Evidence-Augmented Generation Reasoning for Extremely Low-Resource Language Decipherment

Xiaoyu Zhu[1,2], **Longyuan**[1,2], **Rui Qi**[1,2], **Jinan Xu**[1,2†]
[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation
(Beijing Jiaotong University), Ministry of Education
[2]School of Computer Science and Technology, Beijing Jiaotong University
{24120420, jaxu}@bjtu.edu.cn

Abstract

Inspired by linguistic Olympiads, extremely low-resource language reasoning presents a unique challenge that enables models to solve problems without prior knowledge. This task mirrors the Rosetta Stone decipherment process, where the goal is to induce and apply linguistic rules from minimal context. Existing methods mainly rely on naive in-context learning that fails to handle the complexity and diversity of language rules. To mitigate this issue, we propose a framework that combines dynamic knowledge construction with task-aware evidence augmentation. First, we use large language models (LLMs) to generate a diverse set of task-specific examples that instantiate potential linguistic rules for the target low-resource language. Second, we apply a semantic retrieval mechanism to select the most relevant examples as evidence for each test query, preventing context overload and ensuring focused, analogical reasoning. Our method shifts from learning language distributions to dynamically discovering and applying rules. Experimental results on the LINGOLY and Linguini benchmark show that our approach achieves competitive performance across various LLMs, outperforming existing baselines. More importantly, our framework advances extremely low-resource reasoning and provides a generalizable framework for rule induction under knowledge constraints.

1 Introduction

Deciphering unknown linguistic systems is a hallmark of human intelligence, exemplified by the Rosetta Stone (Bozhanov and Derzhanski, 2013). The Rosetta Stone question presents paired examples of a low-resource language alongside its English translation, requiring solvers to deduce the language's grammar and vocabulary solely from those minimal clues (Vogt et al., 2023). For machine learning models, to do this well, it must possess multiple capabilities, including summarizing grammatical, morphological and semantic rules from several examples and applying them to new problems (Lake and Baroni, 2017).

As a conversion to the deciphering problem of the Rosetta Stone, recent benchmarks like LINGOLY (Bean et al., 2024) and Linguini (Sánchez et al., 2024) formalize this challenge for machine learning models. These benchmarks construct tasks using low-resource or artificial languages, ensuring that models cannot rely on memorized knowledge, which poses a substantial challenge for current large language models (LLMs) with strong reasoning capabilities. As shown in Figure 1, each problem is designed as a Rosetta-style task, where the model needs to infer linguistic rules solely from the provided context or require deductive reasoning, such as translating unseen sentences, aligning word pairs, or inferring morphological rules. Since the problems come from extremely low-resource or extinct languages, the LLMs do not have knowledge of those languages and can only attempt reasoning and deduction without prior knowledge of the target language. As a result, the central scientific problem is: ***How can models induce and apply linguistic rules from minimal contextual examples in unfamiliar languages?***

Recent works on deciphering low-resource languages with LLMs have explored few-shot and chain-of-thought prompting (Wei et al., 2022), where models are asked to generalize from a handful of translation exemplars to uncover grammatical rules in unseen languages. However, this approach only captures surface patterns, which fails to capture deeper cross-linguistic patterns, generally limited with weak capabilities for the target language. On the other hand, the advanced approach is analogical prompting (Ramji and Ramji, 2025), where auxiliary exemplars are automatically generated in related, higher-resource languages and combined with target examples. This two-stage reasoning procedure enables models such as GPT-4o

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 14–29

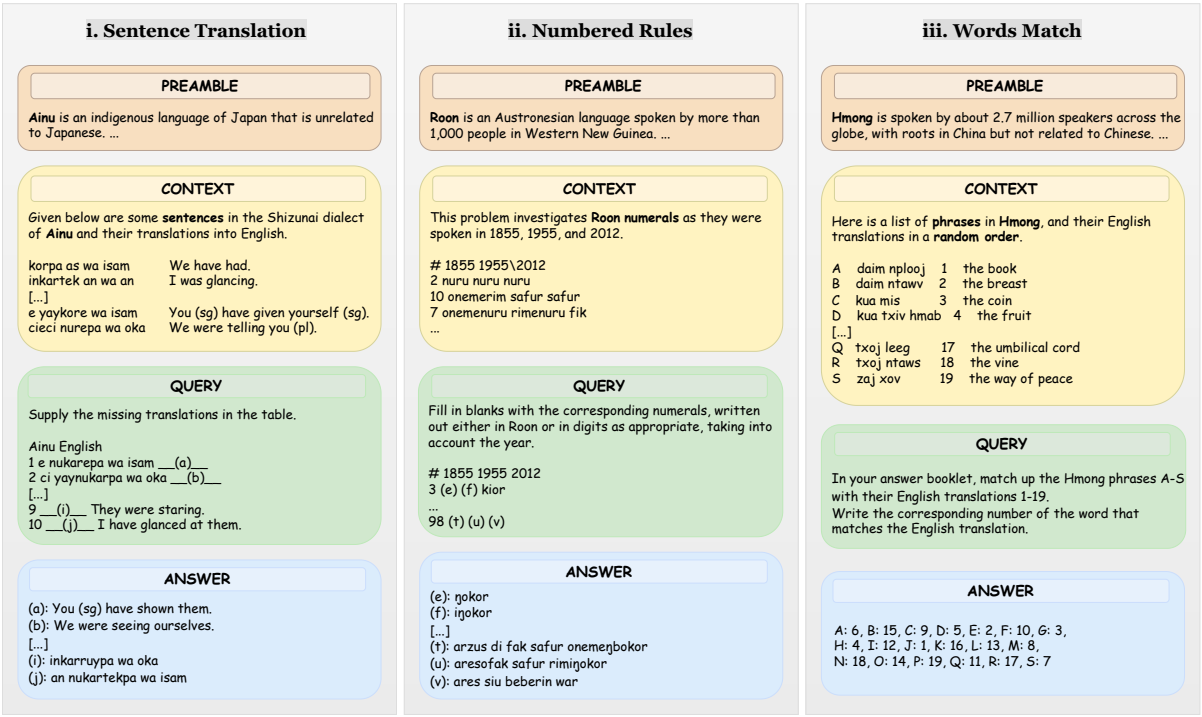


Figure 1: Rosetta Stone linguistic problem example. Each question contains four parts, PREAMBLE: introducing the background information of the current low resource language, CONTEXT: a small number of translation pairs between the current language and English, QUERY: the question set based on context, and ANSWER: the correct answer corresponding to query.

and Llama-3.1-405B to leverage their latent multilingual knowledge more effectively. Although analogical and chain-of-thought prompting improve over naive few-shot learning, models still struggle with multi-step reasoning, complex morphosyntactic generalizations, and robust handling of languages absent from pretraining. This gap underscores the need for novel approaches to reasoning-based decipherment in low-resource settings.

To alleviate this gap, we introduce an evidence-augmented framework through two stages: **Dynamic Knowledge Construction** and **Task-Aware Evidence Augmentation**, which realizes the decipherment reasoning with dynamic and flexible linguistic rules instead of factual data. In particular, for the **Dynamic Knowledge Construction**, we leverage LLMs to generate diverse rule-guided exemplars, forming a dynamic knowledge base. These exemplars are not used as extra training data; they are intermediate knowledge that will later be filtered and transformed into concise evidence for each query. The first stage, Dynamic Knowledge Construction, is necessary to proactively fill the fundamental knowledge vacuum by generating a rich set of rule-embodying examples, moving beyond the limited patterns present in the original context. However, the sheer volume of generated knowledge necessitates the second stage, Task-Aware Evidence Augmentation, which acts as a critical filter to mitigate information overload and ensure that the model's reasoning is guided by the most pertinent analogies for each specific problem.

For the **Task-Aware Evidence Augmentation**, to reduce inefficiency and noise, we retrieve the top-K most relevant examples for each test query, enabling focused analogical reasoning. In particular, the necessity of retrieval in our framework is grounded in the two theoretical foundations: **(1) Cognitive Load Theory.** The limited context window of transformers mirrors human working memory (Leng et al., 2024). By retrieving only the top-K most relevant examples, we reduce extraneous cognitive load, enabling the model to focus on the key rule patterns necessary for solving the task. **(2) Analogical Reasoning Theory.** Human problem-solving often relies on analogy, transferring knowledge from similar past cases (Wang et al., 2024; Musker et al., 2024; Wei et al., 2022). We leverage a retriever as an evidence-seeking mechanism for identifying analogous cases (e.g. "What was the correct form for an adjective describing a 'house' in a similar sentence?"), allowing the model to apply

analogical reasoning effectively. Beyond benchmark performance, the broader implications of this work are profound. Success in low-resource language inference contributes to the revitalization of endangered languages, improves multilingual adaptability in real-world settings, and offers a computational account of how humans induce rules from limited evidence. Overall, our contributions are both methodological and conceptual:

- We formalize low-resource language decipherment as a dynamic rule induction task and achieves competitive performance on the LINGOLY and Linguini, compared with existing methods.
- Our proposed framework combines rule generation and retrieval to simulate human-like decipherment, which is applicable to different types of Rosetta Stone problems, such as sentence translation and words match.
- Experimental results demonstrate that the approach is effective based on both open-source and commercial LLMs with various sizes and can further stimulate LLMs with strong inherent reasoning capabilities.

2 Related Work

The approaches that can be transferred to solve the decryption problem of the Rosetta Stone are to infer linguistic rules and conduct deductive reasoning. The first branch of the approaches is *in-context learning* (ICL), where the ability of LLMs to learn from examples provided in the prompt (Brown et al., 2020; Li, 2023; Luo et al., 2024). Standard approaches to our task, such as directly providing context examples, rely on this capability. However, the performance of ICL is often limited by the number of examples available in the context (typically 10). In contrast, our work enhances ICL by massively scaling up the number of potential examples through generation and then intelligently selecting the most relevant ones. This augmentation goes beyond the constraints of the original problem's limited context, allowing for more efficient and focused reasoning. The second branch of the approaches is the *Retrieval-Augmented Generation* (RAG) (Lewis et al., 2020; Sánchez et al., 2024; Zhao et al., 2024; Zhang et al., 2024; Fan et al., 2025) which integrates parametric knowledge stored in model weights with non-parametric knowledge from external corpora via a retriever. Based on the paradigm of RAG, we transfer this methodology into alleviating the issue of the Rosetta Stone decryption.

Moreover, in comparison to the popular RAG methods, the knowledge database needs to be constructed from-scratch in our scenarios. LLMs have been used to generate synthetic data for training or enhancement (Anil et al., 2023), while our approach does not generate data to modify the model weights (e.g., by fine-tuning). Instead, we generate contextual knowledge on-the-fly, which serves as non-parametric, in-context clues (Wang et al., 2023). This is akin to "self-generated prompts" or "knowledge distillation" from the model, focused on instantiating linguistic rules inferred directly from the problem context. Our approach enables flexible, real-time generation of task-specific knowledge without the need for fine-tuning (Zhou et al., 2022), distinguishing it from typical synthetic data generation approaches.

There are several other reasoning tasks related to decryption which requires rule induction. **Cryptography and Puzzle Solving.** Deciphering an unknown language is akin to breaking a cipher (e.g., substitution ciphers) (Jakobsen, 1995), where the model must find mappings and patterns between the unknown language and a known one. **Inductive Logic Programming.** Inductive Logic Programming (ILP) (Muggleton, 1991; Mooney, 1996; Raedt, 2008) aims to learn logical rules from positive and negative examples, typically requiring formalizing features into logical predicates, a process that is challenging for complex, low-resource linguistic phenomena. **Meta-Learning and Multi-Task Learning.** Training models on a distribution of related tasks (e.g., many language inference problems) to acquire a "learning-to-learn" capability is a promising direction (Hospedales et al., 2022; Gharoun et al., 2024). However, this approach requires large, diverse training data and risks overfitting to the seen task distribution. Our method, by contrast, is a zero-shot, in-context approach that requires no additional training, making it more generalizable to entirely novel languages and tasks.

In summary, prior efforts have either emphasized parametric learning (ICL, meta-learning) or static augmentation (synthetic data generation). However, these approaches struggle when the target distribution is entirely unseen and cannot be inferred from memorized knowledge. Our method differs fundamentally in that it treats the problem as an

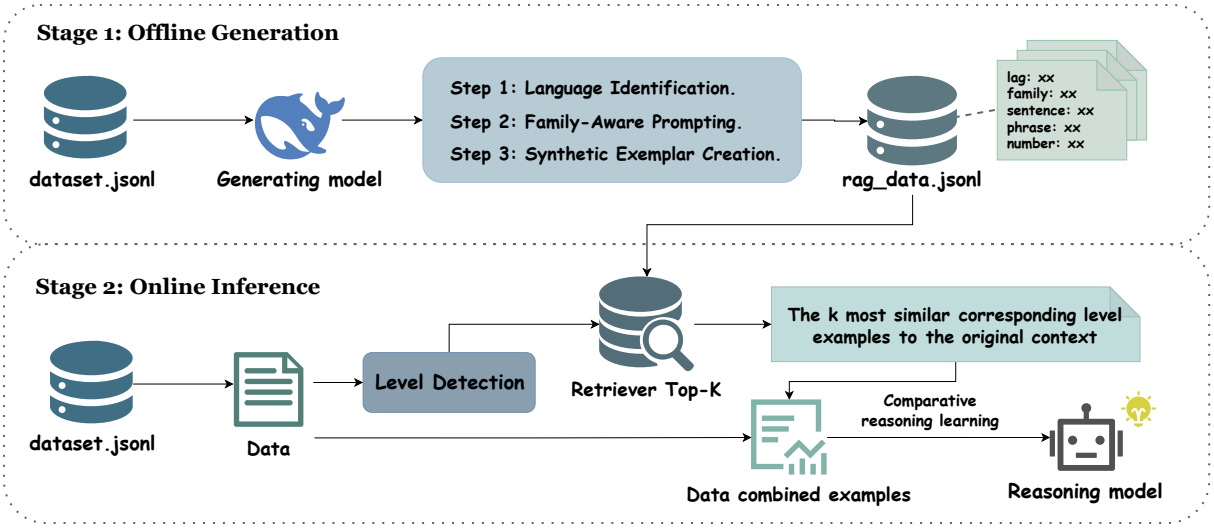


Figure 2: Overview of EASEG. During the Offline Generation stage, we identify the low-resource language L and generate 15 synthetic exemplars per level. During the Online Inference stage, for each test example we retrieve the top-k exemplars and prepend them to the prompt.

online reasoning task: knowledge is not pre-stored but dynamically constructed and filtered. This distinction allows our framework to better simulate the human decipherment process, where knowledge is actively hypothesized and revised in real time rather than recalled from a fixed memory. In this sense, our work bridges the gap between information retrieval, analogical reasoning, and computational linguistics.

3 Method

3.1 Overview of Evidence-Augmented Synthetic Exemplar Generation

We denote the low-resource language reasoning dataset as $D = (x_1, y_1), ..., (x_n, y_n)$, where each instance x_i consists of

- **context** C_i: a short passage containing a few surface forms (sentences, phrases or numerals) in an unseen language L_i
- **query** Q_i: a question about the underlying rule of L_i.
- **gold answer** y_i.

The model involves C_i and Q_i at test stage and predicts y_i without any prior knowledge of L_i. Our goal is to lift the model's performance on y_i by automatically enriching its working memory with synthetic, task-relevant exemplars of L_i.

To alleviate this problem, we propose EASEG (Evidence-Augmented Synthetic Exemplar Generation), a two-stage pipeline, as illustrated in 2. The framework consists of two components:

- **Offline Generation (§3.1.1).** Deepseek-R1 writes synthetic exemplars for every language L in D at three linguistic levels (sentence, phrase, number). These exemplars are stored in a retrieval index R.
- **Online Inference (§3.1.2).** Given a test instance (C, Q) in language L and level ℓ, the system retrieves the top-k most semantically similar exemplars from R and concatenates them to C before feeding the enriched prompt to the reasoning model.

3.1.1 Offline Generation of R

We populate the retrieval index R in three steps.

Step 1: Language Identification. For every example $(x, y) \in D$, we utilize DeepSeek-R1 with a zero-shot prompt: "Identify the low-resource language in the following context. Return only the ISO 639-3 code.", to obtain L.

Step 2: Family-Aware Prompting. We also query DeepSeek-R1 for the language family F(L). This meta-information is later used to bias exemplar generation towards typologically similar languages, increasing the chance of capturing relevant morpho-syntactic patterns.

Step 3: Synthetic Exemplar Creation. For each pair $(L, F(L))$ we prompt DeepSeek-R1 to produce 5 × 3 = 15 synthetic exemplars: 5 sentence-level ⟨low-resource, English⟩ translation pairs, 5 phrase-level pairs and 5 number-level pairs.

Since the languages of each data in the dataset are repeated, the total number of synthetic exemplars in some languages exceeds 15. All exemplars

are assigned a key-value record and appended to R. The prompt used to generate additional examples is in Appendix A.

3.1.2 Online Inference

Given a test instance (C, Q) we perform retrieval as follows.

Step 1: Level Detection.

We classify the task granularity ℓ with a rule-based trigger set:
$tokens \in \{'phrases', 'words', 'phrase', 'word'\}$
$\longrightarrow \ell = phrase$
$tokens \in \{'numbers', 'numerical', 'number'\}$
$\longrightarrow \ell = number$
$otherwise \longrightarrow \ell = sentence.$

Step 2: Evidence-Seeking.

We query R with $lag = L$ and $level = \ell$, then rank the exemplars by the similarity of the retriever to C. The top-k exemplars $E_1 \ldots E_k$ are concatenated to C in order of similarity.

Step 3: Reasoning.

Put the retrieved enhanced test cases (C, q) into the designed prompt and input them to model for reasoning. The prompts used to ultimately enable the model to reason are in Appendix C.

3.2 Dynamic Knowledge Construction

To construct the knowledge database from-scratch for decipherment, we adopt a dynamic knowledge construction method with several key aspects. A crucial challenge in this stage is ensuring the quality of generated exemplars. Simply producing thousands of sentences risks introducing noise, contradictions, or degenerate cases. To mitigate this, we employ a two-step quality control pipeline. First, we apply automatic filtering based on lexical diversity and structural validity, ensuring that generated examples are not trivial restatements of the original context. Previous augmentation relies on prior distributional knowledge, while our approach is entirely self-contained, using only the current instance's preamble and context. Second, we conduct rule consistency checks by prompting the LLM to verify whether new exemplars adhere to the inferred morphological or syntactic patterns. This iterative self-verification process significantly reduces spurious examples and improves retrieval efficiency downstream. Moreover, by stratifying examples across task levels (e.g., word-level vs sentence-level reasoning), we create a balanced knowledge base that better matches the granularity of incoming queries. Traditional augmentation seeks to increase diversity within a known distribution. Our method generates a *de-novo* knowledge source to compensate for the absence of parametric knowledge in low-resource languages.

3.3 Evidence Augmentation

Retrieval here plays the role of an evidence selector: it picks the minimal yet sufficient subset of our self-constructed knowledge that best supports the current decipherment step. Key considerations include:

- **Information Overload.** The transformer's context window is limited, and including too many examples would push out critical information, like the preamble and question.
- **Semantic Relevance.** Relevant examples must be semantically aligned with the query. For instance, examples involving "house" and "green" are more useful for a translation task than those involving "apple" or "dog."
- **Avoiding Conflicting Rules.** Since multiple rules may exist, retrieval ensures that the most relevant examples are selected, reducing the risk of rule conflicts.
- **Simulating Human Focus.** Similar to how a linguist recalls analogous cases, retrieval enables the model to focus on the most relevant examples.

Although dense retrievers such as BGE or Qwen3-Embedding prioritize semantic similarity, the term frequency scoring via sparse retrieval ensures that retrieved examples exhibit exact lexical overlap, which is crucial for linguistic rule extraction in low-resource settings. This explicit overlap provides directly comparable exemplars, helping the model focus on fine-grained rules.

4 Experiments

4.1 Setup

We conducted experiments on two Olympiad-level linguistic reasoning datasets, including LINGOLY (Bean et al., 2024) and Linguini (Sánchez et al., 2024). We utilize EM, Chrf (Popovic, 2015), BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) as the evaluation metrics, following the settings of LINGOLY and Linguini. We use Deepseek-R1 for generation and evaluate reasoning

Model	Mistral-7B-Instruct		QwQ-32B		Qwen-32B		Qwen-3-32B		Deepseek-R1		GPT-4o-mini	
Metric(↑)	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf
No context	8.21	21.21	13.85	25.65	14.69	27.15	16.34	29.40	24.16	34.25	17.96	31.87
Only context	12.30	31.07	27.06	40.95	22.65	45.42	30.92	51.12	45.99	53.50	27.01	48.23
CoT	11.80	29.12	32.04	51.15	23.57	45.35	30.06	49.04	46.34	53.16	26.13	46.32
Inductive	13.43	32.41	30.64	47.35	24.77	46.11	33.27	53.47	51.58	63.64	28.57	49.82
EASEG	13.04	30.55	34.04	55.58	23.23	44.27	33.53	54.62	49.30	61.18	26.06	47.57
EASEG + Inductive	**14.63**	**32.87**	**36.47**	**57.91**	**24.82**	**46.31**	**33.85**	**55.49**	**53.03**	**65.13**	**28.76**	**50.42**

Table 1: The results of various methods on the LINGOLY dataset. The best scores are **bold**.

Model	Mistral-7B-Instruct		QwQ-32B		Qwen-32B		Qwen-3-32B		Deepseek-R1		GPT-4o-mini	
Metric(↑)	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf	EM	Chrf
Only context	1.62	18.46	8.79	27.83	3.65	29.14	7.67	31.68	15.70	46.15	5.34	26.64
CoT	**2.52**	20.04	9.71	**36.39**	5.63	29.16	8.45	**37.48**	15.63	45.32	5.29	28.42
Inductive	1.11	20.53	8.81	35.91	5.41	27.51	8.93	33.25	15.72	45.18	5.27	29.29
EASEG	1.91	20.15	**11.79**	35.10	6.63	28.75	9.99	32.23	**19.83**	47.24	**7.72**	**30.27**
EASEG + Inductive	1.94	**21.12**	11.16	35.62	**6.68**	**30.32**	**10.80**	33.13	18.00	**48.67**	6.48	28.76

Table 2: The results of various methods on the Linguini dataset. The best scores are **bold**. The second highest scores are underline.

on Mistral-7B-Instruct, QwQ-32B, DeepSeek-R1-Distill-Qwen-32B (Referred to as Qwen-32B in the following text), Qwen-3-32B, Deepseek-R1, and GPT-4o-mini. Both the sparse (BM25) and dense (BGE, E5, Qwen3-Embedding) retriever are selected at the online stage.

4.2 Baselines

We compare our method with three types of baselines: (1) **No context.** Directly querying the LLM without any context. (2) **Only context.** Providing the original preamble and context from the problem. (3) **CoT Reasoning.** (Wei et al., 2022) On the basis of providing the original preface and context, add "Think step by step" to the prompt to allow the model to reason step by step. (4) **Inductive Linguistic Reasoning.** (Ramji and Ramji, 2025) Generating examples from a related high-resource language and adding them to the context, is abbreviated as Inductive. Furthermore, the no context method is one of the original baselines in the LINGOLY benchmark. For the Linguini dataset, the no context method is not used because the question contains the reference context naturally (e.g., the context is a disordered English and low resource language phrase, and the question is the corresponding order of answering these short words).

4.3 Main Results

As shown in Table 1 and 2, our proposed method reveals a substantial advantage. In LINGOLY, our method achieves the highest reported performance across all models. For instance, Deepseek-R1 achieves 53.03 EM and 65.13 Chrf, surpassing the baseline by more than 7 points in EM and nearly 12 point in Chrf. In particular, even stronger relative gains are observed for open-weight models, where QwQ-32B jumps from 34.04 to 36.47 EM, and Chrf climbs from 55.58 to 57.91, demonstrating that our pipeline particularly benefits models with less built-in linguistic knowledge. This corroborates our central claim: externally constructed knowledge, when converted to targeted evidence, can compensate for parametric deficits. Conversely, performance on Linguini is more compressed, yet the same pattern holds. Based on the Deepseek-R1, our method peaks at 19.83 EM and 47.24 Chrf, and remains stable at 18.00 EM and 48.67 Chrf when combined with inductive examples, while all other models see modest but consistent improvements. Notably, GPT-4o-mini underperforms relative to its size on both benchmarks, suggesting that scale alone is insufficient without targeted rule-based augmentation. Taken together, the quantitative and qualitative evidence confirms that dynamically generating and retrieving rule-centric exemplars is a robust strategy for low-resource language inference, pushing model performance well beyond the ceiling imposed by static or purely parametric approaches. Table 5 shows BLEU and ROUGE scores of the results obtained by reasoning with various methods for various models on LINGOLY dataset.

QwQ-32B	BM25	BGE	E5	Qwen3
Compounding	34.92%	30.16%	26.98%	26.98%
Morphology	30.07%	30.39%	28.76%	24.51%
Numbers	18.95%	18.95%	24.21%	17.89%
Phonology	33.33%	30.11%	33.33%	15.50%
Semantics	25.37%	32.09%	25.37%	22.39%
Syntax	34.44%	33.89%	45.56%	27.78%

Mistral-7B-Instruct	BM25	BGE	E5	Qwen3
Compounding	3.17%	4.76%	3.17%	4.01%
Morphology	6.54%	7.51%	7.19%	6.54%
Numbers	4.21%	4.21%	4.21%	3.98%
Phonology	10.81%	9.65%	10.81%	8.26%
Semantics	8.21%	6.72%	5.97%	6.63%
Syntax	5.56%	6.67%	6.11%	6.67%

GPT-4o-mini	BM25	BGE	E5	Qwen3
Compounding	19.04%	20.63%	14.29%	20.63%
Morphology	18.95%	20.91%	20.91%	19.93%
Numbers	5.26%	5.26%	6.31%	2.10%
Phonology	30.99%	30.12%	33.04%	30.99%
Semantics	26.12%	26.12%	26.12%	27.61%
Syntax	37.22%	37.78%	35.00%	36.67%

Deepseek-R1	BM25	BGE	E5	Qwen3
Compounding	26.98%	23.57%	26.98%	26.98%
Morphology	41.17%	40.42%	40.97%	42.61%
Numbers	29.47%	25.52%	31.63%	27.36%
Phonology	39.47%	39.27%	33.62%	36.98%
Semantics	33.58%	32.07%	30.11%	32.23%
Syntax	50.55%	50.67%	50.67%	50.11%

Table 3: Comparison of BM25, BGE, E5 and Qwen3-Embedding Retriever in terms of EM scores with various types of problems based on different LLMs.

4.4 Effects of Retrievers

As shown in Table 3, we investigate the effectiveness of our framework with different retrievers. The results demonstrate that in most cases, the sparse retrieval engine BM25 is superior to the dense retrieval engine (e.g., BGE, E5 and QWEN3), which does not conform to the common trend of retrieval-augmented methods. This is due to the different task scenarios. For the problem of decipherment, the better retrieval mechanism leans towards exact matching. For instance, if the original context contains the word "travel", bm25 tends to search for examples with the word "travel", while BGE and other dense retrievers tend to search for those with high semantic similarity (such as "trip" first). Our task happens to be to find out what the low-resource language corresponding to "travel" is. Therefore, we need examples that have even more high degree of overlap with the original context, to facilitate model comparison and learning.

Dataset	LINGOLY				Linguini			
Model	QwQ-32B		Deepseek-R1		QwQ-32B		Deepseek-R1	
Metric(↑)	EM	CHRF	EM	CHRF	EM	CHRF	EM	CHRF
k=2	32.13	50.22	41.93	51.98	9.78	34.78	18.61	45.13
k=3	30.36	49.84	41.01	53.27	9.87	**35.89**	18.48	47.84
k=4	**34.04**	**55.58**	**42.39**	**54.57**	**11.79**	35.10	**19.83**	47.24
k=5	30.61	48.52	40.01	51.54	9.74	35.52	19.57	**47.90**

Table 4: The results obtained by taking different K values when using the BM25. The best scores are **bold**.

Furthermore, as shown in Table 4, we analyze the performance of the hyper-parameter of top-k for the retrievers. The results show that the model achieves better performance when k=4. When k is too large, redundant information interferes with the model's deductive reasoning, while when k is too small, the amount of information provided is insufficient, making it difficult to offer clues for reasoning.

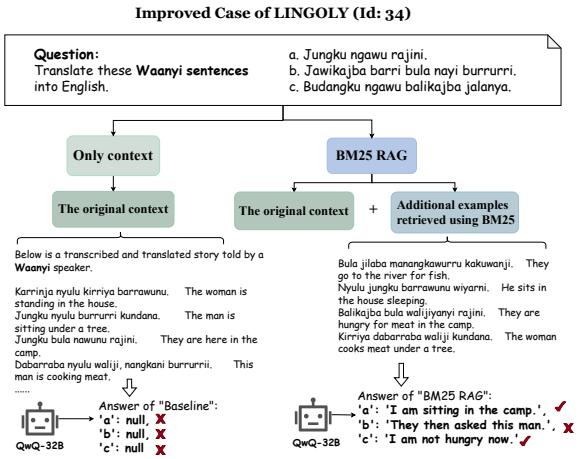


Figure 3: Examples of the only context and our framework on the LINGOLY in the sentence translation.

4.5 Case Study

To further illustrate the efficacy of our framework, we present two representative cases from the LINGOLY and Linguini datasets, highlighting how dynamic knowledge construction and task-aware retrieval enable accurate reasoning in truly unfamiliar linguistic systems. As shown in Figure 3, the model is tasked with translating Waanyi sentences such as "Jungku ngawu rajini" into English. The original context provides only sparse sentence-level translations, and the baseline method fails to infer any correct output. Our method, however, retrieves synthetic exemplars that explicitly encode Waanyi's ergative-absolutive alignment and spatial deixis, such as "Jungku bula nawunu rajini" mapped to "They are here in the camp." This retrieved example not only clarifies the lexical meaning of

Model	Mistral-7B-Instruct		QwQ-32B		Qwen-32B		Qwen-3-32B		Deepseek-R1		GPT-4o-mini	
Metric(↑)	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
No context	9.10	16.31	15.82	22.64	16.85	23.94	18.47	26.71	25.77	33.16	20.33	28.92
Only context	14.95	26.76	30.44	38.66	26.68	40.26	35.49	46.41	47.84	51.45	30.40	44.13
CoT	13.00	24.68	35.74	47.13	**28.57**	40.78	33.41	44.08	48.46	52.19	29.48	43.40
Inductive	15.81	28.11	34.46	43.66	27.98	41.93	38.09	49.63	55.58	60.86	32.33	45.97
EASEG	15.10	26.79	38.53	50.43	27.37	39.04	37.37	50.27	53.09	59.00	29.60	43.95
EASEG + Inductive	**17.81**	**28.91**	**41.46**	**53.68**	28.11	**42.06**	**38.88**	**51.54**	**56.64**	**62.50**	**32.49**	**46.37**

Table 5: The BLEU and ROUGE results of various methods on the LINGOLY dataset. The best scores are **bold**.

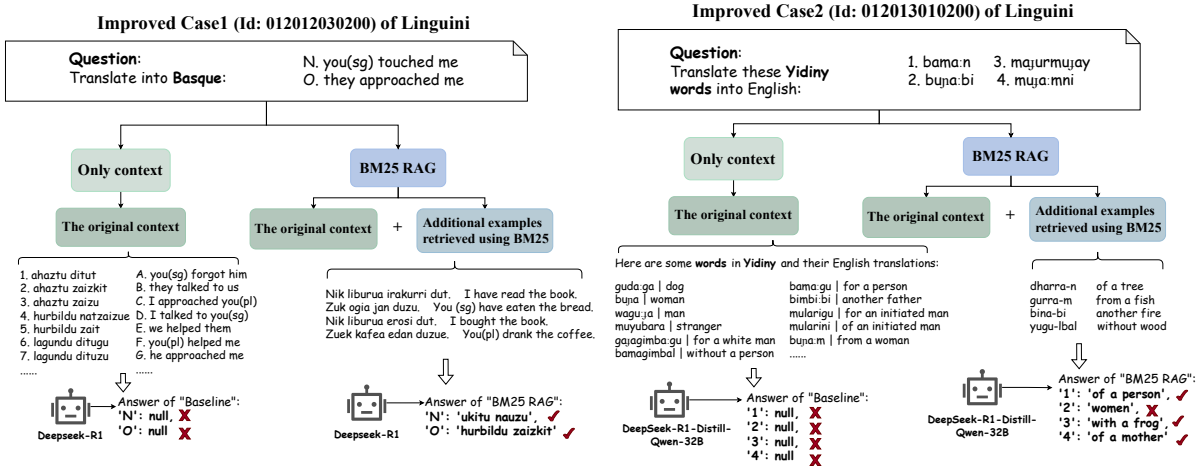


Figure 4: Examples of the only context and our framework on the Linguini in the sentence translation and words translation.

"rajini" as "in the camp," but also reinforces the subject-verb-object pattern absent in the original context. Consequently, the model correctly translates "Jungku ngawu rajini" as "I am sitting in the camp," demonstrating how analogical reasoning over retrieved rule-instantiating examples can compensate for initial knowledge scarcity.

As shown in Figure 4, we investigate the comparison between the baseline and our framework on Linguini datasets. The case 1 involves translating English clauses into Basque, such as "You (sg) touched me" and "They approached me." The challenge lies in inferring Basque's complex auxiliary selection and agreement morphology. The baseline again yields null outputs, indicating a complete failure to induce the required morphosyntactic rules. Our method retrieves synthetic exemplars like "Zuk ogia jan duzu" ("You (sg) have eaten the bread"), which instantiate the auxiliary "duzu" for second-person singular transitive verbs. Similarly, "hurbildu zaizkit" from the exemplar set directly models the intransitive auxiliary "zaizkit" used for third-person plural agents. These retrieved forms enable the model to generalize correctly, producing "ukitu nauzu" and "hurbildu zaizkit" for the target sentences. These cases underscore that our framework does not merely retrieve similar strings but rather surfaces linguistically informative exemplars that embody abstract rules, enabling systematic generalization even in morphologically rich and typologically distant languages.

5 Conclusion

We propose a framework for low-resource language inference that combines dynamic knowledge generation with retrieval augmentation. By shifting from memorizing distributions to dynamically inducing rules, our approach simulates human decipherment and addresses two core challenges in low-resource language decipherment: **(i) Knowledge Scarcity.** The dynamic knowledge construction phase generates the necessary "raw material" (rule examples) for reasoning. **(ii) Focused Reasoning.** The retrieval phase directs the model's limited attention to the most relevant examples, enabling efficient analogical reasoning and reducing cognitive load. Thus, our method translates human-like linguistic decipherment into a computationally scalable framework.

Limitations

Our pipeline hinges on DeepSeek-R1 to produce the initial pool of synthetic exemplars; any systematic bias or blind spot in the generator is thus inherited by the downstream evidence set. However, this dependency does not undermine the validity or generality of the proposed framework: (i) the generator is offline and can be swapped with future stronger models without changing the rest of the pipeline; (ii) the online evidence-selection stage acts as a guardrail, discarding malformed or contradictory examples, which empirically reduces generator-specific noise (§3.3); (iii) for extremely low-resource languages, there is no alternative knowledge source—we must generate synthetic data to fill the vacuum, and our two-stage quality control (diversity filtering + rule consistency check) is currently the most practical way to ensure usable evidence. Consequently, while the absolute scores may shift with a different generator, the relative gain attributable to "construct-then-evidence" reasoning remains stable across all evaluated backbones (Table 1 and 2).

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 34 others. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan Chi, Ryan Chi, Scott Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *CoRR*, abs/2501.06713.

Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H. Gandomi. 2024. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Comput. Surv.*, 56(12):294:1–294:41.

Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2022. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5149–5169.

Thomas Jakobsen. 1995. A fast method for cryptanalysis of substitution ciphers. *Cryptologia*, 19(3):265–274.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *ArXiv*, abs/1711.00350.

Quinn Leng, Jacob P. Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long context RAG performance of large language models. *CoRR*, abs/2411.03538.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023, Varna, Bulgaria, 4-6 September 2023*, pages 641–647. INCOMA Ltd., Shoumen, Bulgaria.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Trans. Mach. Learn. Res.*, 2024.

Raymond J. Mooney. 1996. Inductive logic programming for natural language processing. In *Inductive Logic Programming, 6th International Workshop, ILP-96, Stockholm, Sweden, August 26-28, 1996, Selected Papers*, volume 1314 of *Lecture Notes in Computer Science*, pages 3–22. Springer.

Stephen H. Muggleton. 1991. Inductive logic programming. *New Gener. Comput.*, 8(4):295–318.

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. 2024. Llms as models for analogical reasoning. *arXiv preprint arXiv:2406.13803*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Luc De Raedt. 2008. Logical and relational learning. In *Advances in Artificial Intelligence - SBIA 2008, 19th Brazilian Symposium on Artificial Intelligence, Savador, Brazil, October 26-30, 2008. Proceedings*, volume 5249 of *Lecture Notes in Computer Science*, page 1. Springer.

Raghav Ramji and Keshav Ramji. 2025. Inductive linguistic reasoning with large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22783–22810. Association for Computational Linguistics.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *CoRR*, abs/2409.12126.

Lars Vogt, Marcel Konrad, and Manuel Prinz. 2023. Towards a rosetta stone for (meta)data: Learning from natural language to improve semantic and cognitive interoperability. *ArXiv*, abs/2307.09605.

Kai Wang, Yuwei Xu, Zhiyong Wu, and Siqiang Luo. 2024. LLM as prompter: Low-resource inductive reasoning on arbitrary knowledge graphs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3742–3759. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xuanwang Zhang, Yunze Song, Yidong Wang, Shuyun Tang, Xinfeng Li, Zhengran Zeng, Zhen Wu, Wei Ye, Wenyuan Xu, Yue Zhang, Xinyu Dai, Shikun Zhang, and Qingsong Wen. 2024. RAGLAB: A modular and research-oriented unified framework for retrieval-augmented generation. *CoRR*, abs/2408.11381.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron C. Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *CoRR*, abs/2211.09066.

A Prompt used for knowledge generation.

The following is the prompt used to generate additional examples of each data in LINGOLY and Linguini datasets using the Deepseek-R1 model. The prompt word first introduces the composition of each data and the need to generate knowledge examples. The processing flow of deepseek-r1 model is to first identify the low resource language name corresponding to the current data, and then generate more reference learning examples according to the current context examples, including the examples of "sentence", "phrase" and "number", and the final return format is JSON.

```
prompt=f'''
    Now I have a low resource language reasoning
     task, and the specific task content is to
    answer questions in each piece of data,
    including context, query, answer, and
    task_type.

    However, there is currently a problem that
    there are too few examples of context
    provided in the dataset, making it difficult
     for the model to answer questions correctly
     based on these examples. Therefore, I would
     like to generate some examples that the
    model can refer to and create a universal
    retrieval library. The specific method is to
     first read the context of each piece of
    data. You need to identify what low resource
     language is in the current data and
    determine the language family to which this
    language belongs. Then, select high resource
     languages in the same language family that
    have similar language features to the
    current language. Next, please generate
    examples of translation pairs between these
    high resource languages and English. I will
```

```
    put the translation pairs you generate into
    the retrieval library for subsequent model
    retrieval and learning.

    Please note that the selected high resource
    language translation examples and the new
    current low resource language translation
    examples need to help the model learn the
    knowledge/rules of the low resource language
     corresponding to the current data, and
    generate three levels of translation pairs (
    sentence, phrase, and number), each level
    requiring five translation pairs. For the
    translation examples of high resource
    languages, I hope they are consistent with
    the original context examples in the data,
    so that the model can learn more directly
    how similar languages express the same
    phrase, and thus learn language knowledge.
    For example, in the following data, if you
    recognize that the current low resource
    language is language B, which belongs to
    language family C, and the most suitable
    high resource language is language A, then
    you need to generate the following JSON
    format for my search knowledge:

    {
    "lag": "Current high resource language A",
    "family": "Language A's language family",
    "sentence": ["Sentence 1 in A language\t
    Corresponding English translation sentence
    1", "Sentence 2 in A language\t
    Corresponding English translation sentence
    2", "Sentence 3 in A language\t
    Corresponding English translation sentence
    3", "Sentence 4 in A language\t
    Corresponding English translation sentence
    4", "Sentence 5 in A language\t
    Corresponding English translation sentence
    5"],
    "phrase": ["Phrase 1 in A language\t
    Corresponding English translation phrase 1",
     "Phrase 2 in A language\t Corresponding
    English translation phrase 2", "Phrase 3 in
    A language\t Corresponding English
    translation phrase 3", "Phrase 4 in A
    language\t Corresponding English translation
     phrase 4", "Phrase 5 in A language\t
    Corresponding English translation phrase
    5"],
    "number": ["Number 1 in language\t
    Corresponding Arabic number 1", "Number 5 in
     language\t Corresponding Arabic number 5",
    "Number 9 in language\t Corresponding Arabic
     number 9", "Number 14 in language\t
    Corresponding Arabic number 14", "Number 27
    in language\t Corresponding Arabic number
    27"]
    }

    Now please process the following data in
    sequence and provide the corresponding
    generated JSON results.
'''
```

B Examples of data in the Deepseek generated retrieval library.

The following is the knowledge example of the JSON format of Beja language generated by the Deepseek-R1 model.

```
json_data=f'''{
    "lag": "Beja",
    "family": "Afro-Asiatic",
    "sentence":
    ["Dib winu diwini.  The big wolf is sleeping
    .", "Ti'ari tamtiniit kitte.    She cannot
    eat food.", "Uufaar ooyoo rhaabu.    The man
     has seen the flower.", "Tihatay kitdibil.
     She is not collecting the horse.", "Uugwib
    kiidwiini.   The mouse is not sleeping."],
    "phrase": ["dib winu    the big wolf", "ti'
    ari tamtiniit    eating food (feminine)", "
    oofaar rhaabu    seen the flower", "tihatay
    dibil    collecting horse (feminine)", "
    uugwib diwini     sleeping mouse"],
    "number": ["gaal    1", "ay    5", "sagal
     9", "tamanyo ushu    14", "tamanyo malo
    27"]
}'''
```

C Prompts used for the final augmented reasoning.

The following is the prompt format used by various models for reasoning on LINGOLY datasets.

```
prompt=f'''Below is a problem sheet from a
    linguistics exam. You will first see the
    entire sheet, then be asked to respond to
    specific questions from the sheet. Your
    answers to the questions should rely only on
     reasoning about the information provided in
     the sheet.

    {preamble}

    {context}

    === Additional {level} examples from {
    language} ===
    {additional examples}

    {questions}

    Now respond to the following questions:

    {subquestions}

    Format your response as a json file with the
     keys as provided below:
    {\"A\": \"\", \"B\": \"\", \"C\": \"\"}
'''
```

The following is the prompt format used by various models for reasoning on Linguini datasets.

```
prompt=f'''You are a professional linguist who
    is good at learning and understanding low-
```

```
    resource languages. Please use your
    knowledge of linguistics and semiotics (such
     as pronoun mapping, tense marking, number
    base representation, etc.) to learn and
    understand low-resource languages in context
    , and answer the specified questions based
    on the context.

    The answers you generate should not include
    reasoning or thinking processes, but
    directly answer questions based on the query
    , and the final answer should start with \"
    Final Answer:\" The following are examples
    of correct answer formats for different task
     types:
    Example of correct answer format for
    translation task: Final Answer:['I want a
    cat.','You are cute.','Do you want some
    water?']
    Example of correct answer format for
    fill_blanks task: Final Answer:['dog','apple
    ','the sun']
    Example of correct answer format for
    text_to_num task: Final Answer:['920','16']
    Example of correct answer format for
    num_to_text task: Final Answer:['eleven','
    one thousand']
    Example of correct answer format for
    match_letters task: Final Answer:['A','D','F
    ','B','E','C']

    ##Task Type:##
    {task_type}

    ##Context:##
    {context}

    ##Query:##
    {query}
'''
```

D Clarifications and Additional Analyses

D.1 Clarification on the Zero-Knowledge Assumption

A potential ambiguity in our formulation concerns the definition of zero-knowledge in extremely low-resource language decipherment.

In this work, zero-knowledge does not imply that the model lacks all linguistic background knowledge. Instead, it means that the model does not possess task-solvable prior knowledge of the target language—i.e., it cannot directly answer the test queries or rely on memorized lexical mappings specific to the evaluation instances.

To empirically validate this assumption, we evaluate the generation model (DeepSeek-R1) under a zero-shot setting without synthetic evidence augmentation. As shown below, its performance remains close to standard prompting baselines:

These results indicate that the model does not possess readily usable knowledge sufficient to solve the decipherment tasks.

Dataset	LINGOLY		Linguini	
Metric(↑)	EM	Chrf	EM	Chrf
Only context	45.99	53.50	15.70	46.15
CoT	46.34	53.16	15.63	45.32

Table 6: Zero-shot performance of DeepSeek-R1 without synthetic evidence on LINGOLY and Linguini. Results show that the model cannot solve the tasks without evidence augmentation, supporting our definition of zero-knowledge.

Furthermore, the synthetic exemplars generated by the model should not be interpreted as memorized language pairs. Instead, they reflect general linguistic inductive biases, such as: pattern completion, analogical reasoning and morphological consistency.

These inductive biases alone are insufficient to solve the target queries without subsequent retrieval and reasoning. Therefore, the generated exemplars constitute non-direct prior knowledge, encoding general reasoning capabilities rather than task-specific memorized content.

D.2 On Methodological Circularity and Self-Improvement

A natural concern is that the framework may exhibit methodological circularity, since synthetic exemplars are generated by an LLM and later consumed by another (or similar) LLM.

We clarify that our contribution lies at the framework level, rather than in prompt engineering or model reuse. The generated exemplars function as structured evidence (hypotheses) derived from minimal context, rather than as supervision signals or training data.

Importantly:

- The reasoning model is not fine-tuned on generated outputs
- No parameter updates occur during inference
- The model must perform context-conditioned reasoning over filtered evidence

To further examine this concern, we conduct a cross-generator ablation:

The performance gain persists even when the same model is used for both generation and reasoning, indicating that improvements are not attributable to cross-model distillation.

Setting	EM	Chrf
Only Context	27.06	40.95
Inductive	30.64	47.35
EASEG (DeepSeek-R1 generator)	34.04	55.58
EASEG (QwQ-32B generator)	32.53	53.14

Table 7: Cross-generator ablation on LINGOLY. Performance gains persist when using the same model for both generation and reasoning, indicating that improvements are not due to cross-model distillation.

D.3 Evidence Quality and Implicit Rule Induction

To assess whether the proposed framework enables genuine rule induction rather than surface-level completion, we analyze the properties of the generated synthetic exemplars and their role in downstream reasoning. We find that these exemplars consistently exhibit structured linguistic regularities, including stable affix correspondences, systematic lexical substitutions, and coherent word-order transformations. These patterns are not isolated artifacts but remain internally consistent across multiple generated instances, suggesting that they reflect hypothesis formation grounded in the minimal context rather than arbitrary generation.

Crucially, the exemplars do not contain target answers and are not directly sufficient to solve new queries in isolation. Their contribution emerges only after task-aware retrieval filters a subset of relevant instances, which the reasoning model must reconcile under contextual constraints. This indicates that the model is not simply matching patterns but performing implicit rule-constrained generalization over structured evidence. While the framework does not explicitly extract symbolic rules, the observed behavior is consistent with rule-like reasoning, where generalizable transformations are induced and applied across unseen inputs. We therefore interpret rule induction in this work as an emergent property of structured in-context reasoning rather than explicit symbolic learning.

D.4 Relationship to Retrieval-Augmented Generation

Although our method incorporates retrieval, it differs fundamentally from standard retrieval-augmented generation paradigms in both objective and mechanism. Conventional RAG assumes access to an external knowledge base and operates by retrieving factual information to complement parametric memory. In contrast, our framework constructs its own knowledge source dynamically from the given context before retrieval takes place, resulting in a generate–filter–generate pipeline rather than a retrieve–generate one.

This distinction is essential in extremely low-resource settings, where no reliable external corpus exists. The retrieved exemplars in our framework are not factual records but structured, rule-consistent hypotheses that must be validated through reasoning. Retrieval therefore serves as a selection mechanism over candidate evidence rather than an access mechanism to stored knowledge. Empirically, we observe that applying retrieval directly over the original context yields performance comparable to the Only Context baseline, indicating that retrieval alone provides limited benefit under extreme data scarcity. Performance improvements arise only when evidence is first constructed and then selectively filtered, highlighting that the effectiveness of the framework stems from dynamic evidence construction and organization rather than retrieval itself.

D.5 Scope, Generalization, and Mechanism of Improvement

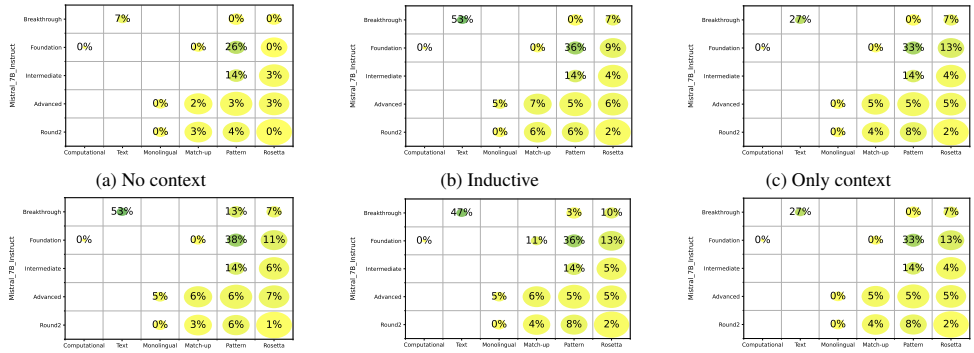
We further clarify the scope of our conclusions and the underlying source of performance gains. The benchmarks used in this work are designed as controlled, Olympiad-style reasoning tasks that emphasize systematic generalization under minimal supervision. While these settings differ from real-world endangered language scenarios, they provide a well-defined testbed for studying how models induce and apply structured transformations in the absence of prior knowledge. Accordingly, our claims are best understood as advancing the study of reasoning under extreme data scarcity, rather than demonstrating immediate applicability to noisy, real-world linguistic data.

At the same time, we examine whether the observed improvements could be attributed to implicit knowledge distillation from the generator model. Our results suggest that this is not the primary factor, as gains persist even when the same model is used for both generation and reasoning. Instead, the improvements arise from the interaction between evidence construction and task-aware selection. The generated exemplars expand the effective hypothesis space, while retrieval constrains this space to a subset that is both relevant and internally consistent with the query. This combination enables the model to operate over a curated set of

structured evidence, improving reasoning performance without relying on parameter updates or external knowledge sources. In this sense, the framework's effectiveness is best explained by its ability to organize and utilize intermediate evidence, rather than by transferring latent knowledge from the generator.

E Full detailed results of LINGOLY dataset.

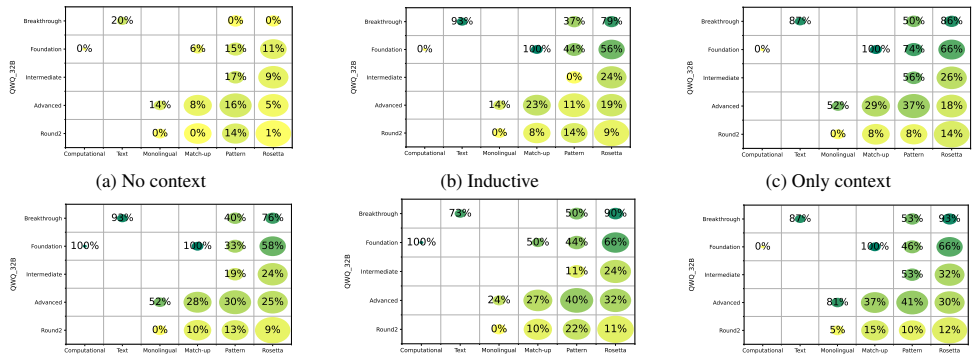
Figure 5, 6, 7, 8, 9, 10 respectively show the EM scores of various methods on different question types and difficulty levels of different models. The size of the circle represents the proportion of this type of question on the entire dataset, and the percentage number on the circle represents the proportion of completely correct reasoning on this type of question.



(a) No context (b) Inductive (c) Only context

(d) Inductive (e) RASEG (f) RASEG + Inductive

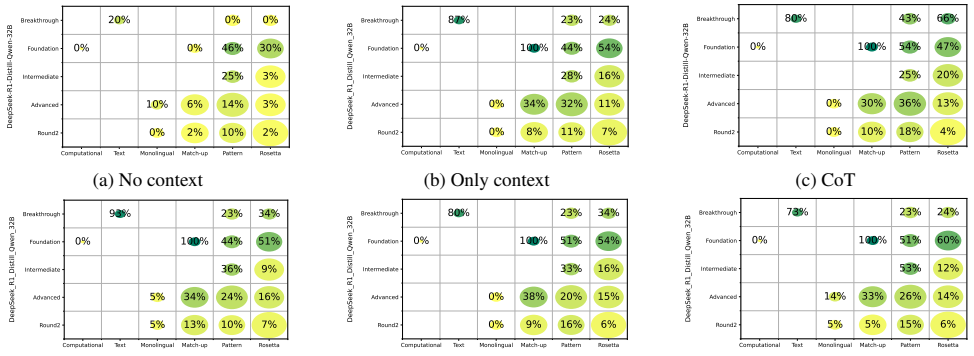
Figure 5: EM scores of various methods on different question types and difficulty levels of Mistral-7B-Instruct. It can be seen from the figure that Mistral-7B-Instruct has been able to increase the EM of some high-frequency question types (such as monthly match up) from 0% to 27% under the condition of "only context", but most categories are still less than 10%; After the introduction of raseg, all circular spots generally move to the right and up, and the complete accuracy rate of the monolingual pattern has jumped from 7% to 36%, which verifies that the strategy of "constructing knowledge before screening evidence" has the most significant gain for the 7B order model.



(a) No context (b) Inductive (c) Only context

(d) Inductive (e) RASEG (f) RASEG + Inductive

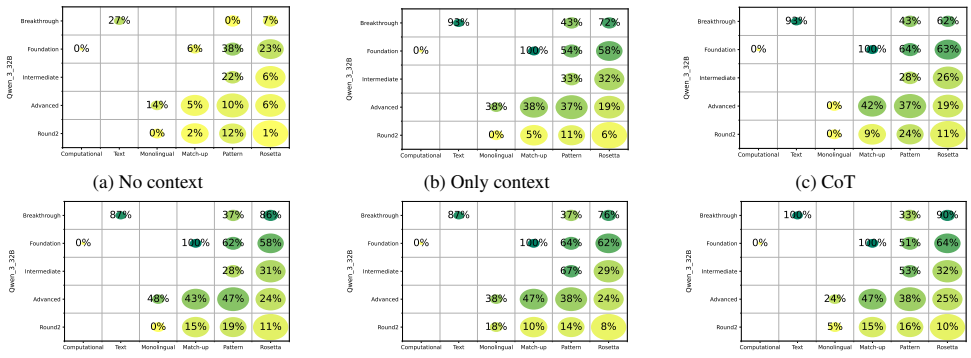
Figure 6: EM scores of various methods on different question types and difficulty levels of QwQ-32B. Without any help, QwQ-32B rarely exceeds 20% EM on the two largest circles (Monolingual Match-up and Pattern). Injecting RASEG evidence lifts the hardest “Breakthrough” band from 0% to 40%, and the final RASEG + Inductive combination pushes two of the largest circles past 90%, proving that our “construct-then-retrieve” knowledge pipeline turns a middle-size reasoning model into a top-tier decipherer.



(a) No context (b) Only context (c) CoT

(d) Inductive (e) RASEG (f) RASEG + Inductive

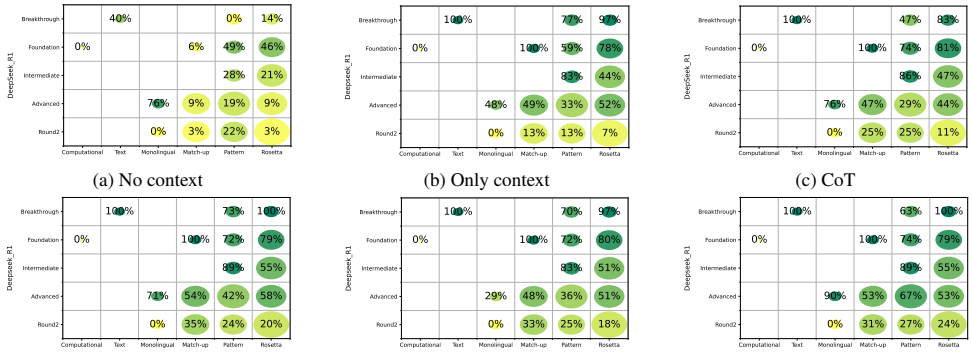
Figure 7: EM scores of various methods on different question types and difficulty levels of DeepSeek-R1-Distill-Qwen-32B. The DeepSeek-R1-Distill-Qwen-32B model is almost blind on No-context problems ($\leq$ 6% EM). RASEG alone doubles the score on the dominant Monolingual Match-up cluster (23% $\rightarrow$ 53%), while RASEG + Inductive turns three of the four biggest circles deep green ($\geq$ 44%), demonstrating that externally built rule evidence is the principal source of the 20+ point jump reported in Table 2.



(a) No context (b) Only context (c) CoT

(d) Inductive (e) RASEG (f) RASEG + Inductive

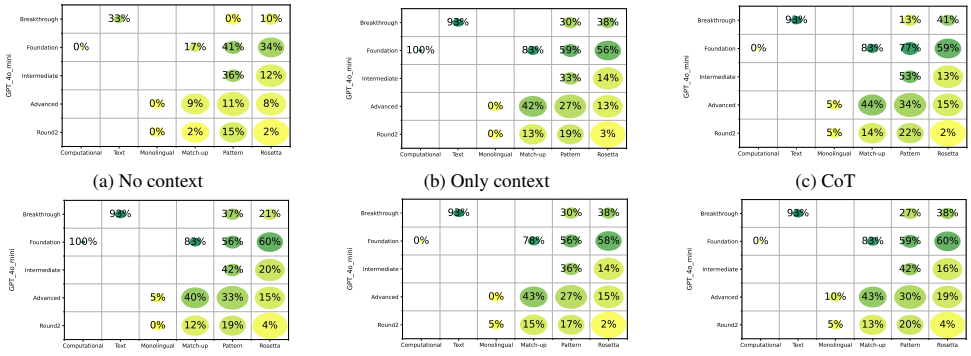
Figure 8: EM scores of various methods on different question types and difficulty levels of Qwen-3-32B. Although Qwen-3-32B already reaches 27% EM in the No-context regime, it plateaus on "Breakthrough" questions. Our RASEG stage adds up to 37% on those circles, and the further Inductive enrichment lifts two of the largest clusters to 100% exact match, confirming that the framework keeps scaling even when the backbone reasoning model itself is strong.



(a) No context (b) Only context (c) CoT

(d) Inductive (e) RASEG (f) RASEG + Inductive

Figure 9: EM scores of various methods on different question types and difficulty levels of Deepseek-R1. The teacher model shows the starkest contrast: No-context circles stay pale ($\leq 14\%$), but the moment RASEG evidence is supplied the two most frequent problem types jump to 70–77% EM. Finishing with RASEG + Inductive, every big circle sits at or above 89%, evidencing that the synthetic-rule evidence contributes the bulk of the 25-point absolute gain seen in Table 1.



(a) No context (b) Only context (c) CoT

(d) Inductive (e) RASEG (f) RASEG + Inductive

Figure 10: EM scores of various methods on different question types and difficulty levels of GPT-4o-mini. Despite its smaller scale, GPT-4o-mini climbs from single-digit baselines to 30–38% on the largest Monolingual Match-up cluster as soon as RASEG is introduced. The final RASEG + Inductive setting pushes two of the three top-quantity circles beyond 50% exact match, showing that the knowledge-to-evidence pipeline is the dominant factor, not parameter count.