their relatively simpler grammatical inflection. In Russian and Hungarian, LLaMA outperforms Qwen, likely reflecting stronger prior task knowledge; while both models struggle in low-resource Amharic and Irish languages.¹⁰

6 Discussion

We investigate the influence of example selection strategies on model performance and the resulting error patterns. In particular, we analyze HC and LCS on the 10-shot results, with a focus on lemma identification, i.e., the recovery of canonical forms when segmenting in the derivational task. To evaluate how selection strategies perform across varying levels of morphological difficulty, we categorize target words according to their segmentation complexity. We distinguish between two types: surface-level segmentation, in which

¹⁰We report the exact accuracy scores in Appendix H.

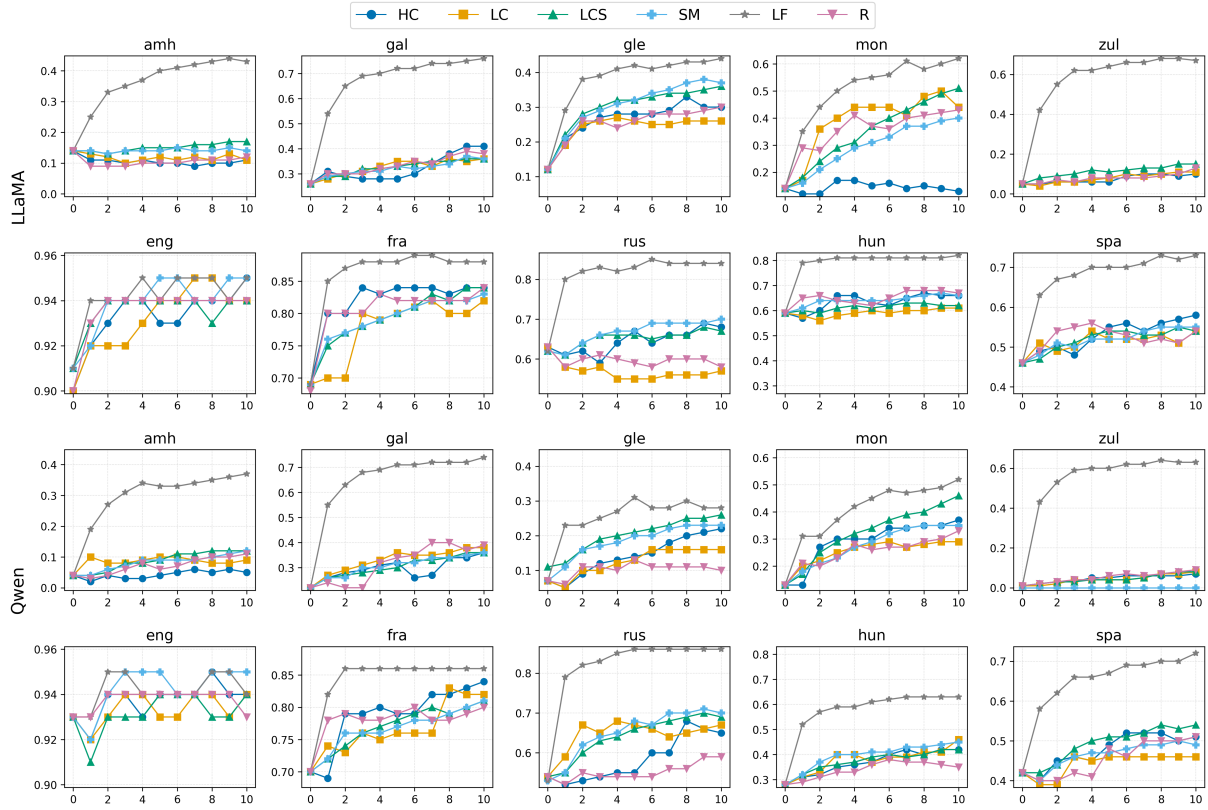


Figure 2: **Inflection task: Accuracy** of different retrieval strategies under varying number of shots. HC: high-confidence; LC: low-confidence; LCS: Longest Common Substring; SM: Semantic; LF: Lemma-Feature Match; R: Random

the lemma is recovered by stripping affixes, and canonical segmentation, which requires orthographic transformations that account for internal word formation processes. While surface-level segmentation is relatively straightforward and can be resolved through pattern matching, canonical segmentation is more challenging, as it requires the model to infer underlying morphological alternations. As shown in Table 7, both LCS and HC methods show substantially higher error rates on canonical segmentation than on surface-level segmentation across all languages, indicating that surface-level forms are considerably easier for the models to identify correctly.

To further characterize model failures, we analyzed morpheme fertility – defined as the number of morphemes produced when a word is decomposed into its lemma and affixes. Figure 4 shows that morpheme fertility scores are comparable across both models. However, HC tends to under-segment words in both surface and canonical forms relative to LCS. For canonical forms, both LCS and HC are less prone to over-segmentation than to under-segmentation.

In the inflection task, the behavior of the example selection strategies diverges from the patterns observed in the derivational task. While the distance-based SES method prioritizes surface similarity between lemmas and i-CES relies on model confidence, neither approach explicitly considers inflectional features. Consequently, the LF strategy significantly outperforms the other methods, as it directly incorporates inflectional features into the example selection process. We hypothesize that SM might also perform reasonably well in this setting, since it tends to retrieve examples whose lemmas belong to the same part-of-speech category and may therefore undergo similar inflectional transformations. Similarly, the LCS strategy retrieved examples that share surface overlap with the query and may occasionally reflect similar inflectional patterns. However, despite retrieving examples that appear relevant, the model frequently fails to produce the correct inflected form.

The fundamental difference lies in what character-level similarity captures across the two tasks: in derivation, surface form overlap more reliably signals shared morphological structure, mak-

ing LCS an effective proxy. In reinflection, however, generating the correct output requires sensitivity to paradigmatic features that are not always recoverable from surface similarity alone. Furthermore, listing the relevant features provides an exact specification, whereas a string overlap only provides an approximation. As a result, LCS examples may not be sufficient to provide the model with the inflectional cues necessary to support accurate form generation.

Combined Strategies (CM) Given that i-CES and SES capture complementary signals for exemplar selection (model confidence vs. query similarity), we investigate whether combining them yields additional improvements. We evaluate a combined approach using HC and LCS. HC supports broad morphological patterns while LCS provides input-specific guidance, making their combination potentially complementary. For each query, candidate examples were retrieved independently: LCS candidates are ranked by their lexical similarity to the query, while HC candidates are ranked by their probability. Unlike the i-CES method, which selects high-confidence exemplars by retrieving instances where the model predicts incorrectly, the combined method selects instances with high posterior probability regardless of prediction correctness. We retain the top 50 examples from each ranked list and construct the final few-shot demonstrations by iteratively selecting examples found in both candidate sets.

The results in Figure 5 and 6 show improvement over random and HC across derivational and inflectional tasks for both models. However, while CM offers a clear improvement over these baselines, it does not surpass the best-performing methods, i.e., LCS for derivation and LF for inflection.

Inflectional segmentation To determine whether the challenges in the inflection task are inherent to *generation*, where the model must produce a word given lemma and features, we introduce an auxiliary inflectional segmentation experiment to assess the models’ ability for *analysis*. In this experiment, we ask the model to perform inflectional segmentation. Since inflectional segmentation datasets are only available for high-resource languages, we conduct this experiment on those languages only. The results are presented in Figure 3.

Unlike the inflectional task discussed in Section 5, the segmentation task shows a clear benefit from both SES and i-CES selection strategies. Con-

sistent with the results observed in the derivational task, the LCS and SM methods achieve the best performance and significantly outperform HC and LC approaches and the random baseline. However, the performance of i-CES remains largely comparable to random selection. Overall, the segmentation task is challenging in the zero-shot setting. Nevertheless, performance improves substantially when carefully selected examples are incorporated using our proposed methods. In contrast, performance on the inflection generation task is relatively strong under zero-shot, and more than 10% improvement is observed with the selected demonstrations.

Many Shots: Using a large number of demonstrations may reduce the need for task-specific fine-tuning. Zhang et al. (2025) report comparable performance when models are prompted and fine-tuned with the same number of examples. Similarly, Lampinen et al. (2025) has shown that ICL can generalize effectively across tasks compared to fine-tuning. Following these findings, we conduct a many-shot experiment on LLaMA for French, Russian, Galician, Irish, and Mongolian. As shown in Table 3, the many-shot setting does not surpass the best selection approaches, LCS, in the derivation task and LF in the inflectional task. Our result aligns with Mirbostani et al. (2023), where carefully selected examples can reach peak performance with fewer examples compared to a randomly selected many-shot method.

Model calibration We investigate whether exposing the model to high-confidence (HC) and low-confidence (LC) instances can improve model calibration, i.e., the degree to which the model’s confidence scores align with its actual predictions. To evaluate this, we measure the Expected Calibration Error (ECE) for selected languages; see Appendix H.

While accuracy shows a gradual improvement as the number of shots increases, ECE does not decrease correspondingly, meaning the model’s confidence estimates fail to become more reliable despite the additional context provided by HC and LC examples. This difference between accuracy and calibration scores implies that selecting instances based on confidence alone does not effectively guide the model toward better-calibrated predictions. Consequently, i-CES proves insufficient as a standalone strategy for mitigating miscalibration.

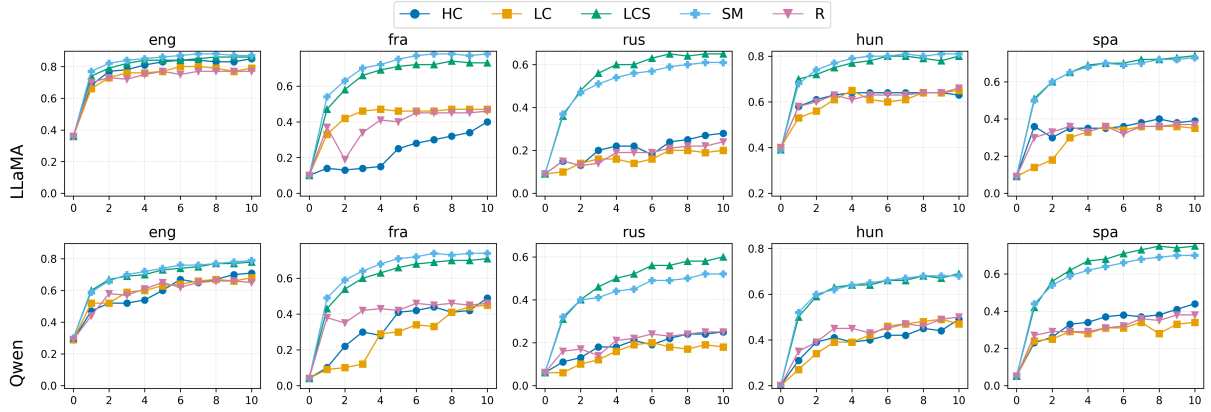


Figure 3: **Inflectional segmentation task: Accuracy** of different retrieval strategies under varying numbers of shots. HC: High-confidence; LC: Low-confidence; LCS: Longest Common Substring; SM: Semantic; R: Random.

Task		Lang.	25-shot	50-shot	250-shot	500-shot	1000-shot	10-shot (LCS)
Derivation	fra		0.26	0.29	0.32	0.27	0.16	0.41
	rus		0.59	0.58	0.58	0.52	0.40	0.73
	gal		0.42	0.42	0.46	0.47	–	0.52
	gle		0.29	0.28	0.27	0.26	0.28	0.38
	mon		0.39	0.39	0.27	0.17	–	0.53
								10-shot (LF)
Inflection	fra		0.84	0.84	0.80	0.18	0.75	0.88
	rus		0.61	0.65	0.63	0.55	0.37	0.84
	gal		0.44	0.57	0.63	0.61	0.59	0.76
	gle		0.36	0.38	0.38	0.40	0.32	0.44
	mon		0.47	0.55	0.61	0.62	0.59	0.62

Table 3: **Derivational and Inflection Task: Accuracy** of the LLaMA model on Many-Shot setting - **Random baseline**.

7 Conclusion

In this work, we study morphological competence across both high-resource and low-resource languages, covering derivational and inflectional tasks using the LLaMA 3.3 70B and Qwen2.5 32B Instruct models. We compare iterative-confidence-based exemplar selection (i-CES) and similarity-based exemplar selection (SES) methods. The findings show that the similarity-based methods outperform confidence-based methods. Moreover, lexical similarity (LCS) is more effective than semantic similarity (SM) for extracting linguistic knowledge, emphasizing the importance of surface-form overlap between the demonstration examples and the query. We show that, for generating inflected forms, providing few-shot demonstrations with explicit morphological features is significantly more beneficial than relying on surface-level overlap. Unlike

i-CES, SES selects query-specific exemplars, yielding more relevant demonstrations and better results. i-CES instead serves as a proxy for generalization, and its weaker performance suggests limited generalization, particularly in low-resource settings.

8 Limitations

We evaluate our methods on two different architectures and sizes: LLaMA 3.3 70B and Qwen2.5 32B models. A broader cross-model analysis would be necessary to confirm the universal robustness of the selection strategies. While we prioritize low-resource languages, our analysis is constrained to five languages due to the limited availability of morphological annotations, particularly for derivational morphology. Expanding this scope to a more typologically diverse set of languages would facilitate a more comprehensive evaluation of our methods. These constraints reflect challenges in computational morphology: the scarcity of gold-standard linguistic resources for underrepresented languages.

Furthermore, our study is restricted to a limited number of potential training examples to simulate a low-resource scenario. Experiments with different training pool sizes could provide further insights.

Finally, we did not explore the full space of combining the different strategies, but only tested one combination. While our experiments show that similarity-based exemplar selection (SES) provides relevant demonstration examples resulting in improved performance, further combinations between selection strategies might provide further improvements. We leave this exploration, potentially in combination with a varying training pool size, as an investigation for future work.

9 Acknowledgements

This work was co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b279bb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

References

- Dang Anh, Limor Raviv, and Lukas Galke. 2024. [Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Baldridge and Alexis Palmer. 2009. [How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation](#). In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. [MorphyNet: a large multilingual database of derivational and inflectional morphology](#). In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. [Prompting language models for linguistic structure](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. [Self-icl: Zero-shot in-context learning with self-generated demonstrations](#). *Preprint*, arXiv:2305.15035.
- Marion Di Marco and Alexander Fraser. 2024. [Subword segmentation in LLMs: Looking at inflection and consistency](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.
- Micha Elsner and David Liu. 2025. Prompt and circumstance”: a word-by-word llm prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the The 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*, pages 1–14.
- Tanja Gaustad and Martin J. Puttkammer. 2022. [Linguistically annotated dataset for four official south african languages with a conjunctive orthography: Isindebele, isixhosa, isizulu, and siswati](#). *Data in Brief*, 41:107994.
- Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. [Can we teach language models to gloss endangered languages?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.
- Leander Gierbach. 2022. [SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation](#). In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 124–130, Seattle, Washington. Association for Computational Linguistics.
- Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. [\(un\)solving morphological inflection: Lemma overlap artificially inflates models’ performance](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

- Coleman Haley. 2024. [The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure?](#) In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 174–178, Mexico City, Mexico. Association for Computational Linguistics.
- Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. [Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.
- Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305.
- Haris Jabbar. 2023. Morphpiece: A linguistic tokenizer for large language models. *arXiv preprint arXiv:2307.07262*.
- Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023. [Morphological inflection: A reality check](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.
- Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. 2025. [On the generalization of language models from in-context learning and finetuning: a controlled study](#). Preprint, arXiv:2505.00661.
- Jaehyeok Lee, Keisuke Sakaguchi, and JinYeong Bak. 2025. [Self-training meets consistency: Improving LLMs’ reasoning with consistency-driven rationale evaluation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10519–10539, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2024. [Self-augmented in-context learning for unsupervised word translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 743–753, Bangkok, Thailand. Association for Computational Linguistics.
- Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. [Active learning principles for in-context learning with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.
- Seyed Morteza Mirbostani, Yasaman Boreshban, Salam Khalifa, SeyedAbolghasem Mirroshandel, and Owen Rambow. 2023. [Deep active learning for morphophonological processing](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 793–803, Toronto, Canada. Association for Computational Linguistics.
- Anssi Moisio, Mathias Creutz, and Mikko Kurimo. 2024. [LLMs’ morphological analyses of complex FST-generated Finnish words](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 242–254, Bangkok, Thailand. Association for Computational Linguistics.
- Saliha Muradoglu, Michael Ginn, Miikka Silfverberg, and Mans Hulden. 2024. [Resisting the lure of the skyline: Grounding practices in active learning for morphological inflection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 47–55, Bangkok, Thailand. Association for Computational Linguistics.
- Saliha Muradoglu and Mans Hulden. 2022. [Eeny, meeny, miny, moe. how to choose data for morphological inflection](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Rita Ramos, Evelyn Asiko Chimoto, Maartje Ter Hove, and Natalie Schluter. 2025. [GrammarMT: Improving machine translation with grammar-informed in-context learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29920–29940, Vienna, Austria. Association for Computational Linguistics.
- Alexey Sorokin and Regina Nasyrova. 2025. [LLMs in alliance with edit-based models: advancing in-context learning for grammatical error correction by specific example selection](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 517–534, Vienna, Austria. Association for Computational Linguistics.
- Jim Su, Justin Ho, George Broadwell, Sarah Moeller, and Bonnie Dorr. 2024. [A comparison of fine-tuning and in-context learning for clause-level morphosyntactic alternation](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 179–187, Mexico City, Mexico. Association for Computational Linguistics.

- Tsedeniya Kinfе Temesgen, Marion Di Marco, and Alexander Fraser. 2025. [Extracting linguistic information from large language models: Syntactic relations and derivational knowledge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27210–27226, Suzhou, China. Association for Computational Linguistics.
- Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. [Holmes: A benchmark to assess the linguistic competence of language models](#). *Transactions of the Association for Computational Linguistics*, 12:1616–1647.
- Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. [Counting the bugs in ChatGPT’s wugs: A multilingual investigation into the morphological capabilities of a large language model](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.
- Marion Weller-Di Marco and Alexander Fraser. 2024. [Analyzing the understanding of morphologically complex words in large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1009–1020, Torino, Italia. ELRA and ICCL.
- Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.
- Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. [LingGym: How far are LLMs from thinking like field linguists?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1340, Suzhou, China. Association for Computational Linguistics.
- Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024. [Multiple sources are better than one: Incorporating external knowledge in low-resource glossing](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4537–4552, Miami, Florida, USA. Association for Computational Linguistics.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. [CrossFit: A few-shot learning challenge for cross-task generalization in NLP](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. [In-context example selection via similarity search improves low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. [Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaoqing Zhang, Ang Lv, Yuhang Liu, Flood Sung, Wei Liu, Jian Luan, Shuo Shang, Xiuying Chen, and Rui Yan. 2025. [More is not always better? enhancing many-shot in-context learning with differentiated and reweighting objectives](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30539–30552, Vienna, Austria. Association for Computational Linguistics.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. [Active example selection for in-context learning](#). *Preprint*, arXiv:2211.04486.
- Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. [NEO-BENCH: Evaluating robustness of large language models with neologisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

Appendices

A Languages

- Amharic (Semitic) is a Semitic language spoken in Ethiopia. It is highly inflected, particularly for verbs, using prefixes, suffixes, and changes to the typical Semitic consonantal root-and-pattern type ¹¹.
- English (Germanic) is a West-Germanic language spoken across the world. Its morphological inflection is only minimal: Nouns mark number (singular, plural), and verbs are inflected for tense and 3rd person singular.
- French (Romance) is a Romance language of the Indo-European family. French is an official language in 26 countries, with speakers worldwide in about 50 countries¹². Nouns and adjectives are inflected for number (singular, plural) and gender (masculine, feminine).

¹¹<https://en.wikipedia.org/wiki/Amharic>

¹²https://en.wikipedia.org/wiki/French_language

Verbal inflection contains tense, aspect, mood, person, and number.

- Galician (Romance) is a West Iberian language (evolved from Latin), spoken in parts of Spain. Nominal inflection contains gender and number; verbs may inflect for number, person, tense, and mood (finite verbs) and gender and number (participles).¹³
- Hungarian (Uralic) is a Uralic language spoken in Hungary and parts of several neighboring countries. Verbs are conjugated for definiteness, tense, mood, person, and number, while nouns take one of 18 case suffixes, most of which correspond to prepositions in English.¹⁴
- Irish (Celtic) is a Celtic language spoken in the Republic of Ireland as an official language. It is an inflecting language. Nouns are inflected in gender, number, and case; verbs are conjugated for person, tense, and mood. Verbs also have a verbal noun, which takes the place of an infinitive, and a verbal adjective that functions similarly to a past participle.¹⁵
- Mongolian (Mongolic) is a Mongolic language and the official language of Mongolia. It is mainly a suffixing language. Mongolian lacks grammatical gender nouns, and its nouns take one of seven to nine cases, depending on the analysis. Several of its derivational verbal suffixes are highly productive.¹⁶
- Russian (Slavic) is a Slavic language and the official language of Russia. It has a highly inflectional morphology, particularly among nominals: nouns are inflected for six main cases across two numbers, and are marked for grammatical gender. Verbs are inflected for three persons and two numbers, and conjugate across two simple tenses¹⁷.
- Spanish (Romance) is a Romance language and the official language of Spain and much of Latin America. Its noun and adjective systems

exhibit two genders and two numbers. Spanish verbs are complex and highly inflected. Spanish inflects verbs to convey mood, tense, voice, and aspect, as well as to agree with person and number, resulting in roughly 50 conjugated forms per verb¹⁸.

- Zulu (Bantu) is a Southern Bantu language widely spoken in South Africa. Its nouns fall into morphological classes, or genders (16 in total), each distinguished by separate prefixes for singular and plural forms. The verbal system combines tense and aspect. Prefixation can also apply to the root itself to derive further words.¹⁹

Derivation across these languages happens mainly through affixation. English, French, Galician, Irish, Spanish, and Russian mainly use suffixation. For agglutinative languages, Hungarian, Mongolian, and Zulu, word formation happens by stacking derivational suffixes. Amharic, however, uses root-and-pattern morphology, fitting a consonantal root into different vowel templates.

B Dataset

Derivational Analysis: For high-resource languages, we use the SIGMORPHON²⁰ dataset, which provides derivational segmentation for each word. For low-resource languages, we used UNIMORPH²¹ for Galician and Irish, for Mongolian sourced from MORPHYNET²², Amharic from Temesgen et al. (2025), and Zulu from Gaustad and Puttkammer (2022).

SIGMORPHON includes 9 languages. For each language, they provide word categories, namely inflection-only, Derivation-only, Inflection and Compound, Root words, Derivation and compound, Inflection and Derivation, Compound-only, and Inflection, Derivation, Compound. Hence, we are primarily interested in derivational and inflectional categories; we select those labeled accordingly.

In addition, SIGMORPHON provides segmentation without part-of-speech (POS) classes; therefore, we obtain POS class information from UniMorph. We then preprocess the dataset by retaining unique surface forms and removing ambiguous

¹³Information from <https://universaldependencies.org/gl/>

¹⁴https://en.wikipedia.org/wiki/Hungarian_grammar

¹⁵<http://www.glottopedia.org/index.php/Irish>

¹⁶https://en.wikipedia.org/wiki/Mongolian_language

¹⁷https://en.wikipedia.org/wiki/Russian_grammar

¹⁸https://en.wikipedia.org/wiki/Spanish_language

¹⁹https://en.wikipedia.org/wiki/Zulu_language

²⁰<https://github.com/sigmorphon/2022SegmentationST>

²¹<https://unimorph.github.io/>

²²<https://github.com/kbatsuren/MorphyNet>

words, i.e., words that can belong to more than one POS class. We frame this task by prompting the model to generate a derivational segmentation for a given input word (See Section D for prompt instructions and examples).

Inflectional Analysis: For both high- and low-resource languages, we use UNIMORPH to obtain morphological segmentation for inflected forms, and MORPHYNET for Mongolian. UniMorph does not uniformly provide both derivational and inflectional annotations for all languages; we restrict our experiments to those languages for which data are available for both tasks. We also ensure that all surface word forms are unique and replace morphological tags with their full explanations to make their meanings clearer for the models. We prompt the model to provide the inflected form of a given input word, given a lemma and morphological features (See Section D for prompt instructions and examples).

Official test splits are preserved; otherwise, we create a split with 1K train and test samples. Table 4 shows the number of samples per language.

Lang	Derivation		Inflection	
	Train	Test	Train	Test
amh	854	1,000	1,000	1,000
gal	661	1,000	1,000	1,000
gle	1,000	1,000	1,000	1,000
mon	548	1,000	1,000	1,000
zul	1000	974	1,000	1,000
eng	1,000	1,000	1,000	1,000
fra	1,000	1,000	1,000	1,000
hun	1,000	1,000	1,000	1,000
rus	1,000	560	1,000	1,000
spa	1,000	1,000	1,000	1,000

Table 4: **Dataset size** for derivation and inflection tasks across languages.

C Levenshtein Vs Longest match Substring

To select an appropriate similarity measure, we compared the performance of the longest common substring (LCS) and Levenshtein distance (LV) in the few-shot, $K=[0, 10]$, setting. Both metrics yielded comparable results, with no consistent advantage for either method. The results in Table 5 shows that the differences between the two approaches are negligible. Given this comparable performance, we adopt LCS for our experiments.

D Prompt Instruction

Table 6 presents the prompt template used for all experiments, and models receive the question directly

	Task	Metric	eng	spa	fra	hun	rus	gal	gle	mon	avg
LLaMA	Der.	LV	0.53	0.45	0.38	0.54	0.68	0.45	0.34	0.41	0.47
		LCS	0.53	0.46	0.37	0.55	0.68	0.44	0.34	0.37	0.47
	Inf.	LV	0.94	0.52	0.79	0.61	0.65	0.32	0.30	0.34	0.56
		LCS	0.94	0.52	0.79	0.61	0.65	0.32	0.30	0.35	0.56
Qwen	Der.	LV	0.48	0.40	0.31	0.47	0.64	0.45	0.35	0.38	0.44
		LCS	0.47	0.40	0.32	0.47	0.63	0.43	0.33	0.35	0.43
	Inf.	LV	0.68	0.32	0.32	0.29	0.48	0.22	0.16	0.29	0.34
		LCS	0.67	0.35	0.35	0.29	0.47	0.23	0.12	0.25	0.34

Table 5: **Longest Common Substring (LCS)** and **Levenshtein distance (LV)**: Average Accuracy over K-shots

after the instruction.

Input examples for each evaluated task are presented below.

Derivation:

Query: incompatibility

Gold answer: in|compatible|ity

Inflection:

Query: Lemma: stock;

Feature(s): Participle, Past

Gold answer: stocked

We conduct additional experiments on high-resource languages, in which the models are prompted to provide inflectional segmentation shown in the example below.

Inflectional Segmentation:

Query: resolved

Gold answer: resolve|ed

E Surface vs. Canonical Segmentation

We analyze derivational words according to two levels of morphological complexity: *surface-level segmentation* and *canonical segmentation*.

Surface-level segmentation identifies the lemma by directly removing affixes from a surface word, without requiring any changes to the stem. For example, *careless* as *care* + *-less*. These cases are relatively straightforward and can often be resolved through simple pattern matching.

In contrast, **canonical segmentation** requires recovering the underlying base form through orthographic transformations. For example, *decision* is canonically segmented as *decide* + *-ion*, where the stem undergoes a spelling change, and *ability* is segmented as *able* + *-ity*, which involves the deletion of the final *e*. Similarly, *studies* can be segmented into the base form *study* through a $y \rightarrow i$

Task	Generation
Derivation	Provide the derivational analysis for the given word based on the morphological rules of LANGUAGE. Segment the given word into its canonical form and its valid affix. The morphemes should be separated by the 'l' symbol. Use the following format exactly: <i>Answer: [answer separated by 'l']</i> Provide ONLY the answer, with no explanation, and you should follow the format strictly.
Inflection	Generate the inflected form of the given lemma based on the morphological rules of LANGUAGE. For the given lemma and inflectional features, produce the correct inflected form of the word. Use the following format exactly: <i>Answer: [inflected_form]</i> . Provide ONLY the answer, with no explanation, and follow the format strictly.
Inflection segmentation	Perform inflectional segmentation of the given word based on the morphological rules of LANGUAGE. Identify and segment only the inflectional morphemes from the given word. The morphemes should be separated by the symbol. Use the following format exactly: <i>Answer: [answer separated by]</i> . Provide ONLY the answer, with no explanation, and follow the format strictly.

Table 6: Prompt instruction for **derivation** and **inflection** task.

alternation. These transformations make canonical segmentation challenging, as the model must infer underlying morphological alternations.

Table 7 reports the error rates for both morphological complexity levels. We compare LCS and HC methods on both LLaMA and Qwen models.

F Morpheme Fertility

Morpheme fertility, as defined by Jabbar (2023), is the average number of subwords a tokenizer splits a word into. In this work, we adopt a conceptually aligned definition that measures word complexity through segmentation length, where morpheme fertility is defined as the number of morphological units within a word, i.e., the stem and all affixes. To better understand model errors, we examine instances that are incorrectly segmented by the model. This allows us to assess whether the model captures the expected number of morphemes in a word even when it fails to correctly identify the lemma or affixes. Therefore, we measure the fertility rate, i.e., how well the predicted segmentation matches the underlying morphological composition of the word, to capture whether the number of predicted segments corresponds to the correct number of decompositions. The evaluation is conducted with respect to word complexity, as described in Section E on surface vs. canonical segmentation, and is reported under both LCS- and HC-based settings on the 10-shot results. See Figure 4

G Analysis of Combining i-CES and SES Methods

We include additional experiments pertaining to the combined method (CM), derived from intersecting i-CES and SES selections. Specifically, we examine a combination of the High-Confidence (HC) and Longest Common Substring (LCS) approaches, selecting demonstrations that appear in both sets.

HC supports broad morphological patterns while LCS provides input-specific guidance, making their combination potentially complementary. Unlike the i-CES method, which selects high-confidence exemplars by retrieving instances where the model predicts incorrectly, the combined method selects instances with high posterior probability regardless of prediction correctness. Results are shown in Figure 5 and Figure 6

H Scores for SES and i-CES for all languages and models

Tables 12, 13, 14, and 15 show the accuracy for the derivation task. Table 8, 9, 10 and 11 reports scores for inflectional task each language. We report the Expected Calibration Error (ECE) scores of the i-CES methods in Table 18 and 19 for selected languages.

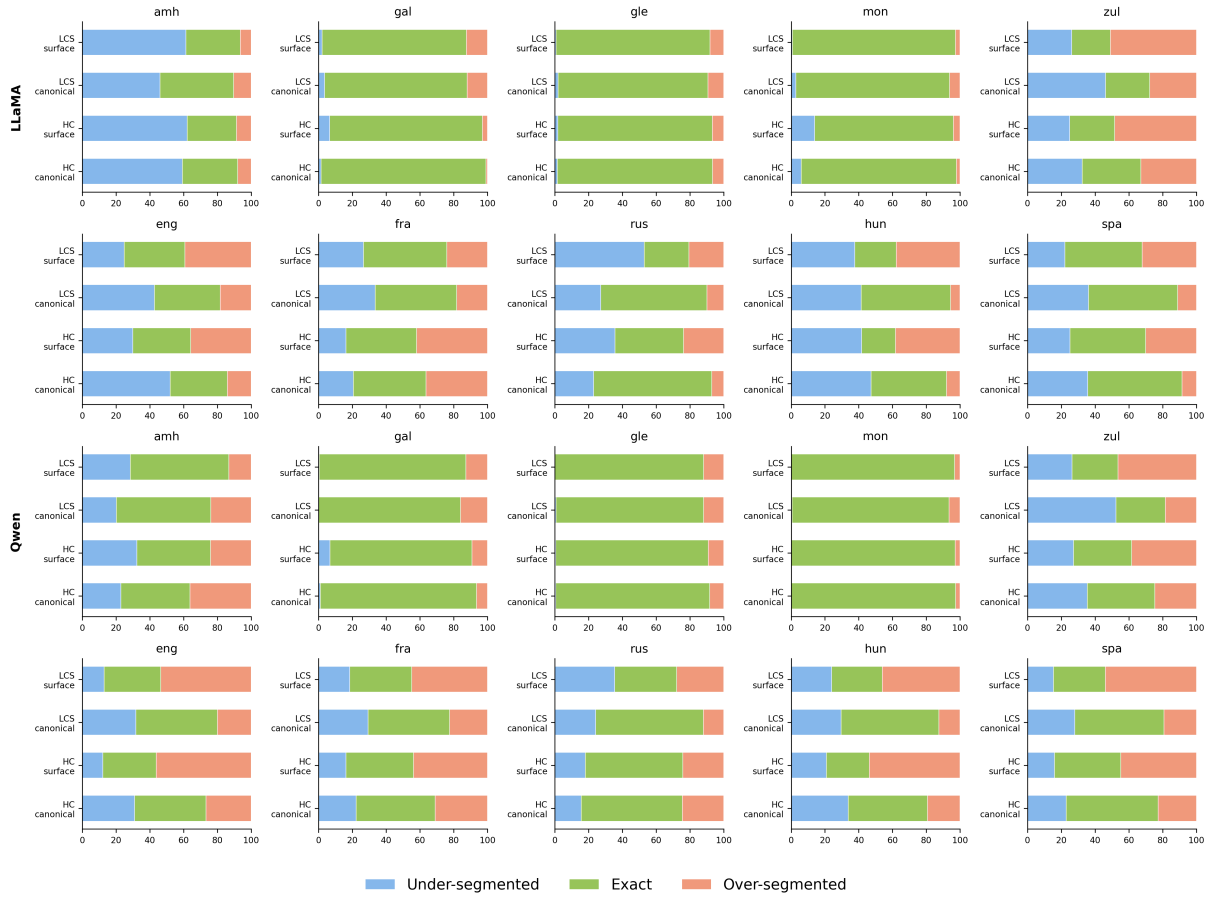


Figure 4: **Morpheme Fertility**: Accuracy of correctly segmented (Exact), over-segmented, and under-segmented morphemes among incorrectly predicted words in the 10-shot setting. Results are reported by word complexity level (surface vs. canonical); HC: high-confidence; LCS: Longest Common Substring

Lang	LLaMA				Qwen				Number of Instances	
	LCS surface	LCS canonical	HC surface	HC canonical	LCS surface	LCS canonical	HC surface	HC canonical	surface	canonical
amh	56.60	80.00	91.70	98.30	62.80	84.20	72.90	97.90	576	424
eng	35.00	74.10	37.00	79.00	42.00	72.40	47.70	84.80	757	243
fra	53.50	64.70	62.10	86.80	58.60	67.10	73.00	77.30	507	493
gal	32.70	69.10	53.70	76.50	40.10	62.20	54.40	71.40	566	434
gle	50.30	79.70	65.70	88.50	51.00	77.40	66.40	86.90	610	390
mon	44.90	47.70	62.20	67.30	48.00	52.30	65.20	76.00	325	675
rus	18.50	42.00	32.20	58.50	21.50	43.00	39.20	63.20	367	193
hun	31.70	81.90	46.40	93.50	38.70	81.90	57.10	96.30	785	215
spa	38.90	54.80	45.30	68.00	45.30	56.90	50.30	78.90	360	640
zul	57.30	42.60	82.50	67.10	56.80	49.40	89.30	76.50	634	340

Table 7: **Error rate** on **Surface** and **Canonical** words on LCS and HC methods (*lower better*).

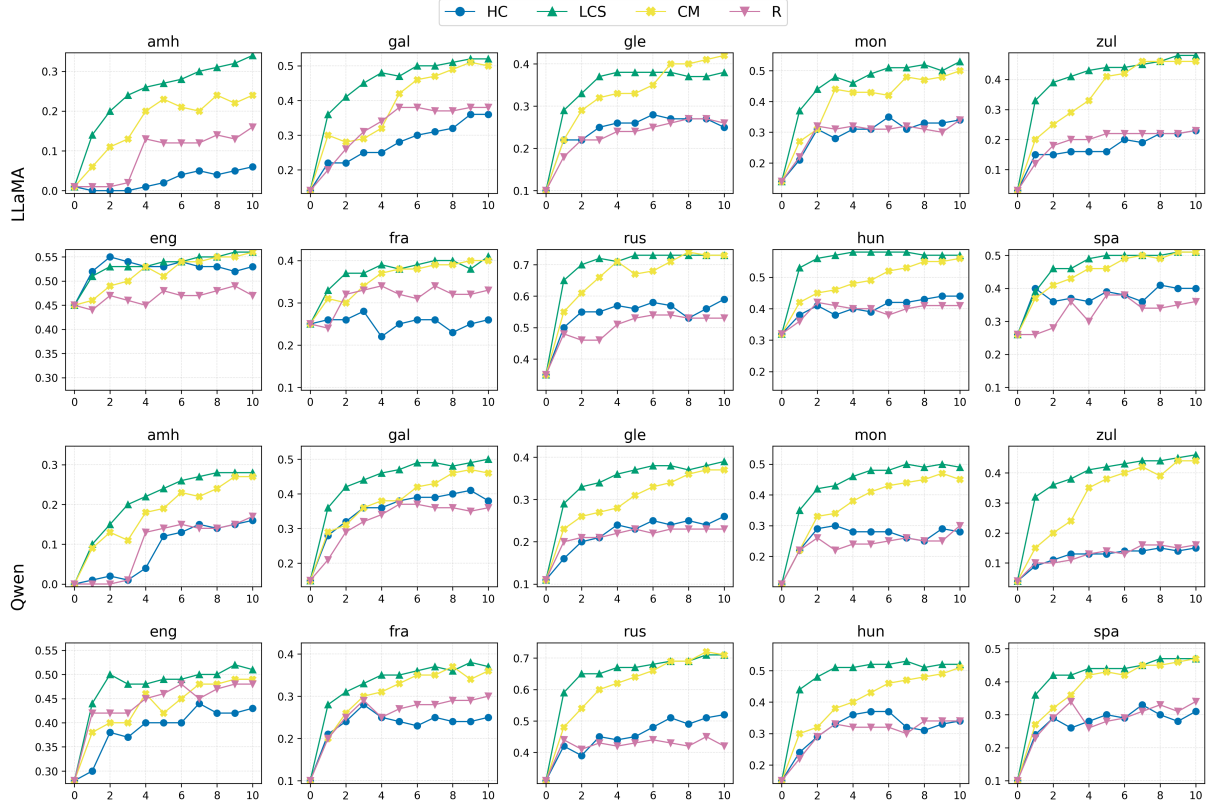


Figure 5: **Derivational task**: Accuracy of different retrieval strategies under varying numbers of shots. CM: Combined; HC: high-confidence; LCS: Longest Common Substring; R: Random.

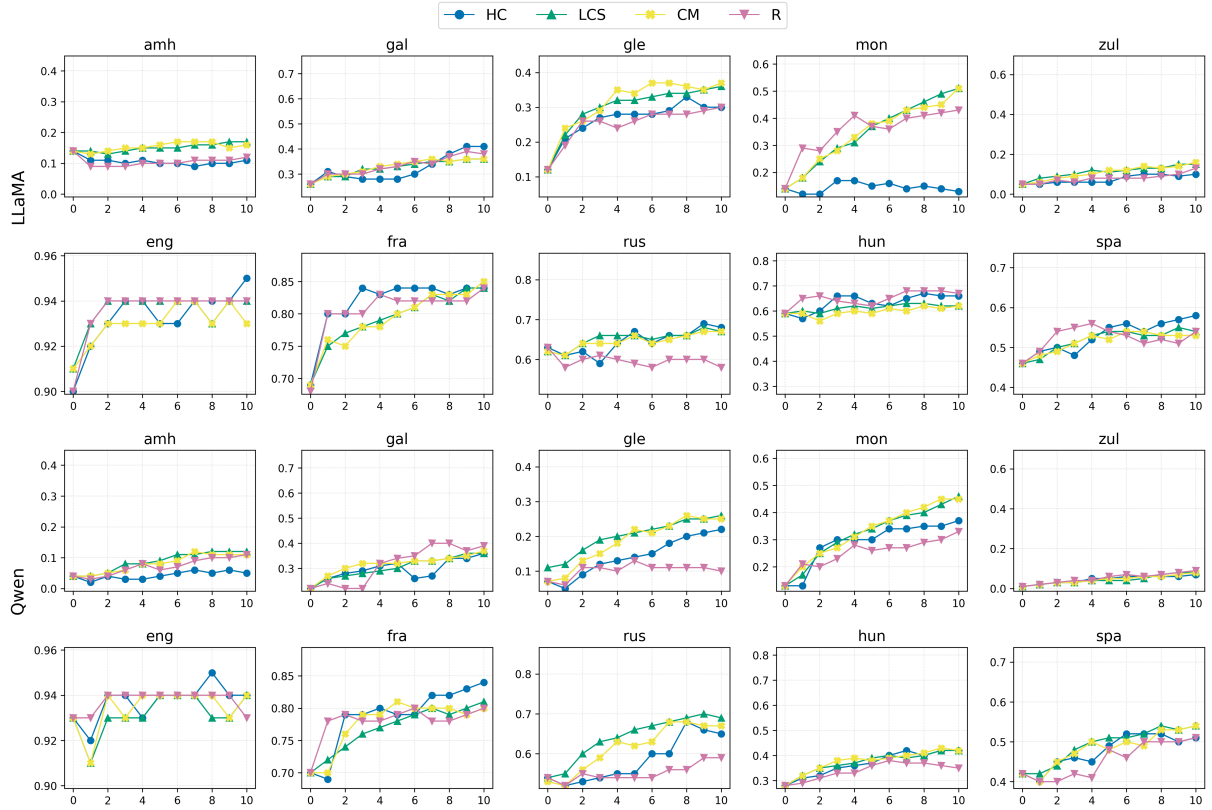


Figure 6: **Inflection task**: Accuracy of different retrieval strategies under varying numbers of shots. CM: Combined; HC: high-confidence; LCS: Longest Common Substring; R: Random; LF: lemma and feature match

Table 8: **Inflection task:** Accuracy results for LLaMA Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.90	0.92	0.92	0.92	0.93	0.94	0.94	0.95	0.95	0.94	0.94
	HC	0.90	0.92	0.93	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.95
	LV	0.91	0.94	0.94	0.94	0.95	0.95	0.94	0.95	0.96	0.95	0.95
	LCS	0.91	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94
	SM	0.91	0.92	0.94	0.94	0.94	0.95	0.95	0.94	0.94	0.95	0.95
	R	0.90	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	CM	0.91	0.92	0.93	0.93	0.93	0.93	0.94	0.94	0.93	0.94	0.93
	LF	0.91	0.94	0.94	0.94	0.95	0.94	0.95	0.95	0.95	0.94	0.95
fra	LC	0.69	0.70	0.70	0.80	0.79	0.80	0.81	0.82	0.80	0.80	0.82
	HC	0.69	0.80	0.80	0.84	0.83	0.84	0.84	0.84	0.83	0.84	0.84
	LV	0.69	0.76	0.77	0.79	0.79	0.80	0.81	0.82	0.82	0.83	0.82
	LCS	0.69	0.75	0.77	0.78	0.79	0.80	0.81	0.83	0.82	0.84	0.84
	SM	0.69	0.76	0.77	0.78	0.79	0.80	0.81	0.82	0.82	0.82	0.83
	R	0.68	0.80	0.80	0.80	0.83	0.82	0.82	0.82	0.82	0.82	0.84
	CM	0.69	0.76	0.75	0.78	0.78	0.80	0.81	0.83	0.83	0.83	0.85
	LF	0.69	0.85	0.87	0.88	0.88	0.88	0.89	0.89	0.88	0.88	0.88
hun	LC	0.59	0.58	0.56	0.58	0.59	0.60	0.59	0.60	0.60	0.61	0.61
	HC	0.59	0.57	0.60	0.66	0.66	0.63	0.62	0.65	0.67	0.66	0.66
	LV	0.59	0.60	0.60	0.62	0.61	0.60	0.61	0.61	0.62	0.62	0.62
	LCS	0.59	0.60	0.59	0.61	0.62	0.61	0.62	0.63	0.63	0.62	0.62
	SM	0.59	0.61	0.64	0.64	0.64	0.64	0.64	0.65	0.66	0.67	0.66
	R	0.59	0.65	0.66	0.64	0.63	0.62	0.65	0.68	0.68	0.68	0.67
	CM	0.59	0.59	0.56	0.59	0.60	0.59	0.61	0.60	0.62	0.61	0.62
	LF	0.59	0.79	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.82
rus	LC	0.63	0.58	0.57	0.58	0.55	0.55	0.55	0.56	0.56	0.56	0.57
	HC	0.63	0.61	0.62	0.59	0.64	0.67	0.64	0.66	0.66	0.69	0.68
	LV	0.62	0.61	0.65	0.65	0.64	0.66	0.65	0.66	0.66	0.67	0.67
	LCS	0.62	0.61	0.64	0.66	0.66	0.66	0.65	0.66	0.66	0.68	0.67
	SM	0.62	0.61	0.64	0.66	0.67	0.67	0.69	0.69	0.69	0.69	0.70
	R	0.63	0.58	0.60	0.61	0.60	0.59	0.58	0.60	0.60	0.60	0.58
	CM	0.62	0.61	0.64	0.64	0.64	0.66	0.64	0.65	0.66	0.67	0.67
	LF	0.62	0.80	0.82	0.83	0.82	0.83	0.85	0.84	0.84	0.84	0.84
spa	LC	0.46	0.51	0.49	0.50	0.54	0.52	0.52	0.52	0.53	0.51	0.54
	HC	0.46	0.49	0.50	0.48	0.52	0.55	0.56	0.54	0.56	0.57	0.58
	LV	0.46	0.50	0.51	0.52	0.52	0.52	0.52	0.54	0.54	0.55	0.55
	LCS	0.46	0.47	0.50	0.51	0.53	0.54	0.54	0.53	0.53	0.55	0.54
	SM	0.46	0.48	0.51	0.50	0.52	0.52	0.52	0.54	0.55	0.55	0.55
	R	0.46	0.49	0.54	0.55	0.56	0.54	0.53	0.51	0.52	0.51	0.54
	CM	0.46	0.48	0.49	0.51	0.53	0.52	0.54	0.54	0.53	0.53	0.53
	LF	0.46	0.63	0.67	0.68	0.70	0.70	0.70	0.71	0.73	0.72	0.73

Table 9: **Inflection task:** Accuracy results for LLaMA Model (*continued from the pervious table*)

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
amh	LC	0.14	0.13	0.12	0.10	0.11	0.12	0.11	0.12	0.11	0.13	0.11
	HC	0.14	0.11	0.11	0.10	0.11	0.10	0.10	0.09	0.10	0.10	0.11
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.14	0.14	0.13	0.14	0.15	0.15	0.15	0.16	0.16	0.17	0.17
	SM	0.14	0.14	0.13	0.14	0.14	0.14	0.15	0.14	0.14	0.15	0.14
	R	0.14	0.09	0.09	0.09	0.10	0.10	0.10	0.11	0.11	0.11	0.12
	CM	0.14	0.13	0.14	0.15	0.15	0.16	0.17	0.17	0.17	0.15	0.16
	LF	0.14	0.25	0.33	0.35	0.37	0.40	0.41	0.42	0.43	0.44	0.43
gal	LC	0.26	0.28	0.30	0.31	0.33	0.35	0.35	0.33	0.36	0.35	0.36
	HC	0.26	0.31	0.29	0.28	0.28	0.28	0.30	0.34	0.38	0.41	0.41
	LV	0.26	0.28	0.28	0.30	0.31	0.32	0.34	0.34	0.35	0.35	0.36
	LCS	0.26	0.29	0.29	0.32	0.32	0.33	0.34	0.35	0.35	0.36	0.36
	SM	0.26	0.29	0.30	0.31	0.31	0.33	0.32	0.33	0.34	0.37	0.36
	R	0.26	0.30	0.30	0.30	0.32	0.33	0.35	0.34	0.37	0.39	0.38
	CM	0.26	0.29	0.30	0.31	0.33	0.34	0.35	0.36	0.35	0.36	0.36
	LF	0.26	0.54	0.65	0.69	0.70	0.72	0.72	0.74	0.74	0.75	0.76
gle	LC	0.12	0.19	0.25	0.26	0.27	0.26	0.25	0.25	0.26	0.26	0.26
	HC	0.12	0.21	0.24	0.27	0.28	0.28	0.28	0.29	0.33	0.30	0.30
	LV	0.12	0.20	0.27	0.30	0.31	0.32	0.34	0.35	0.36	0.36	0.37
	LCS	0.12	0.22	0.28	0.30	0.32	0.32	0.33	0.34	0.34	0.35	0.36
	SM	0.12	0.21	0.27	0.29	0.31	0.32	0.34	0.35	0.37	0.38	0.37
	R	0.12	0.19	0.26	0.26	0.24	0.26	0.28	0.28	0.28	0.29	0.30
	CM	0.12	0.24	0.26	0.29	0.35	0.34	0.37	0.37	0.36	0.35	0.37
	LF	0.12	0.29	0.38	0.39	0.41	0.42	0.41	0.42	0.43	0.43	0.44
mon	LC	0.14	0.17	0.36	0.40	0.44	0.44	0.44	0.41	0.48	0.50	0.44
	HC	0.14	0.12	0.12	0.17	0.17	0.15	0.16	0.14	0.15	0.14	0.13
	LV	0.14	0.17	0.23	0.29	0.33	0.37	0.40	0.42	0.46	0.49	0.50
	LCS	0.14	0.18	0.24	0.29	0.31	0.37	0.40	0.43	0.46	0.49	0.51
	SM	0.14	0.16	0.21	0.25	0.29	0.31	0.33	0.37	0.37	0.39	0.40
	R	0.14	0.29	0.28	0.35	0.41	0.37	0.36	0.40	0.41	0.42	0.43
	CM	0.14	0.18	0.25	0.28	0.33	0.38	0.39	0.43	0.44	0.45	0.51
	LF	0.14	0.35	0.44	0.50	0.54	0.55	0.56	0.61	0.58	0.60	0.62
zul	LC	0.05	0.04	0.06	0.06	0.07	0.08	0.10	0.09	0.10	0.11	0.11
	HC	0.05	0.05	0.06	0.06	0.06	0.06	0.09	0.10	0.10	0.09	0.10
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.05	0.08	0.09	0.10	0.12	0.11	0.12	0.13	0.13	0.15	0.15
	SM	-	-	-	-	-	-	-	-	-	-	-
	R	0.05	0.05	0.07	0.06	0.08	0.08	0.08	0.08	0.09	0.10	0.13
	CM	0.05	0.06	0.08	0.09	0.10	0.12	0.12	0.14	0.13	0.14	0.16
	LF	0.05	0.42	0.55	0.62	0.62	0.64	0.66	0.66	0.68	0.68	0.67

Table 10: **Inflection task:** Accuracy results for Qwen Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.93	0.92	0.93	0.94	0.94	0.93	0.93	0.94	0.94	0.93	0.94
	HC	0.93	0.92	0.94	0.94	0.93	0.94	0.94	0.94	0.95	0.94	0.94
	LV	0.93	0.93	0.93	0.95	0.93	0.94	0.94	0.94	0.94	0.94	0.94
	LCS	0.93	0.91	0.93	0.93	0.93	0.94	0.94	0.94	0.93	0.93	0.94
	SM	0.93	0.92	0.94	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.95
	R	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.93
	CM	0.93	0.91	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.93	0.94
	LF	0.93	0.93	0.95	0.95	0.94	0.94	0.94	0.94	0.95	0.95	0.94
fra	LC	0.70	0.74	0.73	0.76	0.75	0.76	0.76	0.76	0.83	0.82	0.82
	HC	0.70	0.69	0.79	0.79	0.80	0.79	0.79	0.82	0.82	0.83	0.84
	LV	0.70	0.72	0.73	0.76	0.76	0.76	0.79	0.80	0.80	0.80	0.81
	LCS	0.70	0.72	0.74	0.76	0.77	0.78	0.79	0.80	0.79	0.80	0.81
	SM	0.70	0.72	0.76	0.76	0.76	0.77	0.78	0.78	0.79	0.80	0.81
	R	0.70	0.78	0.79	0.78	0.78	0.79	0.80	0.78	0.78	0.79	0.80
	CM	0.70	0.70	0.76	0.79	0.79	0.81	0.80	0.80	0.80	0.79	0.80
	LF	0.70	0.82	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
hun	LC	0.28	0.31	0.33	0.40	0.40	0.37	0.39	0.39	0.41	0.41	0.46
	HC	0.28	0.31	0.32	0.35	0.36	0.37	0.40	0.42	0.40	0.42	0.42
	LV	0.28	0.32	0.35	0.38	0.38	0.39	0.40	0.42	0.42	0.42	0.44
	LCS	0.28	0.32	0.35	0.36	0.37	0.39	0.40	0.39	0.40	0.42	0.42
	SM	0.28	0.32	0.37	0.40	0.40	0.41	0.41	0.43	0.43	0.44	0.45
	R	0.28	0.29	0.31	0.33	0.33	0.36	0.38	0.37	0.37	0.36	0.35
	CM	0.28	0.32	0.35	0.38	0.39	0.38	0.39	0.40	0.41	0.43	0.42
	LF	0.28	0.52	0.57	0.59	0.59	0.61	0.62	0.63	0.63	0.63	0.63
rus	LC	0.54	0.59	0.67	0.65	0.68	0.67	0.66	0.64	0.65	0.66	0.67
	HC	0.54	0.52	0.53	0.54	0.55	0.55	0.60	0.60	0.68	0.66	0.65
	LV	0.54	0.55	0.60	0.64	0.65	0.67	0.68	0.70	0.70	0.71	0.71
	LCS	0.54	0.55	0.60	0.63	0.64	0.66	0.67	0.68	0.69	0.70	0.69
	SM	0.53	0.55	0.62	0.64	0.65	0.68	0.67	0.70	0.70	0.71	0.70
	R	0.54	0.52	0.55	0.54	0.54	0.54	0.54	0.56	0.56	0.59	0.59
	CM	0.53	0.52	0.56	0.59	0.63	0.62	0.63	0.68	0.68	0.67	0.67
	LF	0.53	0.79	0.82	0.83	0.85	0.86	0.86	0.86	0.86	0.86	0.86
spa	LC	0.42	0.39	0.39	0.46	0.45	0.46	0.46	0.46	0.46	0.46	0.46
	HC	0.42	0.40	0.45	0.46	0.45	0.49	0.52	0.52	0.52	0.50	0.51
	LV	0.42	0.41	0.45	0.45	0.47	0.48	0.48	0.50	0.50	0.51	0.52
	LCS	0.42	0.42	0.44	0.48	0.50	0.51	0.51	0.52	0.54	0.53	0.54
	SM	0.42	0.40	0.44	0.46	0.47	0.47	0.48	0.49	0.49	0.50	0.49
	R	0.42	0.40	0.40	0.42	0.41	0.48	0.46	0.50	0.50	0.50	0.51
	CM	0.42	0.40	0.45	0.47	0.50	0.48	0.50	0.49	0.53	0.53	0.54
	LF	0.42	0.58	0.62	0.66	0.66	0.67	0.69	0.69	0.70	0.70	0.72

Table 11: **Inflection task:** Accuracy results for Qwen Model (*continued from the pervious table*)

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
amh	LC	0.04	0.10	0.08	0.08	0.09	0.10	0.10	0.09	0.08	0.08	0.09
	HC	0.04	0.02	0.04	0.03	0.03	0.04	0.05	0.06	0.05	0.06	0.05
	LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LCS	0.04	0.04	0.05	0.08	0.08	0.09	0.11	0.11	0.12	0.12	0.12
	SM	0.04	0.04	0.06	0.07	0.09	0.09	0.09	0.09	0.10	0.11	0.12
	R	0.04	0.03	0.04	0.06	0.08	0.06	0.07	0.09	0.10	0.10	0.11
	CM	0.04	0.04	0.05	0.06	0.08	0.08	0.09	0.12	0.11	0.11	0.11
	LF	0.04	0.19	0.27	0.31	0.34	0.33	0.33	0.34	0.35	0.36	0.37
gal	LC	0.22	0.27	0.29	0.31	0.33	0.36	0.35	0.35	0.36	0.38	0.38
	HC	0.22	0.26	0.28	0.29	0.31	0.32	0.26	0.27	0.34	0.34	0.36
	LV	0.22	0.26	0.26	0.29	0.29	0.30	0.32	0.32	0.32	0.34	0.35
	LCS	0.22	0.26	0.27	0.28	0.29	0.30	0.33	0.33	0.34	0.36	0.36
	SM	0.22	0.26	0.26	0.30	0.30	0.32	0.32	0.34	0.34	0.35	0.36
	R	0.22	0.24	0.22	0.22	0.32	0.34	0.35	0.40	0.40	0.37	0.39
	CM	0.22	0.27	0.30	0.32	0.32	0.32	0.33	0.33	0.34	0.35	0.37
	LF	0.22	0.55	0.63	0.68	0.69	0.71	0.71	0.72	0.72	0.72	0.74
gle	LC	0.07	0.05	0.10	0.10	0.12	0.13	0.16	0.16	0.16	0.16	0.16
	HC	0.07	0.05	0.09	0.12	0.13	0.14	0.15	0.18	0.20	0.21	0.22
	LV	0.07	0.12	0.16	0.19	0.20	0.22	0.23	0.25	0.25	0.26	0.27
	LCS	0.11	0.12	0.16	0.19	0.20	0.21	0.22	0.23	0.25	0.25	0.26
	SM	0.07	0.11	0.16	0.17	0.18	0.20	0.20	0.22	0.23	0.23	0.23
	R	0.07	0.06	0.11	0.11	0.10	0.13	0.11	0.11	0.11	0.11	0.10
	CM	0.07	0.08	0.13	0.15	0.18	0.22	0.21	0.23	0.26	0.25	0.25
	LF	0.07	0.23	0.23	0.25	0.27	0.31	0.28	0.28	0.30	0.28	0.28
	LC	0.13	0.20	0.22	0.25	0.27	0.28	0.29	0.27	0.28	0.29	0.29
mon	HC	0.13	0.13	0.27	0.30	0.30	0.30	0.34	0.34	0.35	0.35	0.37
	LV	0.13	0.19	0.25	0.29	0.34	0.37	0.42	0.43	0.45	0.48	0.49
	LCS	0.13	0.17	0.25	0.29	0.32	0.34	0.37	0.39	0.40	0.43	0.46
	SM	0.13	0.18	0.21	0.23	0.27	0.29	0.32	0.34	0.35	0.35	0.35
	R	0.13	0.21	0.20	0.23	0.28	0.26	0.27	0.27	0.29	0.30	0.33
	CM	0.13	0.20	0.25	0.27	0.31	0.35	0.37	0.40	0.42	0.45	0.45
	LF	0.13	0.31	0.31	0.37	0.42	0.45	0.48	0.47	0.48	0.49	0.52
zul	LC	0.01	0.01	0.02	0.04	0.04	0.04	0.04	0.05	0.07	0.07	0.08
	HC	0.01	0.02	0.03	0.03	0.05	0.05	0.06	0.06	0.06	0.06	0.07
	LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LCS	0.01	0.02	0.03	0.03	0.04	0.04	0.04	0.05	0.07	0.08	0.08
	SM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R	0.01	0.02	0.03	0.04	0.04	0.06	0.07	0.06	0.07	0.08	0.09
	CM	0.01	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.06	0.07	0.08
	LF	0.01	0.43	0.53	0.59	0.60	0.60	0.62	0.62	0.64	0.63	0.63

Table 12: **Derivation Task:** Accuracy results for LLaMA Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.45	0.46	0.49	0.49	0.48	0.49	0.50	0.50	0.50	0.48	0.49
	HC	0.45	0.52	0.55	0.54	0.53	0.53	0.54	0.53	0.53	0.52	0.53
	LV	0.45	0.49	0.52	0.54	0.54	0.56	0.55	0.55	0.56	0.55	0.56
	LCS	0.45	0.51	0.53	0.53	0.53	0.54	0.54	0.55	0.55	0.56	0.56
	SM	0.45	0.49	0.50	0.51	0.52	0.53	0.52	0.53	0.54	0.54	0.55
	R	0.45	0.44	0.47	0.46	0.45	0.48	0.47	0.47	0.48	0.49	0.47
	CM	0.45	0.46	0.49	0.50	0.53	0.51	0.54	0.54	0.55	0.55	0.56
fra	LC	0.25	0.31	0.31	0.30	0.33	0.33	0.32	0.32	0.34	0.34	0.34
	HC	0.25	0.26	0.26	0.28	0.22	0.25	0.26	0.26	0.23	0.25	0.26
	LV	0.25	0.33	0.36	0.38	0.39	0.40	0.42	0.42	0.42	0.40	0.42
	LCS	0.25	0.33	0.37	0.37	0.39	0.38	0.39	0.40	0.40	0.38	0.41
	SM	0.25	0.32	0.35	0.38	0.39	0.40	0.41	0.41	0.42	0.41	0.42
	R	0.25	0.24	0.32	0.33	0.34	0.32	0.31	0.34	0.32	0.32	0.33
	CM	0.25	0.31	0.30	0.34	0.37	0.38	0.38	0.39	0.39	0.40	0.40
hun	LC	0.32	0.36	0.36	0.37	0.38	0.41	0.40	0.41	0.41	0.42	0.41
	HC	0.32	0.38	0.41	0.38	0.40	0.39	0.42	0.42	0.43	0.44	0.44
	LV	0.32	0.50	0.55	0.56	0.55	0.57	0.57	0.57	0.57	0.58	0.57
	LCS	0.32	0.53	0.56	0.57	0.58	0.58	0.58	0.58	0.57	0.57	0.57
	SM	0.32	0.51	0.53	0.55	0.55	0.55	0.55	0.56	0.56	0.55	0.56
	R	0.32	0.36	0.42	0.41	0.40	0.40	0.38	0.40	0.41	0.41	0.41
	CM	0.32	0.42	0.45	0.46	0.48	0.49	0.52	0.53	0.55	0.55	0.56
rus	LC	0.35	0.49	0.51	0.54	0.55	0.55	0.56	0.57	0.57	0.58	0.57
	HC	0.35	0.50	0.55	0.55	0.57	0.56	0.58	0.57	0.53	0.56	0.59
	LV	0.35	0.66	0.68	0.70	0.73	0.73	0.72	0.72	0.72	0.73	0.73
	LCS	0.35	0.65	0.70	0.72	0.71	0.73	0.73	0.73	0.73	0.73	0.73
	SM	0.35	0.64	0.67	0.68	0.71	0.71	0.73	0.72	0.73	0.73	0.74
	R	0.35	0.48	0.46	0.46	0.51	0.53	0.54	0.54	0.53	0.53	0.53
	CM	0.35	0.55	0.61	0.66	0.71	0.67	0.68	0.71	0.74	0.73	0.73
spa	LC	0.26	0.37	0.39	0.41	0.43	0.44	0.44	0.43	0.41	0.43	0.41
	HC	0.26	0.40	0.36	0.37	0.36	0.39	0.38	0.36	0.41	0.40	0.40
	LV	0.26	0.41	0.44	0.47	0.48	0.47	0.47	0.50	0.50	0.49	0.51
	LCS	0.26	0.39	0.46	0.46	0.49	0.50	0.50	0.50	0.50	0.51	0.51
	SM	0.26	0.37	0.43	0.45	0.46	0.45	0.47	0.48	0.49	0.49	0.50
	R	0.26	0.26	0.28	0.36	0.30	0.38	0.38	0.34	0.34	0.35	0.36
	CM	0.26	0.37	0.41	0.43	0.46	0.46	0.49	0.50	0.49	0.51	0.51

Table 13: **Derivation Task:** Accuracy results for LLaMA Model (*continued from the pervious table*)

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
amh	LC	0.01	0.02	0.04	0.04	0.05	0.15	0.10	0.12	0.13	0.20	0.12
	HC	0.01	0.00	0.00	0.00	0.01	0.02	0.04	0.05	0.04	0.05	0.06
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.01	0.14	0.20	0.24	0.26	0.27	0.28	0.30	0.31	0.32	0.34
	SM	0.01	0.05	0.07	0.09	0.11	0.13	0.13	0.14	0.15	0.15	0.16
	R	0.01	0.01	0.01	0.02	0.13	0.12	0.12	0.12	0.14	0.13	0.16
	CM	0.01	0.06	0.11	0.13	0.20	0.23	0.21	0.20	0.24	0.22	0.24
gal	LC	0.14	0.18	0.23	0.26	0.31	0.30	0.34	0.33	0.34	0.34	0.32
	HC	0.14	0.22	0.22	0.25	0.25	0.28	0.30	0.31	0.32	0.36	0.36
	LV	0.14	0.37	0.43	0.46	0.50	0.49	0.52	0.51	0.52	0.52	0.53
	LCS	0.14	0.36	0.41	0.45	0.48	0.47	0.50	0.50	0.51	0.52	0.52
	SM	0.14	0.31	0.36	0.38	0.41	0.42	0.43	0.44	0.44	0.43	0.43
	R	0.14	0.20	0.26	0.31	0.34	0.38	0.38	0.37	0.37	0.38	0.38
	CM	0.14	0.30	0.28	0.29	0.32	0.42	0.46	0.47	0.49	0.51	0.50
gle	LC	0.10	0.10	0.19	0.21	0.19	0.22	0.21	0.18	0.23	0.21	0.24
	HC	0.10	0.22	0.22	0.25	0.26	0.26	0.28	0.27	0.27	0.27	0.25
	LV	0.10	0.29	0.32	0.35	0.38	0.38	0.38	0.37	0.39	0.39	0.38
	LCS	0.10	0.29	0.33	0.37	0.38	0.38	0.38	0.38	0.37	0.37	0.38
	SM	0.10	0.27	0.31	0.33	0.36	0.37	0.39	0.37	0.38	0.39	0.37
	R	0.10	0.18	0.22	0.22	0.24	0.24	0.25	0.26	0.27	0.27	0.26
	CM	0.10	0.22	0.29	0.32	0.33	0.33	0.35	0.40	0.40	0.41	0.42
mon	LC	0.14	0.16	0.16	0.20	0.20	0.26	0.25	0.25	0.25	0.25	0.34
	HC	0.14	0.21	0.31	0.28	0.31	0.31	0.35	0.31	0.33	0.33	0.34
	LV	0.14	0.43	0.47	0.52	0.55	0.54	0.55	0.54	0.55	0.54	0.54
	LCS	0.14	0.37	0.44	0.48	0.46	0.49	0.51	0.51	0.52	0.50	0.53
	SM	0.14	0.33	0.39	0.42	0.44	0.44	0.46	0.47	0.48	0.48	0.49
	R	0.14	0.22	0.32	0.31	0.32	0.31	0.31	0.32	0.31	0.30	0.34
	CM	0.14	0.27	0.31	0.44	0.43	0.43	0.42	0.48	0.47	0.48	0.50
zul	LC	0.03	0.15	0.20	0.19	0.23	0.20	0.21	0.27	0.26	0.24	0.26
	HC	0.03	0.15	0.15	0.16	0.16	0.16	0.20	0.19	0.22	0.22	0.23
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.03	0.33	0.39	0.41	0.43	0.44	0.44	0.45	0.46	0.48	0.48
	SM	-	-	-	-	-	-	-	-	-	-	-
	R	0.03	0.12	0.18	0.20	0.20	0.22	0.22	0.22	0.22	0.22	0.23
	CM	0.03	0.20	0.25	0.29	0.33	0.41	0.42	0.46	0.46	0.46	0.46

Table 14: **Derivation Task:** Accuracy results for Qwen Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.28	0.38	0.39	0.43	0.45	0.45	0.46	0.45	0.45	0.46	0.46
	HC	0.28	0.30	0.38	0.37	0.40	0.40	0.40	0.44	0.42	0.42	0.43
	LV	0.28	0.46	0.47	0.48	0.49	0.50	0.50	0.51	0.52	0.52	0.52
	LCS	0.28	0.44	0.50	0.48	0.48	0.49	0.49	0.50	0.50	0.52	0.51
	SM	0.28	0.44	0.47	0.46	0.47	0.47	0.47	0.49	0.49	0.50	0.50
	R	0.28	0.42	0.42	0.42	0.45	0.46	0.48	0.45	0.47	0.48	0.48
	CM	0.28	0.38	0.40	0.40	0.46	0.42	0.45	0.48	0.48	0.49	0.49
fra	LC	0.10	0.16	0.20	0.15	0.14	0.14	0.14	0.16	0.18	0.19	0.21
	HC	0.10	0.21	0.24	0.28	0.25	0.24	0.23	0.25	0.24	0.24	0.25
	LV	0.10	0.28	0.31	0.33	0.33	0.34	0.33	0.35	0.37	0.38	0.37
	LCS	0.10	0.28	0.31	0.33	0.35	0.35	0.36	0.37	0.36	0.38	0.37
	SM	0.10	0.26	0.31	0.33	0.33	0.33	0.34	0.36	0.35	0.36	0.34
	R	0.10	0.20	0.25	0.29	0.25	0.27	0.28	0.28	0.29	0.29	0.30
	CM	0.10	0.20	0.26	0.30	0.31	0.33	0.35	0.35	0.37	0.34	0.36
hun	LC	0.15	0.36	0.34	0.36	0.37	0.37	0.39	0.40	0.38	0.39	0.40
	HC	0.15	0.24	0.29	0.33	0.36	0.37	0.37	0.32	0.31	0.33	0.34
	LV	0.15	0.42	0.47	0.49	0.50	0.52	0.52	0.52	0.53	0.52	0.53
	LCS	0.15	0.44	0.48	0.51	0.51	0.52	0.52	0.53	0.51	0.52	0.52
	SM	0.15	0.41	0.46	0.49	0.49	0.49	0.49	0.50	0.51	0.51	0.52
	R	0.15	0.22	0.29	0.33	0.32	0.32	0.32	0.30	0.34	0.34	0.34
	CM	0.15	0.30	0.32	0.38	0.40	0.43	0.46	0.47	0.48	0.49	0.51
rus	LC	0.31	0.45	0.46	0.46	0.46	0.49	0.54	0.52	0.54	0.54	0.57
	HC	0.31	0.42	0.39	0.45	0.44	0.45	0.48	0.51	0.49	0.51	0.52
	LV	0.31	0.58	0.64	0.67	0.67	0.69	0.69	0.69	0.71	0.70	0.71
	LCS	0.31	0.59	0.65	0.65	0.67	0.67	0.68	0.69	0.69	0.71	0.71
	SM	0.31	0.57	0.61	0.64	0.65	0.65	0.66	0.66	0.68	0.67	0.68
	R	0.31	0.44	0.41	0.43	0.42	0.43	0.44	0.43	0.42	0.45	0.42
	CM	0.31	0.48	0.54	0.60	0.62	0.64	0.66	0.69	0.69	0.72	0.71
spa	LC	0.10	0.24	0.24	0.27	0.30	0.28	0.29	0.29	0.28	0.31	0.27
	HC	0.10	0.24	0.29	0.26	0.28	0.30	0.29	0.33	0.30	0.28	0.31
	LV	0.10	0.37	0.41	0.42	0.44	0.44	0.44	0.45	0.46	0.48	0.47
	LCS	0.10	0.36	0.42	0.42	0.44	0.44	0.44	0.45	0.47	0.47	0.47
	SM	0.10	0.29	0.37	0.39	0.40	0.40	0.41	0.43	0.44	0.44	0.44
	R	0.10	0.23	0.29	0.34	0.26	0.28	0.29	0.31	0.33	0.31	0.34
	CM	0.10	0.27	0.32	0.36	0.42	0.43	0.42	0.45	0.45	0.46	0.47

Table 15: **Derivation Task:** Accuracy results for Qwen Model (*continued from the pervious table*)

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
amh	LC	0.00	0.10	0.08	0.08	0.08	0.08	0.10	0.10	0.10	0.13	0.13
	HC	0.00	0.01	0.02	0.01	0.04	0.12	0.13	0.15	0.14	0.15	0.16
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.00	0.10	0.15	0.20	0.22	0.24	0.26	0.27	0.28	0.28	0.28
	SM	0.00	0.06	0.09	0.12	0.13	0.15	0.16	0.18	0.18	0.18	0.19
	R	0.00	0.00	0.00	0.01	0.13	0.14	0.15	0.14	0.14	0.15	0.17
	CM	0.00	0.09	0.13	0.11	0.18	0.19	0.23	0.22	0.24	0.27	0.27
gal	LC	0.15	0.33	0.32	0.37	0.39	0.37	0.37	0.38	0.38	0.38	0.38
	HC	0.15	0.28	0.32	0.36	0.36	0.38	0.39	0.39	0.40	0.41	0.38
	LV	0.15	0.37	0.44	0.47	0.50	0.51	0.52	0.50	0.51	0.50	0.50
	LCS	0.15	0.36	0.42	0.44	0.46	0.47	0.49	0.49	0.48	0.49	0.50
	SM	0.15	0.31	0.35	0.36	0.40	0.41	0.42	0.43	0.43	0.44	0.44
	R	0.15	0.21	0.29	0.32	0.34	0.37	0.37	0.36	0.36	0.35	0.36
	CM	0.15	0.29	0.31	0.36	0.38	0.38	0.42	0.43	0.46	0.47	0.46
gle	LC	0.11	0.21	0.23	0.24	0.26	0.25	0.26	0.27	0.26	0.25	0.25
	HC	0.11	0.16	0.20	0.21	0.24	0.23	0.25	0.24	0.25	0.24	0.26
	LV	0.11	0.30	0.35	0.38	0.38	0.39	0.39	0.39	0.41	0.40	0.40
	LCS	0.11	0.29	0.33	0.34	0.36	0.37	0.38	0.38	0.37	0.38	0.39
	SM	0.11	0.29	0.31	0.34	0.33	0.33	0.34	0.35	0.36	0.35	0.36
	R	0.11	0.20	0.21	0.21	0.22	0.23	0.22	0.23	0.23	0.23	0.23
	CM	0.11	0.23	0.26	0.27	0.28	0.31	0.33	0.34	0.36	0.37	0.37
mon	LC	0.11	0.15	0.16	0.17	0.17	0.18	0.23	0.22	0.22	0.22	0.23
	HC	0.11	0.22	0.29	0.30	0.28	0.28	0.28	0.26	0.25	0.29	0.28
	LV	0.11	0.40	0.44	0.47	0.50	0.49	0.50	0.49	0.50	0.50	0.49
	LCS	0.11	0.35	0.42	0.43	0.46	0.48	0.48	0.50	0.49	0.50	0.49
	SM	0.11	0.30	0.37	0.39	0.40	0.41	0.42	0.44	0.45	0.45	0.45
	R	0.11	0.22	0.26	0.22	0.24	0.24	0.25	0.26	0.25	0.25	0.30
	CM	0.11	0.22	0.33	0.34	0.38	0.41	0.43	0.44	0.45	0.47	0.45
zul	LC	0.04	0.12	0.11	0.15	0.16	0.17	0.18	0.18	0.17	0.18	0.18
	HC	0.04	0.09	0.11	0.13	0.13	0.13	0.14	0.14	0.15	0.14	0.15
	LV	-	-	-	-	-	-	-	-	-	-	-
	LCS	0.04	0.32	0.36	0.38	0.41	0.42	0.43	0.44	0.44	0.45	0.46
	SM	-	-	-	-	-	-	-	-	-	-	-
	R	0.04	0.10	0.10	0.11	0.13	0.14	0.13	0.16	0.16	0.15	0.16
	CM	0.04	0.15	0.20	0.24	0.35	0.38	0.40	0.42	0.39	0.44	0.44

Table 16: **Inflection segmentation task:** Accuracy results for LLaMA Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.36	0.66	0.73	0.76	0.76	0.77	0.80	0.80	0.79	0.77	0.79
	HC	0.36	0.68	0.77	0.78	0.81	0.83	0.84	0.84	0.83	0.83	0.85
	LV	0.36	0.77	0.81	0.85	0.86	0.87	0.87	0.88	0.88	0.88	0.88
	LCS	0.36	0.74	0.79	0.82	0.84	0.84	0.84	0.85	0.86	0.86	0.86
	SM	0.36	0.77	0.82	0.84	0.85	0.86	0.87	0.88	0.88	0.87	0.87
	R	0.36	0.70	0.73	0.72	0.75	0.77	0.75	0.77	0.77	0.77	0.77
	CM	0.36	0.74	0.78	0.80	0.82	0.83	0.83	0.86	0.86	0.85	0.86
	LF	-	-	-	-	-	-	-	-	-	-	-
fra	LC	0.10	0.33	0.42	0.46	0.47	0.46	0.46	0.46	0.47	0.47	0.47
	HC	0.10	0.14	0.13	0.14	0.15	0.25	0.28	0.30	0.32	0.34	0.40
	LV	0.10	0.49	0.61	0.66	0.70	0.72	0.73	0.74	0.73	0.73	0.75
	LCS	0.10	0.47	0.58	0.66	0.69	0.71	0.72	0.72	0.74	0.73	0.73
	SM	0.10	0.54	0.63	0.70	0.72	0.75	0.77	0.78	0.78	0.77	0.78
	R	0.10	0.37	0.19	0.34	0.41	0.40	0.45	0.45	0.45	0.45	0.46
	CM	0.10	0.30	0.50	0.59	0.62	0.68	0.70	0.71	0.74	0.73	0.74
	LF	-	-	-	-	-	-	-	-	-	-	-
hun	LC	0.40	0.53	0.56	0.61	0.65	0.61	0.60	0.61	0.64	0.64	0.65
	HC	0.40	0.58	0.61	0.63	0.64	0.64	0.64	0.64	0.64	0.64	0.63
	LV	0.39	0.64	0.69	0.73	0.74	0.76	0.78	0.77	0.77	0.77	0.78
	LCS	0.39	0.70	0.72	0.75	0.77	0.78	0.80	0.80	0.79	0.78	0.80
	SM	0.39	0.68	0.74	0.77	0.79	0.80	0.80	0.81	0.80	0.81	0.81
	R	0.40	0.58	0.60	0.63	0.61	0.63	0.63	0.63	0.64	0.64	0.66
	CM	0.39	0.57	0.64	0.69	0.72	0.73	0.75	0.76	0.79	0.80	0.80
	LF	-	-	-	-	-	-	-	-	-	-	-
rus	LC	0.09	0.10	0.14	0.16	0.16	0.14	0.16	0.20	0.20	0.19	0.20
	HC	0.09	0.15	0.13	0.20	0.22	0.22	0.18	0.24	0.25	0.27	0.28
	LV	0.09	0.34	0.48	0.55	0.60	0.60	0.61	0.62	0.61	0.63	0.64
	LCS	0.09	0.36	0.48	0.56	0.60	0.60	0.63	0.65	0.64	0.65	0.65
	SM	0.09	0.37	0.47	0.51	0.54	0.56	0.57	0.59	0.60	0.61	0.61
	R	0.09	0.15	0.13	0.14	0.19	0.19	0.19	0.21	0.22	0.22	0.24
	CM	0.09	0.21	0.33	0.38	0.42	0.49	0.57	0.58	0.61	0.62	0.65
	LF	-	-	-	-	-	-	-	-	-	-	-
spa	LC	0.09	0.14	0.18	0.30	0.33	0.36	0.34	0.36	0.36	0.36	0.35
	HC	0.09	0.36	0.30	0.35	0.35	0.35	0.36	0.38	0.40	0.38	0.39
	LV	0.09	0.49	0.57	0.64	0.66	0.68	0.70	0.71	0.71	0.73	0.73
	LCS	0.09	0.51	0.60	0.65	0.69	0.70	0.70	0.72	0.72	0.73	0.74
	SM	0.09	0.50	0.60	0.65	0.68	0.70	0.69	0.70	0.72	0.72	0.73
	R	0.09	0.30	0.33	0.36	0.33	0.36	0.32	0.36	0.36	0.37	0.37
	CM	0.09	0.42	0.47	0.52	0.60	0.66	0.68	0.69	0.70	0.71	0.72
	LF	-	-	-	-	-	-	-	-	-	-	-

Table 17: **Inflection segmentation task:** Accuracy results for Qwen Model

lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
eng	LC	0.29	0.52	0.52	0.59	0.60	0.63	0.64	0.66	0.67	0.66	0.68
	HC	0.29	0.47	0.52	0.52	0.54	0.60	0.67	0.65	0.67	0.70	0.71
	LV	0.30	0.62	0.70	0.72	0.74	0.74	0.75	0.77	0.78	0.79	0.78
	LCS	0.30	0.60	0.67	0.69	0.70	0.73	0.74	0.75	0.77	0.77	0.78
	SM	0.30	0.59	0.66	0.70	0.72	0.74	0.76	0.76	0.77	0.78	0.79
	R	0.29	0.44	0.58	0.57	0.61	0.65	0.62	0.65	0.66	0.66	0.65
	CM	0.30	0.54	0.66	0.69	0.70	0.72	0.73	0.74	0.75	0.76	0.78
	LF	-	-	-	-	-	-	-	-	-	-	-
fra	LC	0.04	0.09	0.10	0.12	0.29	0.30	0.34	0.33	0.41	0.44	0.45
	HC	0.04	0.10	0.22	0.30	0.28	0.41	0.42	0.44	0.41	0.42	0.49
	LV	0.04	0.45	0.56	0.63	0.68	0.71	0.74	0.75	0.76	0.75	0.75
	LCS	0.04	0.43	0.54	0.60	0.63	0.66	0.68	0.69	0.70	0.70	0.71
	SM	0.04	0.49	0.59	0.64	0.68	0.71	0.72	0.74	0.73	0.74	0.74
	R	0.04	0.38	0.35	0.42	0.43	0.42	0.46	0.45	0.46	0.45	0.46
	CM	0.04	0.30	0.44	0.57	0.57	0.63	0.66	0.70	0.69	0.71	0.71
	LF	-	-	-	-	-	-	-	-	-	-	-
hun	LC	0.20	0.27	0.34	0.39	0.39	0.42	0.46	0.47	0.48	0.49	0.47
	HC	0.20	0.31	0.39	0.41	0.39	0.40	0.42	0.42	0.45	0.44	0.49
	LV	0.20	0.51	0.56	0.58	0.60	0.62	0.66	0.67	0.66	0.64	0.65
	LCS	0.20	0.50	0.59	0.63	0.64	0.64	0.66	0.66	0.68	0.67	0.69
	SM	0.20	0.52	0.60	0.62	0.64	0.65	0.66	0.67	0.68	0.68	0.68
	R	0.20	0.35	0.39	0.45	0.45	0.43	0.45	0.47	0.46	0.49	0.50
	CM	0.20	0.38	0.43	0.49	0.55	0.61	0.61	0.64	0.66	0.66	0.67
	LF	-	-	-	-	-	-	-	-	-	-	-
rus	LC	0.06	0.06	0.10	0.12	0.16	0.19	0.20	0.18	0.17	0.19	0.18
	HC	0.06	0.11	0.13	0.18	0.18	0.21	0.19	0.22	0.24	0.24	0.25
	LV	0.06	0.29	0.41	0.46	0.51	0.52	0.55	0.56	0.57	0.58	0.57
	LCS	0.06	0.31	0.40	0.46	0.50	0.52	0.56	0.56	0.58	0.58	0.60
	SM	0.06	0.32	0.40	0.41	0.44	0.45	0.49	0.49	0.50	0.52	0.52
	R	0.06	0.16	0.17	0.14	0.21	0.22	0.24	0.23	0.24	0.25	0.25
	CM	0.06	0.18	0.24	0.34	0.40	0.44	0.51	0.52	0.54	0.53	0.56
	LF	-	-	-	-	-	-	-	-	-	-	-
spa	LC	0.05	0.24	0.25	0.29	0.28	0.31	0.31	0.34	0.28	0.33	0.34
	HC	0.05	0.23	0.26	0.33	0.34	0.37	0.38	0.37	0.38	0.41	0.44
	LV	0.05	0.41	0.53	0.61	0.62	0.67	0.70	0.70	0.71	0.72	0.74
	LCS	0.05	0.42	0.56	0.62	0.67	0.68	0.71	0.73	0.75	0.74	0.75
	SM	0.05	0.44	0.54	0.59	0.62	0.64	0.66	0.68	0.69	0.70	0.70
	R	0.05	0.27	0.29	0.29	0.29	0.31	0.32	0.36	0.35	0.38	0.38
	CM	0.05	0.34	0.42	0.51	0.58	0.60	0.66	0.68	0.73	0.71	0.74
	LF	-	-	-	-	-	-	-	-	-	-	-

Table 18: **Derivation and Inflection tasks:** ECE results of LLaMA model

Task	lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
Derivation	eng	HC	0.48	0.55	0.56	0.55	0.55	0.55	0.55	0.55	0.55	0.53	0.54
		LC	0.48	0.49	0.52	0.49	0.47	0.49	0.49	0.51	0.51	0.49	0.51
	fra	HC	0.28	0.32	0.33	0.34	0.31	0.33	0.33	0.34	0.31	0.32	0.32
		LC	0.28	0.36	0.35	0.33	0.34	0.36	0.35	0.37	0.36	0.38	0.38
	gal	HC	0.20	0.35	0.31	0.33	0.31	0.34	0.37	0.37	0.37	0.39	0.41
		LC	0.20	0.28	0.33	0.33	0.39	0.39	0.42	0.40	0.41	0.40	0.38
	gle	HC	0.13	0.28	0.29	0.32	0.33	0.33	0.35	0.33	0.34	0.33	0.32
		LC	0.13	0.16	0.24	0.28	0.24	0.29	0.28	0.26	0.30	0.28	0.31
	hun	HC	0.36	0.41	0.45	0.43	0.44	0.44	0.46	0.46	0.47	0.47	0.46
		LC	0.36	0.42	0.41	0.42	0.43	0.45	0.44	0.45	0.46	0.45	0.47
	mon	HC	0.10	0.16	0.19	0.26	0.22	0.24	0.25	0.26	0.28	0.27	0.27
		LC	0.10	0.15	0.24	0.24	0.24	0.25	0.28	0.31	0.31	0.32	0.34
	rus	HC	0.38	0.50	0.54	0.55	0.58	0.55	0.56	0.55	0.55	0.54	0.55
		LC	0.38	0.50	0.51	0.52	0.54	0.54	0.55	0.54	0.55	0.55	0.53
	spa	HC	0.30	0.41	0.38	0.39	0.37	0.41	0.37	0.37	0.42	0.40	0.41
		LC	0.30	0.39	0.42	0.42	0.44	0.45	0.45	0.45	0.44	0.44	0.42
Inflection	eng	HC	0.55	0.63	0.60	0.60	0.59	0.58	0.58	0.59	0.59	0.58	0.58
		LC	0.55	0.58	0.61	0.63	0.64	0.64	0.67	0.67	0.67	0.64	0.61
	fra	HC	0.50	0.51	0.56	0.55	0.55	0.53	0.60	0.58	0.56	0.53	0.52
		LC	0.50	0.55	0.50	0.51	0.49	0.52	0.51	0.50	0.49	0.48	0.48
	gal	HC	0.34	0.38	0.35	0.34	0.35	0.35	0.36	0.41	0.42	0.40	0.42
		LC	0.34	0.36	0.38	0.36	0.40	0.38	0.38	0.37	0.39	0.38	0.36
	gle	HC	0.15	0.22	0.25	0.27	0.29	0.29	0.29	0.30	0.31	0.29	0.31
		LC	0.15	0.23	0.26	0.26	0.26	0.27	0.24	0.25	0.27	0.26	0.26
	hun	HC	0.44	0.46	0.46	0.47	0.46	0.45	0.42	0.44	0.46	0.45	0.47
		LC	0.44	0.44	0.44	0.43	0.43	0.43	0.45	0.46	0.45	0.47	0.46
	mon	HC	0.16	0.14	0.14	0.18	0.18	0.18	0.18	0.15	0.15	0.15	0.14
		LC	0.16	0.20	0.38	0.38	0.40	0.41	0.40	0.38	0.43	0.44	0.39
	rus	HC	0.44	0.45	0.46	0.45	0.44	0.45	0.46	0.47	0.47	0.49	0.48
		LC	0.44	0.45	0.45	0.47	0.46	0.47	0.48	0.49	0.49	0.47	0.46
	spa	HC	0.41	0.40	0.38	0.38	0.39	0.43	0.41	0.42	0.42	0.44	0.45
		LC	0.41	0.39	0.39	0.38	0.42	0.40	0.42	0.42	0.42	0.41	0.43

Table 19: **Derivation and Inflection tasks:** ECE results of Qwen model

Task	lang	method	0shot	1shot	2shot	3shot	4shot	5shot	6shot	7shot	8shot	9shot	10shot
Derivation	eng	HC	0.33	0.38	0.46	0.44	0.47	0.46	0.47	0.49	0.48	0.47	0.48
		LC	0.33	0.43	0.44	0.47	0.49	0.49	0.50	0.50	0.49	0.49	0.50
	fra	HC	0.13	0.27	0.29	0.31	0.29	0.27	0.28	0.30	0.30	0.30	0.30
		LC	0.13	0.21	0.26	0.19	0.19	0.21	0.19	0.22	0.23	0.24	0.25
	gal	HC	0.19	0.35	0.36	0.39	0.42	0.43	0.42	0.42	0.42	0.43	0.42
		LC	0.19	0.40	0.41	0.41	0.42	0.42	0.42	0.42	0.42	0.43	0.43
	gle	HC	0.14	0.21	0.26	0.26	0.29	0.27	0.30	0.30	0.29	0.29	0.31
		LC	0.14	0.27	0.28	0.29	0.31	0.31	0.32	0.33	0.32	0.31	0.31
	hun	HC	0.16	0.29	0.34	0.37	0.39	0.39	0.39	0.35	0.34	0.35	0.36
		LC	0.16	0.37	0.37	0.38	0.39	0.39	0.42	0.42	0.41	0.42	0.42
	mon	HC	0.09	0.20	0.26	0.22	0.24	0.24	0.23	0.20	0.15	0.19	0.21
		LC	0.09	0.15	0.16	0.26	0.25	0.25	0.27	0.24	0.27	0.27	0.27
	rus	HC	0.29	0.42	0.41	0.43	0.44	0.44	0.44	0.49	0.47	0.49	0.49
		LC	0.29	0.43	0.45	0.45	0.45	0.46	0.51	0.50	0.52	0.51	0.55
	spa	HC	0.12	0.30	0.33	0.30	0.31	0.34	0.32	0.34	0.33	0.32	0.35
		LC	0.12	0.28	0.30	0.33	0.34	0.33	0.33	0.34	0.33	0.35	0.32
Inflection	eng	HC	0.36	0.67	0.69	0.71	0.66	0.70	0.69	0.69	0.69	0.65	0.64
		LC	0.36	0.69	0.68	0.61	0.68	0.66	0.67	0.69	0.67	0.66	0.66
	fra	HC	0.33	0.57	0.56	0.56	0.56	0.52	0.57	0.52	0.52	0.53	0.52
		LC	0.33	0.57	0.60	0.58	0.60	0.64	0.64	0.57	0.55	0.52	0.54
	gal	HC	0.12	0.30	0.32	0.34	0.32	0.31	0.29	0.32	0.37	0.35	0.36
		LC	0.12	0.31	0.32	0.35	0.35	0.32	0.35	0.35	0.35	0.32	0.34
	gle	HC	0.06	0.04	0.06	0.10	0.09	0.12	0.14	0.15	0.17	0.18	0.19
		LC	0.06	0.04	0.07	0.09	0.11	0.12	0.13	0.15	0.15	0.15	0.14
	hun	HC	0.18	0.28	0.32	0.31	0.30	0.30	0.33	0.33	0.29	0.31	0.30
		LC	0.18	0.29	0.29	0.30	0.32	0.29	0.31	0.30	0.33	0.31	0.31
	mon	HC	0.09	0.12	0.19	0.27	0.23	0.23	0.29	0.31	0.30	0.31	0.33
		LC	0.09	0.18	0.22	0.21	0.24	0.25	0.25	0.24	0.26	0.26	0.26
	rus	HC	0.29	0.42	0.44	0.46	0.45	0.46	0.50	0.51	0.54	0.53	0.52
		LC	0.29	0.37	0.51	0.49	0.46	0.47	0.47	0.50	0.48	0.50	0.53
	spa	HC	0.21	0.39	0.42	0.41	0.42	0.43	0.44	0.45	0.44	0.45	0.42
		LC	0.21	0.38	0.39	0.44	0.45	0.45	0.46	0.47	0.46	0.46	0.45