BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning

Anonymous ACL submission

Abstract

This paper introduces BMIKE-53, a comprehensive benchmark for cross-lingual in-context knowledge editing (IKE) across 53 languages, unifying three knowledge editing (KE) datasets: zsRE, CounterFact, and WikiFactDiff. Cross-lingual KE, which requires knowledge edited in one language to generalize across others while preserving unrelated knowledge, remains underexplored. To address this gap, we systematically evaluate IKE under zero-shot, one-shot, and few-shot setups, incorporating tailored metric-specific demonstrations. Our findings reveal that model scale and demonstration alignment critically govern cross-lingual IKE efficacy, with larger models and tailored demonstrations significantly improving performance. Linguistic properties, particularly script type, strongly influence performance variation across languages, with non-Latin languages underperforming due to issues like language confusion.

1 Introduction

Large language models (LLMs) have demonstrated remarkable abilities to encode vast amounts of knowledge during pre-training, enabling them to perform well across a range of tasks (Min et al., 2022; Zhang et al., 2023; Zhou et al., 2024). However, this knowledge remains static, becoming outdated as the world evolves, necessitating mechanisms to update models with new facts while preserving their overall performance (Cao et al., 2021; Dhingra et al., 2022). Traditional approaches, such as fine-tuning, are computationally expensive and impractical for closed-source or large-scale models (Dai et al., 2022). These limitations have motivated the emergence of knowledge editing (KE)—a technique for selectively modifying LLMs to incorporate new knowledge while maintaining the integrity of unrelated knowledge (Zhang et al., 2024b).

Recent advancements in KE have explored gradient-free methods inspired by in-context learning (ICL), where LLMs learn through prompts and demonstrations without requiring parameter updates (Zheng et al., 2023). These methods are efficient and particularly suitable for scenarios where direct access to model parameters is restricted. However, existing gradient-free KE research primarily focuses on monolingual settings, leaving the potential for cross-lingual KE largely unexplored. Cross-lingual KE, a more challenging task, as illustrated in Figure 1, requires knowledge edited in one language (e.g., English) to generalize effectively to semantically equivalent queries across diverse target languages while preserving unrelated knowledge.

Figure 1: Examples of cross-lingual in-context knowledge editing.

This study addresses the critical gap in cross-lingual knowledge editing by proposing an comprehensive and multidimensional investigation of in-context knowledge editing (IKE) methods. We introduce **BMIKE-53**, a multilingual benchmark spanning 53 languages and integrating three representative KE datasets: zsRE, which evaluates regular fact modifications; CounterFact, which examines counterfactual knowledge updates; and WikiFactDiff (WFD), which assesses real-world, temporally dynamic knowledge updates. This benchmark is the most comprehensive multilingual KE resource to date, unifying diverse KE datasets into a consistent format and expanding them into multiple languages using LLM-assisted translation. The wide linguistic coverage allows us to systematically analyze cross-lingual differences and their underlying causes.

To evaluate cross-lingual IKE, we implement zero-shot, one-shot, and few-shot setups to explore the impact of demonstration quality and quantity on performance Figure 2. Notably, we propose two few-shot setups: 8-shot mixed demonstrations, which expose the model to diverse query types, and 8-shot metric-specific demonstrations, which target specific query types like locality or portability to enhance performance. These setups allow us to analyze the interplay between demonstration strategies, query types, and cross-lingual transfer. Our findings show that larger models and tailored demonstrations significantly improve performance, especially for complex queries. Linguistic properties, such as syntactic and phonological similarity with English, positively influence performance, while language family has no significant impact. Instead, script type emerges as a critical factor, with non-Latin languages underperforming due to issues like language confusion, where models generate answers in English instead of the target language.

In summary, our contributions are as follows: **i)** We introduce BMIKE-53, the most comprehensive multilingual KE benchmark, covering 53 languages and three diverse KE datasets, which serves as a foundation for evaluating cross-lingual KE methods. **ii)** We extensively evaluate gradient-free cross-lingual KE methods under various IKE setups, providing valuable insights into the effectiveness of in-context learning for cross-lingual knowledge editing. **iii)** We conduct a detailed analysis of factors influencing cross-lingual KE performance, uncovering the impact of linguistic properties, script types, and language confusion on cross-lingual knowledge transfer.

2 Related Work

Traditional KE methods are primarily gradient-based. They typically introduce additional trainable parameters, such asMEND (Mitchell et al., 2021) and SERAC (Mitchell et al., 2022)–or edit specific parameters of the original model, as in-ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023). However, these methods have high computational demands and are difficult to scale. Recently, gradient-free methods, such as IKE (Zheng et al., 2023), MeLLo (Zhong et al., 2023), and ICE (Cohen et al., 2024), have been studied in KE for LLMs. Given the multilingual in-context learning capabilities of English-centric LLMs (Lai et al., 2023; Nie et al., 2024; Zhang et al., 2024a), the potential for cross-lingual KE appears promising. Recent cross-lingual KE work largely employs gradient-based methods (Xu et al., 2023; Wang et al., 2023; Beniwal et al., 2024; Wei et al., 2024). A notable gradient-free work is ReMaKE (Wang et al., 2024b), a cross-lingual retrieval-augmented KE method. However, their method is specifically applied to a rather special KE scenario – batch edit. In this setting, multiple knowledge pieces, such as the entire knowledge base, are edited simultaneously (Appendix C).

3 BMIKE-53

BMIKE-53 spans a wide range of knowledge editing perspectives, from artificial to realistic scenarios, and provides a solid foundation for evaluating cross-lingual KE methods. Additionally, with coverage of 53 languages, it stands as the most comprehensive multilingual KE benchmark to date.

Task	#Test	Q-Len.	A-Len.	#Lang.
zsRE	743	9.02	2.02	
CounterFact	1,031	5.97	1.00	53
WFD	784	4.71	2.55	

Table 1: Statistics of BMIKE-53. Q/A-Len.: Average Text Length of Query/Answer.

3.1 Datasets

As shown in Figure 1, BMIKE-53 is constructed from three monolingual KE datasets: **zsRE**, **CounterFact**, and **WikiFactDiff (WFD)**. Each dataset was selected to represent a distinct perspective of knowledge editing, ensuring the benchmark comprehensively evaluates diverse KE scenarios.

The **zsRE** dataset, originally introduced by Levy et al. (2017), was designed for zero-shot relation extraction and later adapted by De Cao et al. (2021) and Mitchell et al. (2021) for knowledge editing tasks. zsRE focuses on regular, well-defined knowledge items, making it an ideal baseline for evaluating the reliability and generality of KE methods.

The **CounterFact** dataset, introduced by Meng et al. (2022), is designed to evaluate the ability of models to update knowledge with counterfactual

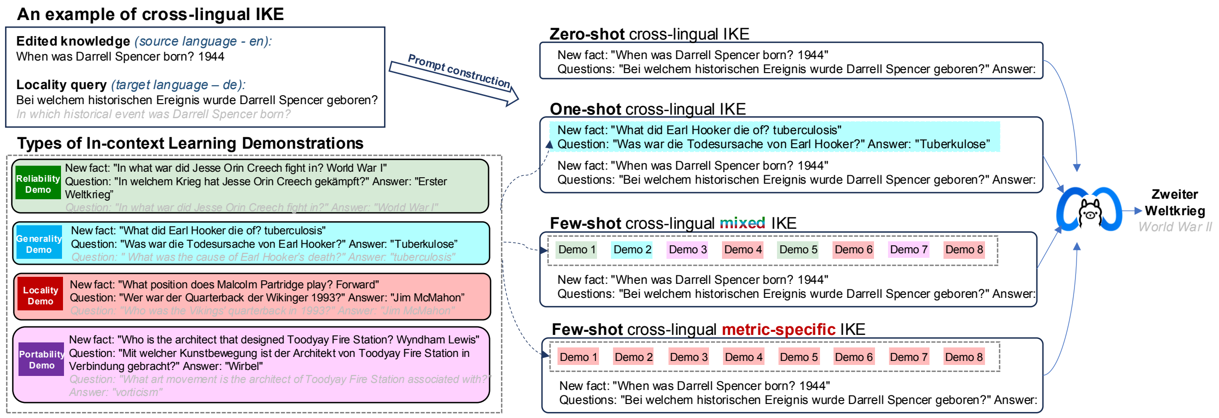

Figure 2: Cross-Lingual IKE setups and demonstration types.

(false) facts. Each entry in CounterFact represents a knowledge triple that has been altered to reflect a hypothetical or fabricated scenario. CounterFact is particularly valuable for assessing the locality of KE methods, as it requires precise updates to counterfactual knowledge without unintended side effects.

The **WikiFactDiff (WFD)** dataset, introduced by Khodja et al. (2024), focuses on real-world, temporally recent knowledge updates. Derived from WikiData (Vrandečić and Krötzsch, 2014), WFD captures changes to knowledge triples that reflect actual updates in the real world, such as changes in political leadership, scientific discoveries, or other evolving facts. WikiFactDiff is essential for assessing the real-world applicability of KE methods, as it introduces the challenge of updating models with temporally recent and realistic knowledge changes.

3.2 Benchmark Construction

The construction of the BMIKE-53 benchmark involves three key steps: unifying data formats, multilingual expansion, and quality control. These steps ensure that the benchmark is consistent, multilingual, and of high quality, enabling robust evaluation of cross-lingual knowledge editing (KE) methods.

Unifying Data Formats To ensure consistency across the three datasets, we standardized their formats into a unified structure to create a cohesive English base dataset. Each data item in the unified format includes an edited knowledge item and four types of test queries: reliability, generality, locality, and portability. The portability queries for zsRE and CounterFact were adopted from the work of Yao et al. (2023), while for WFD, we extracted a knowledge graph from the original WFD dataset and performed one-hop knowledge reasoning within the graph to generate portability queries. All data items are stored in a JSON format, with a consistent data structure across the three datasets. This unified format facilitates seamless integration into the benchmark and enables efficient LLM-assisted translation in the multilingual expansion process.

Structured Multilingual Expansion with LLMs To create a multilingual benchmark for cross-lingual KE tasks, we expanded the English base data into 52 target languages using LLM-assisted structural translation. The language coverage is adopted from similar multilingual datasets like MLAMA (Kassner et al., 2021) and BMLAMA (Qi et al., 2023). App. A.1 provides details regarding the language coverage.

We employed the GPT-4o model[1] via the OpenAI API for the multilingual expansion. The translation process was guided by a structured prompt template (see Appendix A.3), which ensured that the JSON format and structure of the data items were preserved during translation. Specifically: Only the text values within the JSON structure were translated, while the key names remained unchanged. The prompt explicitly instructed the model to output the translated data in plain text, adhering to the original JSON format. We compared several machine translation tools, including NLLB-200 and Google Translate API, but found that LLM-assisted translation was more appropriate for our use case due to its accuracy and flexibility in handling structured data formats (Appendix A.3).

Quality Control To ensure the quality of the multilingual expansion, we implemented a rigorous quality control process involving qualitative eval-

[1] `gpt-4o-2024-08-06`

Model	Setup	zsRE				CounterFact				WikiFactDiff			
		rel	gen	loc	port	rel	gen	loc	port	rel	gen	loc	port
`Llama3.2-3B`	*zero-shot*	50.16	49.03	5.64	5.41	43.51	40.84	7.53	5.61	58.28	57.41	6.77	3.18
	one-shot	71.57	71.25	7.78	14.97	68.34	68.02	6.58	16.30	66.32	65.93	6.31	3.06
	8-shot (mix)	70.54	70.49	8.89	16.43	70.80	70.33	5.11	13.75	64.97	64.60	6.24	4.00
	8-shot (metric)	70.94	70.91	12.23	22.97	67.57	67.14	31.61	31.20	67.77	67.48	9.14	10.72
`Llama3.1-8B`	*zero-shot*	65.53	64.09	9.76	10.05	63.01	60.59	18.68	11.16	67.84	66.40	10.04	4.15
	one-shot	75.27	74.90	13.36	20.81	71.92	71.29	12.66	21.93	70.53	69.84	7.80	4.15
	8-shot (mix)	74.29	74.00	15.46	25.18	75.15	74.42	11.40	23.74	68.57	67.86	8.27	8.87
	8-shot (metric)	74.86	74.79	16.15	32.86	73.88	73.19	47.55	41.17	71.98	71.34	13.84	14.58

Table 2: Main Results. Average cross-lingual IKE performance across 52 languages (F1-score).

uation and quantitative analysis. We conducted a manual review of sampled sentences by native speakers of selected languages, then we used back-translation techniques to provide an overall assessment of translation quality. Specifically, each translated sentence was back-translated into English, and the BLEU score and semantic similarity were calculated. Table 5 shows the results of translation quality control via the back-translation.

4 Experiments

The primary goal of this work is to extensively investigate the performance of cross-lingual in-context knowledge editing (IKE). Using the proposed benchmark BMIKE-53, we aim to explore the factors influencing cross-lingual IKE performance, identify performance tendencies, and analyze variations in cross-lingual behavior across different languages and query types. To achieve this, we first formally define the cross-lingual IKE task and its evaluation framework. We then introduce the different IKE setups and strategies explored in our experiments

4.1 Task: Cross-Lingual In-Context Knowledge Editing

The Cross-Lingual In-Context Knowledge Editing (IKE) task evaluates a language model's ability to incorporate new knowledge (a fact) in one language and apply it across multiple languages while preserving unrelated knowledge. This task leverages in-context learning (ICL) to guide the model in editing and applying knowledge through demonstrations. Below, we formally define the task and the four types of cross-lingual queries used to evaluate the model's performance.

Task Definition Given a language model $\mathcal{M}$; a new fact represented as a query-answer pair $f = (x_s^*, y_s^*)$ in the source language s, where x_s^* is the query and y_s^* is the corresponding answer; a set $\mathcal{X}_s^*$, which contains x_s^* and other semantically equivalent queries in the source language; the translations of $\mathcal{X}_s^*$ into a target language t, denoted as $\mathcal{X}_t^* = \{I_t(x_s) : x_s \in \mathcal{X}_s^*\}$ where $I^t(\cdot)$ is a translator mapping source language queries to their target language counterparts. The task involves evaluating the model's response to a target language query x_t after incorporating the new fact f. Specifically, the model assigns a probability $P_{\mathcal{M}}(y|x_t, f)$ to an answer y given the query x_t and the fact f. The predicted answer is defined as $Pred(x_t, f) = \arg\max_y \mathcal{P}_{\mathcal{M}}(y|x_t, f)$. The goal is for the model to predict the correct translation of the fact answer, $I^t(y_s^*)$, when the query x_t is semantically equivalent to the fact query x_s^*, and to preserve the original knowledge and predict the correct answer y_t for unrelated queries, ensuring no unintended inference from the knowledge editing process.

Cross-Lingual Query Types To comprehensively evaluate the model's cross-lingual knowledge editing capabilities, we define four types of target language queries, each testing a specific aspect of the task. The model should reliably apply the new fact to the exact translation of the original query. A ***reliability*** query x_t is defined as $x_t = I_t(x_s^*)$. ***Generality*** queries test whether the model can generalize the new fact to other semantically equivalent queries in the target language that differ in phrasing or structure. A generality query x_t is defined as $x_t \in \mathcal{X}_t^* \backslash \{I_t(x_s^*)\}$. The expected answer for the reliability and generality query is $Pred(x_t, f) = I_t(y_s^*)$. ***Portability*** queries evaluate whether the model can apply the new fact to related but contextually different queries in the target language. These queries are derived through one-hop knowledge reasoning from the original fact. Let $\mathcal{X}_{s,\text{1-hop}}^*$ denote the one-hop query set in the source language, which includes x_s^* and queries influenced by x_s^* through one-hop reasoning in a knowledge graph. The correspond-

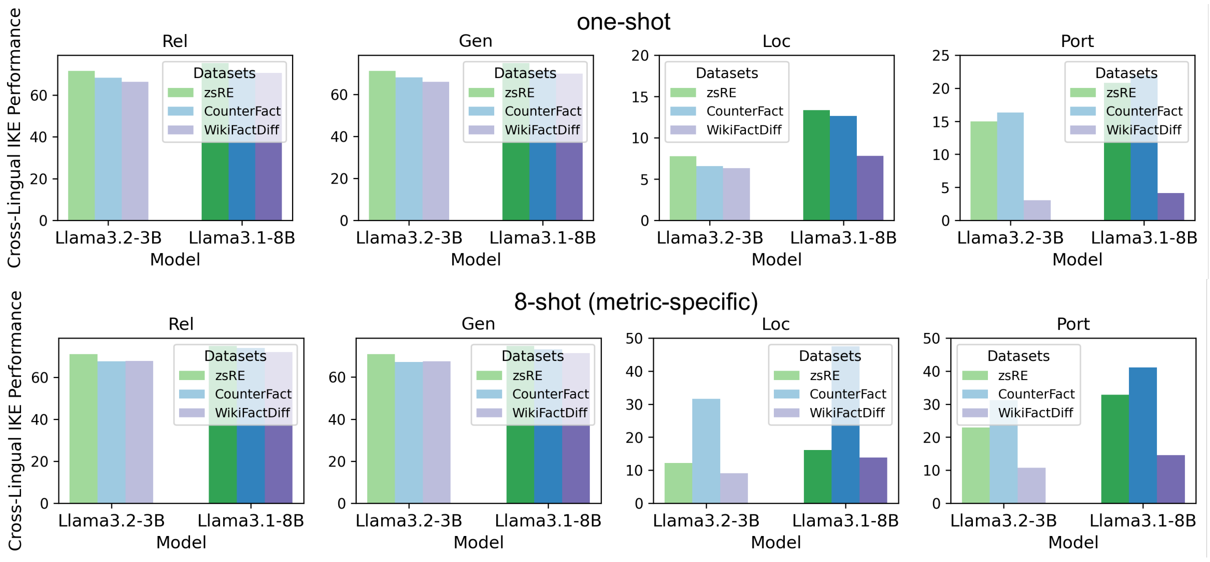

Figure 3: Average cross-lingual IKE performance across languages (F1-score).

ing target language set is $\mathcal{X}^*_{t,\text{1-hop}} = \{I_t(x^*_s) : x_s \in \mathcal{X}^*_{s,\text{1-hop}}\}$. A portability query x_t is defined as $x_t \in \mathcal{X}^*_{t,\text{1-hop}}$. ***Locality*** queries test whether the model can preserve unrelated knowledge while incorporating the new fact. A locality query x_t is defined as $x_t \notin \mathcal{X}^*_{t,\text{1-hop}}$. The expected answer is $Pred(x_t, f) = y_t$.

4.2 Cross-Lingual IKE Setup

IKE Demonstrations In the context of in-context learning (ICL), demonstrations are examples provided in the input prompt to guide the model's behavior. For the IKE task, the demonstrations are designed to teach the model how to perform cross-lingual knowledge editing. Formally, the set of demonstrations is defined as $C = \{c_1, \cdots, c_k\}$, where each demonstration c_i consists of a new fact $f' = (x'_s, y'_s)$ in the source language, a query x'_t in the target language, and the correct answer y'_t in the target language. As an ICL-based KE method, IKE uses demonstrations to guide the model in learning the relationships between the source language fact and the target language queries. As illustrated in Figure 2, the demonstrations are designed to reflect the four query types, enabling the model to learn the appropriate behavior for each type. By including examples of cross-lingual queries and their correct answers, the demonstrations help the model generalize the knowledge editing process across languages.

IKE Setup As illustrated in Figure 2, we evaluate cross-lingual IKE under four distinct setups: zero-shot, one-shot, few-shot mixed, and few-shot metric-specific. In ***zero-shot*** cross-lingual IKE, the model performs cross-lingual knowledge editing without any demonstrations, relying solely on its pre-trained capabilities. In a ***one-shot*** setup, a single randomly selected demonstration is provided to familiarize the model with the task format. This setup is designed to give the model an overview of the task without significantly aiding its ability to complete the task. ***Few-shot mixed*** cross-lingual IKE includes eight demonstrations of mixed types. The goal is to teach the model cross-lingual knowledge editing by exposing it to diverse query types. To enhance performance on specific query types, the ***few-shot metric-specific*** variation provides eight demonstrations of the same query type as the test target.

4.3 Experimental Setting

We conduct experiments using two multilingual LLMs: Llama3.2-3B and Llama3.1-8B. These models were selected for their multilingual capabilities and represent different model sizes, allowing us to analyze the impact of model scale on cross-lingual IKE performance. To measure the cross-lingual IKE performance, we compare the predicted answers with the ground-truth answers using the **F1** score and Exact Match (**EM**) metrics, consistent with prior work (Wang et al., 2023, 2024b). The implementation details are provided in Appendix D.

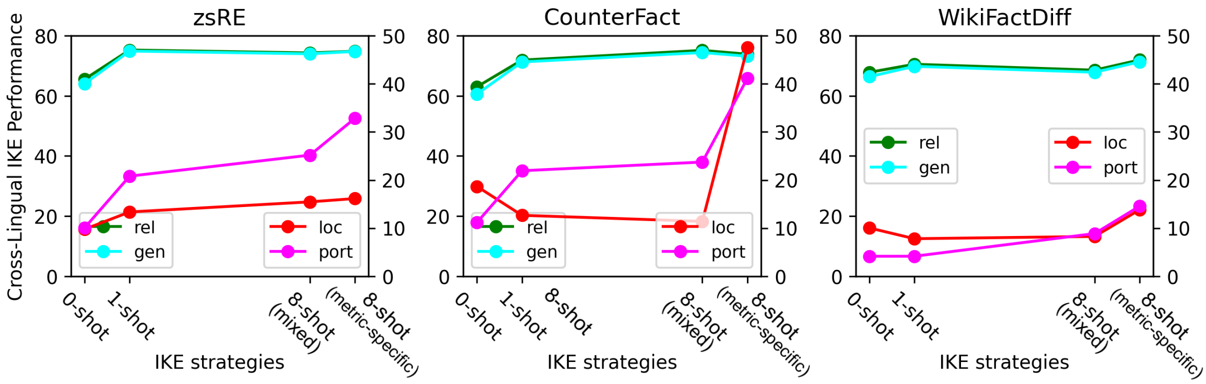

Figure 4: Average cross-lingual IKE performance of Llama3.1-8B across 52 languages (F1-score).

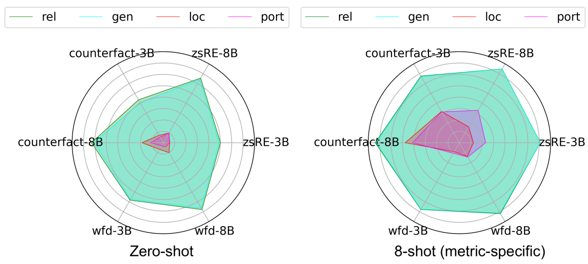

Figure 5: Average cross-lingual IKE performance across 52 languages (F1-score).

5 Multidimensional Analysis of Cross-Lingual IKE

This section provides a multidimensional analysis of cross-lingual IKE performance, focusing on the effects of model size, dataset-specific performance, query type variations, and IKE setup strategies. Using the BMIKE-53 benchmark, we aim to uncover key insights into how these factors influence IKE performance and cross-lingual variations.

5.1 Effects of Model Scale

As shown in Figure 3, larger models demonstrate superior cross-lingual IKE capabilities across all experimental configurations. Notably, larger models show particular advantages in handling queries requiring complex multilingual reasoning and knowledge preservation, evidenced by the performance gap of locality (loc) and portability (port) queries.

5.2 Dataset-Specific Performance Patterns

We observe substantial cross-dataset variance in editing efficacy, especially with loc and port queries, indicative of inherent dataset complexity differences. While WikiFactDiff (WFD) achieves comparable reliability (rel) and generality (gen) scores to zsRE and CounterFact, it shows the lowest port performance. This discrepancy could be attributed to the real-world nature of WFD, where all knowledge–both the original and updated facts–is temporally recent. Portability queries in WFD require the model to reason over a second-order knowledge chain, where the correct answer depends on understanding the relationship between the updated fact and its broader context, CounterFact, on the other hand, benefits the most from metric-specific demonstrations, particularly for locality and portability. The compositional nature of CounterFact's fabricated facts allows the model to leverage targeted demonstrations effectively.

5.3 Query-Type Sensitivity

The four query types exhibit divergent response patterns to IKE strategies. Figure 5 shows that rel and gen achieve near-parity across setups, especially in 8-shot metric-specific IKE, suggesting models effectively align cross-lingual surface forms. In contrast, loc and port perform substantially worse across datasets and setups. However, port demonstrates greater sensitivity to metric-specific demonstrations than loc. Port benefits from metric-specific demonstrations with more pronounced improvements, especially in datasets like zsRE and WFD.

5.4 Impact of Demonstration Strategies

The comparison of zero-shot, one-shot, few-shot mixed, and few-shot metric-specific setups reveals the importance of demonstration quality and quantity in cross-lingual IKE. Figure 4 illustrates how the performance of `Llama3.1-8B` evolves across setups for each query type and dataset. In the zero-shot setup, models rely solely on their pre-trained capabilities, resulting in limited perfor-

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy
zsRE																										
rel	73	53	66	53	55	26	76	80	79	76	78	81	51	73	78	66	45	76	74	75	75	47	58	75	71	26
gen	73	54	67	53	55	25	76	80	78	76	79	82	51	74	78	67	44	75	74	74	74	46	59	75	71	26
loc	69	19	49	12	27	12	68	63	58	55	75	79	29	71	55	59	19	58	72	45	66	22	34	58	62	16
port	65	29	43	37	42	13	69	54	66	46	72	77	31	69	58	35	29	65	73	40	68	26	32	60	60	13
avg	70	39	56	39	45	19	73	69	70	63	76	80	41	72	67	57	34	68	73	59	70	35	46	67	66	20
CounterFact																										
rel	67	65	52	65	67	54	75	82	60	65	80	89	82	77	35	45	75	37	73	50	79	86	82	50	76	51
gen	72	69	54	69	71	56	79	86	61	68	84	94	85	80	38	50	80	40	74	52	82	90	87	52	80	53
loc	86	50	67	56	58	50	76	69	71	74	91	97	75	94	79	67	60	45	90	51	76	50	68	80	64	34
port	58	47	28	49	53	21	70	70	69	48	73	80	50	79	59	42	35	59	76	39	80	38	37	60	55	16
avg	71	58	50	60	62	45	75	77	65	64	82	90	73	82	53	51	62	45	78	48	79	66	69	60	69	38
WikiFactDiff																										
rel	83	22	76	27	21	14	85	80	79	77	85	81	36	85	80	78	24	76	83	82	77	20	31	79	81	10
gen	83	22	77	26	19	13	83	81	80	77	84	82	35	84	80	77	24	77	82	81	78	20	31	79	81	10
loc	72	40	55	27	46	11	92	92	46	65	56	71	33	93	64	50	23	61	87	53	86	16	38	57	66	12
port	80	32	70	12	32	11	75	34	69	58	77	67	21	84	70	60	27	61	87	60	78	35	22	70	76	9
avg	79	29	70	23	29	12	84	72	68	69	75	75	31	87	74	66	25	69	85	69	79	23	31	71	76	10

	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh
zsRE																										
rel	80	74	71	25	56	59	62	51	80	79	75	73	76	60	78	78	63	56	76	27	54	76	60	42	73	61
gen	79	74	71	24	57	58	63	51	80	78	74	73	75	60	78	77	64	55	77	28	53	76	61	41	74	61
loc	78	68	27	16	22	42	43	30	69	83	56	67	64	41	58	62	53	36	68	7	39	61	35	17	65	19
port	75	73	44	8	32	35	40	35	67	76	66	75	67	54	63	59	42	37	68	5	29	65	52	19	54	48
avg	78	72	53	18	42	48	52	41	74	79	68	72	70	54	69	69	55	46	72	17	44	70	52	30	67	47
CounterFact																										
rel	75	69	92	43	81	50	33	31	79	82	61	71	79	66	58	58	61	64	82	46	80	78	54	60	71	83
gen	79	71	97	45	85	52	33	31	83	86	65	72	82	73	62	62	64	65	86	49	85	83	57	64	74	88
loc	77	93	78	30	76	64	45	40	84	99	79	81	87	78	66	76	70	60	92	31	70	86	71	49	76	84
port	69	78	52	12	38	34	43	37	57	76	74	73	70	70	57	52	51	34	75	13	34	62	57	28	59	54
avg	75	78	80	33	70	50	39	35	76	86	70	74	80	72	61	62	62	56	84	35	67	77	60	50	70	77
WikiFactDiff																										
rel	82	85	23	4	23	77	67	62	91	87	79	77	87	30	78	75	75	59	84	5	56	92	21	16	87	35
gen	82	85	23	5	22	77	65	63	91	86	79	72	87	32	78	75	77	53	83	5	56	92	25	16	86	34
loc	73	88	42	14	13	64	61	75	69	82	45	78	86	36	52	65	61	41	44	6	39	61	38	46	60	52
port	70	79	14	3	16	34	37	32	78	86	71	91	85	46	70	79	74	15	82	2	42	79	39	16	77	30
avg	77	84	26	7	19	63	57	58	82	85	69	80	86	36	70	74	72	42	73	4	48	81	31	23	77	38

Table 3: Cross-lingual IKE performance of Llama3.1-8B in 52 languages under the 8-shot metric-specific setup.

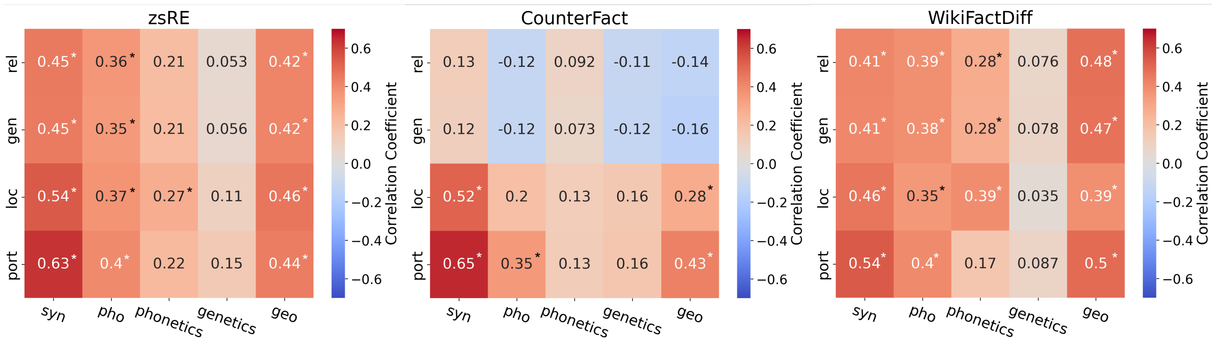

Figure 6: Correlation between linguistic properties and IKE performance of 52 languages. Experimental results of Llama3.1-8B under the 8-shot metric-specific setup. Significant correlations are marked with $*$ ($p < 0.05$).

mance. While one-shot setups provide modest improvements by familiarizing the model with the task format, they fail to address the challenges of loc. A single, randomly selected demonstration often confuses the model, particularly when the demonstration type does not align with the target query type. For loc queries, when encountering rel or gen demonstration types, this misalignment can lead the model to incorrectly repeat the edited knowledge in the target language, further degrading performance. Adding more demonstrations in the few-shot mixed setup yields limited performance gains, particularly for loc queries. However, when demonstrations are tailored to the specific query type being tested, as in the 8-shot metric-specific setup, notable gains are observed This setup achieves the highest scores for loc and portability, as shown in Figure 4. These findings underscore the importance of demonstration quality and specificity in maximizing the benefits of in-context learning for cross-lingual knowledge editing.

6 Language Performance Variance in Cross-Lingual IKE

To analyze cross-lingual performance variation, we use a normalized metric that uses the exact match (EM) of English as a reference to highlight cross-lingual transfer differences. Specifically, we calculate the ratio of each target language's EM performance to English. As visually evident in Table 3, some languages perform particularly poorly, while others achieve results closer to English, motivating further investigation into the factors influencing these differences.

6.1 Correlation Between Language Properties and Performance

We conducted a correlation analysis between language properties (derived using Lang2Vec (?), details in Appendix) and query-type performance across datasets. As revealed in Figure 6, syntactic, phonological, and geographic similarities with English positively correlate with performance, particu-

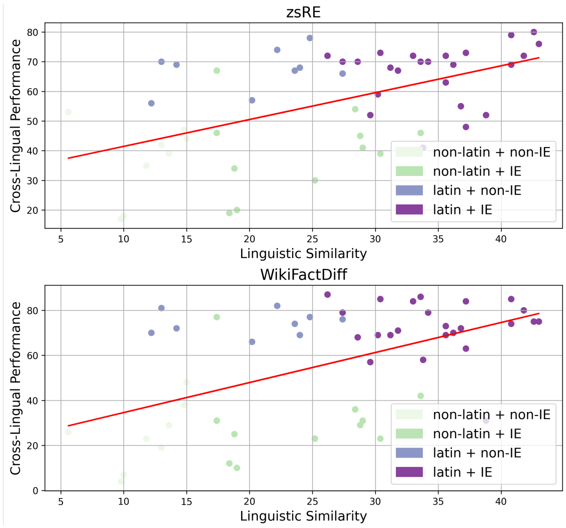

Figure 7: Panorama of per-language performance.

larly for loc and port queries However, two notable exceptions emerge: rel and gen queries in CounterFact show no significant correlation with language properties, likely because these query types achieve uniformly high performance across all languages. Additionally, genetic similarity (language family) shows no meaningful correlation across datasets.

6.2 Impact of Script and Language Family

To further explore the role of script and language family, we grouped languages into four clusters based on script type (Latin vs. non-Latin) and language family (Indo-European vs. non-Indo-European). Figure 7 reveals that script type plays a more critical role than language family. Non-Latin languages, regardless of their family, perform worse than Latin-script languages. This trend is consistent across datasets and query types, as further supported by Figure 8. For example, the non-Latin + non-IE and non-Latin + IE groups exhibit similar performance, both significantly lower than the Latin + IE and Latin + non-IE groups.

6.3 Language Confusion and Script Effects

A qualitative error analysis highlights language confusion as a key factor explaining why script type matters. Language confusion occurs when the model generates answers in English instead of the target language (Marchisio et al., 2024), even when explicitly instructed to use the target language. This issue is particularly prevalent in code-switched prompts and disproportionately affects non-Latin languages. For example, while the model may produce the correct answer in English,

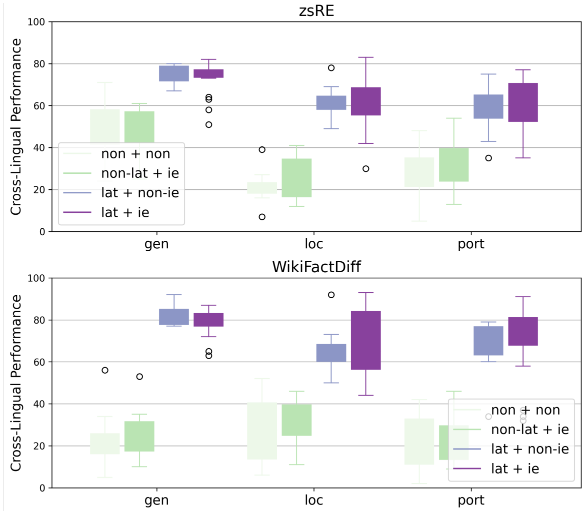

Figure 8: IE: Indo-European language family.

zsRE (Case id: 17) - Locality

New Fact: "What vessel type is USS Flusser (DD-368)? 368-class destroyer"
Question: "连接一个顶点到自身的边的术语是什么？"
What is the term for an edge connecting a vertex to itself?
Model Answer: loop True Answer: 环*loop*

CounterFact (Case id: 128) - Portability

New Fact: "Where does Spike Hughes originate from? Philadelphia"
Question: "与斯派克·休斯来自的城市相关的著名食物是什么？"
What is the famous food associated with the city Spike Hughes is from?
Model Answer: Cheesesteak True Answer:奶酪牛排*Cheesesteak*

Table 4: Examples of language confusion from the 8-shot metric-specific setup with Llama3.1-8B.

it is considered a failure if the output is not in the target language. Table 4 highlights instances of language confusion that occur when queries are presented in Chinese. This problem is exacerbated for non-Latin languages, as their distinct scripts reduce the likelihood of overlapping writing forms between English and the target language, further degrading performance.

7 Conclusion

We present BMIKE-53, a multilingual KE benchmark, and leverage it to investigate the potential of gradient-free in-context learning methods for cross-lingual knowledge editing. Our experiments demonstrate that tailored demonstration strategies significantly enhance KE performance, with metric-specific demonstrations improving locality and portability. Additionally, linguistic properties, particularly script type, strongly influence cross-lingual knowledge transfer. We hope that BMIKE-53 will inspire further research in multilingual KE, advancing the understanding and capabilities of LLMs across diverse linguistic contexts.

Limitation

This research focuses on the application of gradient-free in-context learning (ICL) methods for cross-lingual knowledge editing (KE) using BMIKE-53, the most comprehensive multilingual KE benchmark to date. While our study highlights the effectiveness of ICL-based methods and tailored demonstration strategies, certain areas remain for further exploration. For instance, gradient-based KE methods, which are computationally intensive, were explored very deeply in this investigation as our primary focus was on the efficiency and practicality of gradient-free approaches. Additionally, while BMIKE-53 provides valuable insights into the challenges of linguistic diversity, future work could aim to further optimize performance for non-Latin languages and address language confusion issues. These extensions could deepen our understanding of cross-lingual KE and enhance the applicability of BMIKE-53 as a resource for advancing multilingual LLM research.

Ethic Statement

This research was conducted in accordance with the ACM Code of Ethics. The datasets used in this paper are publicly available. We do not intend to share any Personally Identifiable Data with this paper. Our project may raise awareness of these low-resource languages in the computational linguistics community.

References

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

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

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258.

Hichem Ammar Khodja, Frédéric Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2024. Wikifactdiff: A large, realistic, and temporally adaptable dataset for atomic factual knowledge update in causal language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17614–17624.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–

6677, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schütze. 2024. Decomposed prompting: Unveiling multilingual linguistic structure knowledge in english-centric large language models. *Preprint*, arXiv:2402.18397.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023. Cross-lingual knowledge editing in large language models. *arXiv preprint arXiv:2309.08952*.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024a. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024b. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Mlake: Multilingual knowledge editing benchmark for large language models. *arXiv preprint arXiv:2404.04990*.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024a. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, and Shumin Deng. 2024b. Knowledge editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 33–41, Torino, Italia. ELRA and ICCL.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping

Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A BMIKE-53 Details

A.1 Language Coverage

BMIKE-53 encompasses a total of 53 languages. A comprehensive list of these languages can be found in Table 8. Additionally, the table outlines the linguistic feature similarities between each target language and English.

A.2 Data Entry Example

Figure 10 shows the data item examples of BMIKE-53 for all three datasets.

A.3 LLM-Assisted Translation

Figure 9 shows the prompt template used for our LLM-assisted translation.

Why LLM-Assisted Translation? We compared several machine translation tools, including NLLB-200 and Google Translate API, but found that LLM-assisted translation offered superior performance in terms of:

- Accuracy: LLMs demonstrated better handling of complex linguistic structures and domain-specific terminology.
- Flexibility: LLMs provided greater adaptability for processing structured data formats like JSON.
- Consistency: The structured translation process ensured that the multilingual data remained aligned with the original English data.

By leveraging LLM-assisted translation, we ensured that the multilingual benchmark maintained high linguistic quality and structural integrity.

A.4 Translation Quality Control

We used back-translation techniques to assess the quality of the translations. Specifically, each translated sentence was back-translated into English, We calculated the BLEU score and semantic similarity between the original and the back-translated English text. BLEU Score measures the formal similarity between the original and back-translated sentences. Semantic Similarity evaluates the semantic alignment between the original and back-translated sentences using cosine similarity in a sentence embedding space. The results of the back-translation evaluation are shown in Table 5.

system:
You are an intelligent multilingual translation assistant that can structurally translate English text in a fixed data format into 52 different languages.

user:
Translate the following JSON data item from English to `{target language}`. Keep the original JSON format and structure, translating only the text in the values while keeping the key names unchanged. Output only in plain text without additional formatting text. Use double quotes for key and value names and add the escape character for quote marks in the text of value: `{json_data_item}`

Figure 9: Prompt template for `GPT-4o` as translation assistant.

	zsRE		CounterFact		WFD	
Lang.	**BLEU**	**Sim.**	**BLEU**	**Sim.**	**BLEU**	**Sim.**
es	0.81	0.94	0.82	0.90	0.81	0.90
vi	0.82	0.93	0.77	0.91	0.78	0.90
ru	0.78	0.91	0.71	0.87	0.72	0.87
zh	0.78	0.89	0.76	0.85	0.76	0.85
de	0.84	0.93	0.82	0.92	0.82	0.92

Table 5: Results of Translation Quality Control via Back-Translation.

B BMIKE-53 with Other Baseline Methods

The primary purpose of this paper is to provide a basic gradient-free method for cross-lingual KE. However, it is helpful for better understanding our gradient-free method to compare it with other existing baseline methods, including gradient-based methods. Therefore, we also experimented on BMIKE-53 with commonly used baseline methods, including fine-tuning (FT), LoRA (Hu et al.), ROME (Meng et al., 2022), and KN (Dai et al., 2022) in selected languages. We use the EasyEdit framework (Wang et al., 2024a) to conduct the baseline experiments. Table 6 shows the baseline method results.

C Further Discussion on Related Work

As discussed in §2, a few previous studies have explored gradient-free knowledge editing meth-

Query	Baseline	ar	de	fa	fr	he	hu	ja	ru	tr	zh
Reliability	**FT**	24.9	22.4	20.5	19.6	26.0	35.0	34.4	17.0	24.1	23.4
	LoRA	47.1	46.0	47.0	44.8	43.9	44.1	53.1	35.4	41.3	47.9
	ROME	34.9	36.3	35.8	33.7	25.8	32.2	42.6	24.8	30.7	38.0
	KN	20.1	3.7	24.0	3.7	15.8	5.1	28.0	8.5	5.7	23.2
Generality	**FT**	25.6	23.6	21.7	20.8	25.9	34.5	34.0	17.1	25.2	24.5
	LoRA	44.6	45.4	46.7	44.4	42.7	42.8	52.7	34.8	41.0	47.0
	ROME	33.7	35.1	36.6	32.7	25.3	30.4	43.0	25.4	31.3	38.8
	KN	20.7	3.5	24.8	3.7	16.1	5.5	28.2	8.4	5.8	24.1
Locality	**FT**	37.5	18.7	31.6	15.1	41.0	35.6	39.9	32.0	25.7	39.8
	LoRA	38.1	26.3	37.4	26.1	41.0	35.1	42.3	32.4	29.6	45.5
	ROME	78.9	75.8	79.8	78.3	83.9	87.3	80.8	77.7	83.2	84.6
	KN	51.3	60.2	54.6	57.5	62.8	64.7	56.8	54.7	60.3	61.1
Portability	**FT**	32.7	16.8	28.9	14.1	34.7	18.5	35.1	19.6	19.0	27.6
	LoRA	42.9	27.2	42.8	24.0	40.4	23.7	44.4	28.4	27.8	37.2
	ROME	40.2	26.9	41.1	25.2	36.7	23.5	42.2	26.6	29.0	37.5
	KN	28.3	16.5	29.6	13.6	27.2	15.2	31.0	17.0	18.2	26.5

Table 6: Experimental Results of CounterFact with Gradient-based Baseline Methods. F1 scores are reported.

ods, notably ReMaKE (Wang et al., 2024b). Our work diverges from these existing approaches in several key aspects, which we will elaborate on below. Regarding task setup, ReMaKE employs a cross-lingual retrieval-augmented strategy tailored for batch edits, allowing simultaneous modifications of multiple knowledge pieces, such as an entire knowledge base. Conversely, our approach focuses on individual knowledge edits, which do not involve cross-lingual retrieval. his fundamental difference sets our approach apart and addresses a unique aspect of knowledge editing. In prompt engineering, when designing the demonstrations of ICL, ReMaKE uses translation pairs of source and target language facts to make up the cross-lingual demonstrations, from which the model cannot learn real cross-lingual knowledge editing competencies like portability and locality. In contrast, our MIKE method provides four different types of cross-lingual ICL demonstrations. From these demonstrations, the model can effectively learn real cross-lingual knowledge editing.

D Experiment Implementation Details

Table 7 displays the experiment implementation details. We use the models from HuggingFace (`https://huggingface.co/meta-llama/Llama-3.2-3B`, `https://huggingface.co/meta-llama/Llama-3.1-8B`).

Parameter	Value
Model	`meta-llama/Llama-3.1-8B`, `meta-llama/Llama-3.2-3B`
Max. length	4096
Num. of demonstration	8
Type of demonstration	Reliability, Generality, Locality, Portability
Proportion of demo.	1:3:2:2
GPU Type	NVIDIA A100-SXM4-80GB
Number of GPU	4
Running hours	72

Table 7: Experimental Implementation Details.

E Full Results

We show the full experimental results for all three tasks in Table 9 (zsRE), Table 10 (CounterFact), and Table 11 (WFD), respectively.

lid	language	Family	syn_sim	pho_sim	inv_sim	gen_sim	geo_sim
af	Afrikaans	Indo-European (Germanic)	84.94	81.83	69.10	50.46	86.84
ar	Arabic	Semitic	65.11	70.09	70.81	0.15	97.04
az	Azerbaijani	Turkic	52.00	81.83	67.86	0.19	96.96
be	Belarusian	Slavic	78.64	85.83	70.42	16.80	99.35
bg	Bulgarian	Slavic	85.78	85.83	68.38	13.73	99.01
bn	Bengali	Indo-European (Indo-Aryan)	58.36	76.30	74.38	12.71	88.96
ca	Catalan	Indo-European (Romance)	87.30	85.83	75.22	10.64	99.66
ceb	Cebuano	Austronesian	62.17	76.30	75.22	0.13	81.50
cs	Czech	Slavic	73.99	85.83	66.51	13.73	99.71
cy	Welsh	Indo-European (Celtic)	71.90	81.83	77.85	13.73	99.99
da	Danish	Indo-European (Germanic)	88.01	81.83	77.54	40.90	99.89
de	German	Indo-European (Germanic)	90.26	80.60	76.28	54.49	99.76
el	Greek	Indo-European (Hellenic)	78.31	95.35	64.76	15.03	98.96
es	Spanish	Indo-European (Romance)	82.16	85.83	63.83	9.71	99.59
et	Estonian	Uralic	77.35	85.83	66.94	0.23	99.45
eu	Basque	Isolate	62.36	85.29	56.88	3.33	99.76
fa	Persian	Indo-European (Iranian)	50.03	78.35	72.83	13.73	94.23
fi	Finnish	Uralic	71.08	87.05	70.00	0.19	99.19
fr	French	Indo-European (Romance)	81.18	75.28	74.09	9.71	99.93
ga	Irish	Indo-European (Celtic)	72.01	85.83	69.35	12.71	99.96
gl	Galician	Indo-European (Romance)	80.23	90.46	70.75	10.14	99.65
he	Hebrew	Semitic	75.15	72.55	64.37	0.13	97.16
hi	Hindi	Indo-European (Indo-Aryan)	61.63	78.35	70.91	12.71	91.10
hr	Croatian	Slavic	83.18	85.83	69.67	12.71	99.50
hu	Hungarian	Uralic	69.40	85.83	74.03	0.33	99.46
hy	Armenian	Indo-European (Satem)	63.03	69.66	68.73	19.39	97.23
id	Indonesian	Austronesian	72.66	90.92	75.58	0.12	79.16
it	Italian	Indo-European (Romance)	85.78	85.83	70.00	11.21	99.53
ja	Japanese	Isolate	50.03	66.77	65.40	0.19	85.65
ka	Georgian	Caucasian	68.50	66.93	62.93	0.19	97.09
ko	Korean	Isolate	55.29	74.65	70.94	0.33	86.93
la	Latin	Indo-European (Romance)	78.27	85.83	76.76	15.03	99.47
lt	Lithuanian	Indo-European (Baltic)	69.33	80.42	74.63	19.39	99.44
lv	Latvian	Indo-European (Baltic)	75.39	81.83	75.22	19.39	99.42
ms	Malay	Austronesian	70.49	90.92	72.49	0.15	80.49
nl	Dutch	Indo-European (Germanic)	92.43	81.83	72.24	44.51	99.96
pl	Polish	Slavic	78.64	85.83	65.29	15.03	99.63
pt	Portuguese	Indo-European (Romance)	84.24	90.46	78.68	10.14	99.68
ro	Romanian	Indo-European (Romance)	79.60	90.46	73.42	11.89	99.22
ru	Russian	Slavic	81.18	85.83	64.76	16.80	95.81
sk	Slovak	Slavic	82.16	85.83	70.66	15.03	99.55
sl	Slovenian	Slavic	80.59	85.83	75.58	15.03	99.62
sq	Albanian	Indo-European (Other)	79.60	87.05	72.49	33.48	99.19
sr	Serbian	Slavic	79.60	85.83	72.94	12.71	99.23
sv	Swedish	Indo-European (Germanic)	93.34	81.83	67.98	40.90	99.62
ta	Tamil	Dravidian	51.36	85.29	65.81	0.11	87.95
th	Thai	Kra-Dai	63.95	78.35	74.91	0.11	85.25
tr	Turkish	Turkic	50.68	81.83	66.59	0.14	98.25
uk	Ukrainian	Slavic	84.73	85.83	74.38	15.03	99.28
ur	Urdu	Indo-European (Indo-Aryan)	61.63	85.83	71.57	12.71	92.54
vi	Vietnamese	Austroasiatic	66.04	78.35	74.74	0.19	85.25
zh-cn	Chinese	Sino-Tibetan	71.08	72.55	69.73	0.33	88.42

Table 8: Detailed information of the languages covered by BMIKE-53. The right five columns show the linguistic feature similarities between the target language and English. syn: syntax, pho: phonology, inv: phonetics, gen: phylogenetic, geo: geographic, sim: similarity.

zsRE

```
"en": {
    "case_id": 8,
    "subject": "Chlorophyll Kid",
    "src": "Which fictional universe is Chlorophyll Kid part of?",
    "rephrase": "What fictitious universe is the figure of Chlorophyll Kid associated with?",
    "alt": "Image Universe",
    "loc": "Which language did the recipient of the first Jnanpith Award write in?",
    "loc_ans": "Malayalam",
    "port": "Who is one of the founders of the fictional universe that Chlorophyll Kid is part of?",
    "port_ans": "Todd McFarlane"
},
"th": {
    "case_id": 8,
    "subject": "เด็กคลอโรฟิลล์",
    "src": "จักรวาลนิยายใดที่เด็กคลอโรฟิลล์เป็นส่วนหนึ่งของ?",
    "rephrase": "จักรวาลที่เป็นนิยายใดที่ตัวละครเด็กคลอโรฟิลล์เกี่ยวข้อง?",
    "alt": "ภาพจักรวาล",
    "loc": "ผู้รับรางวัล Jnanpith Award คนแรกเขียนด้วยภาษาอะไร?",
    "loc_ans": "มลายู",
    "port": "ใครคือหนึ่งในผู้ก่อตั้งจักรวาลนิยายที่เด็กคลอโรฟิลล์เป็นส่วนหนึ่งของ?",
    "port_ans": "ท็อดด์ แมคฟาร์เลน"
}
```

CounterFact

```
"en": {
    "case_id": 46,
    "subject": "Maso da San Friano",
    "src": "At which city did Maso da San Friano pass away?",
    "rephrase": "In which city did Maso da San Friano's life come to an end?",
    "old": "Florence",
    "alt": "Vienna",
    "loc": "At which city did Lina Cavalieri pass away?",
    "loc_ans": "Florence",
    "port": "What cultural aspect is Vienna known for that might have influenced Maso da San Friano's work?",
    "port_ans": "Art and music"
},
"th": {
    "case_id": 46,
    "subject": "Maso da San Friano",
    "src": "เมืองไหนที่ Maso da San Friano เสียชีวิต?",
    "rephrase": "ในเมืองไหนที่ชีวิตของ Maso da San Friano สิ้นสุดลง?",
    "old": "ฟลอเรนซ์",
    "alt": "เวียนนา",
    "loc": "เมืองไหนที่ Lina Cavalieri เสียชีวิต?",
    "loc_ans": "ฟลอเรนซ์",
    "port": "ด้านวัฒนธรรมอะไรที่เวียนนามีชื่อเสียงซึ่งอาจมีอิทธิพลต่อผลงานของ Maso da San Friano?",
    "port_ans": "ศิลปะและดนตรี"
}
```

WikiFactDiff

```
"en": {
    "case_id": 27,
    "subject": "James McCarthy",
    "src": "For which team did James McCarthy play?",
    "rephrase": "For which team does James McCarthy play?",
    "old": "Crystal Palace F.C.",
    "alt": "Celtic F.C.",
    "loc": "Which team did Shane Long play for?",
    "loc_ans": "Southampton F.C.",
    "port": "Who is the coach of the team that James McCarthy played for?",
    "port_ans": "Yiannick ferrera"
},
"th": {
    "case_id": 27,
    "subject": "James McCarthy",
    "src": "เจมส์ แม็คคาร์ธี เล่นให้กับทีมไหน?",
    "rephrase": "เจมส์ แม็คคาร์ธี เล่นให้กับทีมไหน?",
    "old": "คริสตัล พาเลซ เอฟ.ซี.",
    "alt": "เซลติก เอฟ.ซี.",
    "loc": "เชน ลอง เล่นให้กับทีมไหน?",
    "loc_ans": "เซาแธมป์ตัน เอฟ.ซี.",
    "port": "ใครคือโค้ชของทีมที่เจมส์ แม็คคาร์ธี เล่นให้?",
    "port_ans": "ยานนิค เฟอเรร่า"
}
```

Figure 10: Data Item Examples of BMIKE-53.

zsRe

Llama3.2-3B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	53.2	32.6	62.0	36.1	44.8	49.4	46.1	68.9	60.4	27.6	63.8	44.5	31.4	39.4	57.0	58.3	29.0	69.2	47.2	41.8	45.8	46.6	48.5	56.4	63.3	45.3	74.0	43.9	37.1	49.5	34.0	55.9	53.0	53.4	65.8	44.9	59.4	41.2	59.2	50.3	62.0	49.0	59.5	45.3	67.7	49.2	38.4	59.8	56.4	36.6	57.6	36.9	94.3	50.2
rel_em	44.7	22.5	55.3	30.0	28.7	15.1	37.6	61.2	54.2	20.0	59.0	38.2	20.7	32.4	49.7	49.7	20.5	63.9	38.9	35.4	37.0	19.2	33.1	49.7	53.8	17.6	69.7	37.8	23.3	17.0	19.0	46.0	44.9	45.2	59.0	38.1	52.8	33.9	51.4	35.9	55.8	43.5	52.6	36.0	63.0	15.5	27.5	53.0	43.2	24.6	50.3	28.4	93.1	39.5
gen_f1	47.6	33.5	59.1	35.4	41.8	48.7	44.1	67.3	59.2	37.6	62.5	45.7	31.9	34.0	53.9	56.7	28.1	65.8	45.4	41.4	45.7	45.0	47.5	54.7	61.2	44.7	72.2	44.6	35.9	48.7	32.9	54.7	52.8	52.2	64.5	44.9	59.3	40.1	58.0	49.7	59.9	46.6	56.9	44.6	66.0	49.1	38.5	56.7	55.5	35.1	55.2	36.7	87.3	49.0
gen_em	39.8	22.8	52.1	30.0	26.1	14.4	35.3	59.1	53.0	30.0	57.5	39.2	21.8	27.9	46.8	47.9	19.8	60.7	38.1	34.9	36.1	18.2	32.8	48.7	52.2	16.0	67.1	38.2	22.2	15.9	18.8	44.5	44.3	44.1	57.6	39.0	52.6	33.0	50.1	35.4	54.0	41.2	49.9	35.2	61.6	14.8	28.8	49.9	42.4	23.5	48.0	27.1	84.1	38.5
loc_f1	4.3	2.5	4.3	14.4	3.2	16.5	4.3	3.6	4.3	0.2	6.0	5.6	2.6	4.8	3.3	2.7	1.8	4.7	5.6	1.4	4.5	10.1	5.2	4.1	4.6	18.8	7.2	5.9	5.1	24.4	5.6	2.5	3.1	2.9	5.2	6.3	5.1	5.8	4.5	3.8	3.6	2.4	1.7	3.0	6.3	19.1	4.0	5.2	3.6	3.8	5.0	4.9	12.9	5.6
loc_em	1.8	0.5	1.9	10.0	0.9	0.3	1.9	1.5	1.6	0.0	3.2	2.7	0.5	3.0	1.8	1.1	0.1	2.7	2.7	0.1	1.9	0.1	1.9	2.3	1.5	0.4	4.5	3.5	0.4	0.4	1.2	1.2	1.1	0.7	3.0	3.0	2.3	3.0	2.6	0.8	1.6	1.4	0.3	1.6	3.5	0.7	1.4	2.6	1.2	0.4	2.2	0.7	7.3	1.8
port_f1	3.5	3.4	5.5	0.0	4.0	19.2	3.4	4.1	4.2	1.8	3.5	2.9	3.1	3.6	3.0	3.6	4.0	3.7	4.1	3.0	4.7	14.6	5.5	3.4	5.7	19.9	6.0	3.3	4.3	24.7	3.1	3.4	3.3	3.1	4.5	3.9	4.2	3.5	4.0	6.4	4.2	2.7	3.4	2.5	5.2	22.6	3.4	4.6	4.6	3.9	4.5	5.0	24.1	5.4
port_em	0.8	0.5	2.6	0.0	0.7	0.3	0.9	1.4	1.5	0.0	0.9	0.8	0.1	0.9	0.4	1.1	0.9	1.1	1.2	1.1	1.6	0.7	1.8	1.2	2.6	0.7	2.4	1.2	0.4	0.4	0.7	0.9	0.7	0.4	1.4	1.1	1.5	1.2	1.8	1.9	1.5	1.1	0.7	0.3	2.2	0.4	0.5	1.6	1.5	0.4	2.0	1.6	17.1	1.1

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	76.5	57.5	69.6	62.6	65.1	57.7	78.5	85.1	80.9	73.2	82.5	85.4	52.3	80.0	77.0	71.2	50.7	78.7	81.2	68.7	73.2	64.9	69.6	75.5	77.5	58.4	84.6	80.3	72.4	59.4	62.4	65.1	69.5	72.4	79.5	81.4	79.1	77.9	81.8	71.0	78.3	76.9	73.0	60.4	83.1	53.4	59.8	77.8	71.0	56.0	76.4	65.7	97.1	71.6
rel_em	70.3	41.9	58.8	46.3	48.3	24.2	70.8	79.1	75.2	67.7	77.5	81.3	40.0	73.2	69.9	62.6	37.4	73.2	72.3	64.1	64.9	35.7	51.5	69.0	70.4	25.8	80.2	73.4	58.1	21.6	45.9	56.1	60.6	64.6	72.9	77.0	73.1	71.6	73.4	57.9	72.6	70.8	61.5	51.2	77.7	17.7	46.8	70.6	59.1	36.7	70.4	53.6	96.4	60.1
gen_f1	77.0	57.7	68.5	62.6	65.6	56.9	77.4	85.0	81.0	73.3	81.8	85.2	51.3	78.9	76.4	71.3	50.1	78.6	80.3	68.2	72.6	64.7	69.6	75.4	77.2	58.1	85.2	80.1	71.9	59.7	62.4	65.1	69.5	72.4	78.8	81.1	77.9	78.4	81.5	70.4	78.6	76.7	72.6	59.8	82.5	53.4	59.2	77.0	71.0	54.9	74.8	65.4	96.8	71.2
gen_em	71.1	42.4	57.5	46.2	49.0	23.8	69.7	78.6	75.6	68.1	76.5	81.2	39.2	72.1	68.8	62.9	36.7	72.8	71.1	63.5	63.3	35.4	51.1	68.8	69.3	25.4	80.9	73.6	57.6	21.6	46.2	56.0	60.5	64.3	72.2	76.6	71.8	72.4	73.6	57.3	72.8	70.9	61.4	50.5	77.4	17.9	46.2	69.9	58.9	35.8	68.7	52.8	96.0	59.8
loc_f1	9.1	3.9	4.5	4.0	4.9	17.6	8.6	4.4	9.1	2.1	10.0	10.6	4.6	9.9	6.7	1.7	3.3	7.0	9.9	3.5	8.2	11.6	5.1	8.7	7.0	19.0	8.0	9.0	6.1	24.2	5.5	2.7	5.5	4.8	8.2	10.9	8.5	12.1	9.7	6.0	7.2	6.5	5.3	4.4	9.7	19.7	4.4	6.4	4.8	4.7	8.3	6.9	18.6	7.8
loc_em	5.7	0.9	1.9	1.6	2.3	0.9	4.7	2.4	5.7	1.2	6.2	7.3	1.6	5.4	4.4	0.7	1.2	4.3	5.3	1.6	5.3	0.7	2.0	5.8	3.8	0.4	5.0	5.4	0.9	0.8	1.4	1.6	2.2	2.3	5.1	6.6	5.3	7.9	6.5	2.8	4.2	4.2	2.7	1.9	6.4	0.7	1.5	3.9	2.7	0.9	3.8	1.9	12.8	3.3
port_f1	17.0	9.5	8.8	11.0	14.5	19.7	22.0	12.2	20.6	7.1	19.2	21.9	11.3	20.3	11.5	6.0	10.3	15.9	23.4	7.6	20.3	19.2	10.7	16.3	15.3	23.7	12.6	24.6	11.7	27.5	7.5	8.2	8.6	7.9	14.7	20.4	17.9	22.1	17.5	16.5	14.0	15.2	11.5	8.4	18.3	23.9	7.9	15.4	15.1	7.3	13.4	15.3	38.6	15.0
port_em	9.8	4.6	4.0	5.1	6.7	1.1	14.9	7.9	14.3	3.1	13.2	14.9	4.6	14.3	6.1	2.4	4.0	9.3	15.3	4.0	13.7	2.4	4.5	10.4	8.9	1.2	7.8	17.4	3.8	0.9	3.0	4.1	4.3	3.8	9.2	13.2	12.6	14.9	11.2	10.0	8.8	9.4	6.8	4.3	12.8	0.5	2.6	10.0	8.5	1.5	8.8	7.7	29.9	7.7

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	74.6	57.5	68.5	60.3	64.1	56.0	80.3	83.7	80.0	75.7	82.4	85.1	51.8	76.6	77.4	74.0	52.3	77.2	80.4	68.8	79.9	68.7	66.8	76.6	77.0	55.3	82.8	78.7	74.2	54.8	58.7	63.1	70.8	65.5	77.9	80.7	80.2	75.8	80.5	67.8	77.8	77.5	71.9	36.7	84.1	60.3	54.1	79.0	62.8	55.4	75.5	71.0	99.3	70.5
rel_em	68.0	42.4	58.8	47.1	49.7	20.2	72.3	77.9	75.1	69.3	76.7	81.3	42.5	69.2	69.3	64.6	39.3	70.9	70.7	63.3	72.3	41.3	47.3	70.0	68.5	21.3	77.8	71.7	50.9	21.1	47.6	54.3	62.6	56.8	70.2	75.9	74.5	68.9	71.7	53.7	71.4	72.1	60.1	28.4	78.8	20.6	44.6	70.9	53.0	35.9	69.0	53.2	98.9	58.9
gen_f1	74.2	58.0	68.5	59.6	63.9	57.2	80.0	83.4	79.8	75.5	82.5	84.9	52.2	76.2	77.4	73.7	52.3	77.1	80.1	68.9	79.1	69.3	66.7	75.8	76.7	55.2	83.3	78.9	74.2	54.8	58.1	63.1	70.9	65.9	78.1	80.9	80.1	75.6	80.3	68.2	78.0	76.8	71.9	36.8	84.1	60.5	53.7	79.0	62.5	56.0	75.1	70.8	99.4	70.5
gen_em	67.4	43.1	58.7	46.6	49.1	21.4	72.1	77.9	74.7	69.0	76.9	81.2	43.2	68.8	69.6	64.5	39.6	70.9	70.3	63.5	71.7	41.8	47.0	69.2	67.8	21.0	78.2	72.0	50.9	20.9	47.0	54.4	63.0	56.6	71.0	76.0	74.1	68.9	71.7	54.1	71.7	71.4	59.9	27.6	78.8	21.6	44.3	70.7	53.6	36.3	68.9	53.5	99.1	58.9
loc_f1	10.9	4.3	7.3	4.8	4.6	17.0	10.9	6.7	11.9	5.1	12.3	11.9	5.4	12.0	8.5	5.2	4.3	7.0	11.1	4.5	9.1	13.8	4.9	7.6	9.6	17.0	11.8	11.4	6.0	23.9	5.1	2.6	5.4	5.7	11.2	12.8	8.4	12.9	10.9	7.7	9.0	9.1	6.3	2.5	10.3	19.2	6.7	9.2	4.5	4.2	10.9	7.3	18.4	8.9
loc_em	7.0	0.9	3.9	1.6	2.0	0.9	6.6	4.7	7.7	3.1	8.3	7.5	2.2	6.7	5.3	3.0	1.4	4.9	6.6	2.6	5.5	1.9	1.4	4.3	5.5	0.1	7.6	7.3	1.5	0.7	1.5	2.0	2.8	2.7	7.3	8.6	5.1	7.8	7.6	4.0	5.8	5.7	3.4	0.7	6.9	0.5	2.8	5.6	2.6	1.1	6.1	1.5	13.2	4.1
port_f1	19.6	12.5	14.6	14.9	16.3	23.2	25.5	9.9	21.6	8.9	21.0	20.4	13.0	26.6	13.0	7.0	13.0	10.7	24.8	9.7	23.4	23.0	13.1	15.2	14.2	23.5	18.6	24.7	16.9	27.9	9.1	9.2	8.6	8.7	16.0	22.0	15.5	24.0	22.9	19.9	16.5	14.4	12.1	5.6	18.3	24.1	8.9	15.5	15.0	9.2	15.5	16.8	34.6	16.4
port_em	12.4	5.4	7.8	7.1	8.5	1.6	17.8	5.7	14.4	4.2	14.1	14.4	6.3	18.4	7.0	2.6	6.3	5.8	17.6	5.1	15.2	3.3	5.5	9.8	8.4	1.1	11.3	17.0	7.0	1.1	3.8	5.1	4.3	3.3	9.3	14.8	9.7	15.9	15.7	11.4	10.9	8.6	5.9	2.6	12.3	0.7	3.5	9.8	10.0	2.4	9.3	8.2	26.1	8.5

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	77.6	59.8	73.7	64.3	66.0	63.1	82.0	86.4	79.1	75.2	84.5	84.7	52.8	78.6	78.2	72.5	52.8	78.1	80.5	72.2	77.7	67.5	67.9	76.1	78.1	29.0	82.5	80.0	74.4	57.6	65.2	62.6	68.3	53.4	82.6	80.6	79.1	78.2	82.0	67.8	78.3	78.9	72.3	43.4	80.8	61.2	57.6	80.2	68.6	55.8	75.4	64.0	99.4	70.9
rel_em	70.7	44.0	64.7	47.9	48.9	26.1	73.9	81.6	74.2	69.0	79.0	80.2	44.2	71.5	70.0	65.1	39.2	72.7	70.9	66.0	68.5	39.1	48.5	70.0	69.0	21.3	77.8	72.8	55.9	22.0	49.4	54.1	58.4	41.7	75.2	75.2	71.9	71.5	72.9	53.7	72.9	73.9	59.9	36.7	74.6	21.0	48.5	72.4	55.1	35.8	71.3	48.7	99.1	59.6
gen_f1	77.1	59.9	73.3	64.2	66.0	62.9	82.1	86.4	79.5	75.8	84.0	84.9	53.4	78.3	78.3	72.6	53.1	78.0	80.4	72.3	77.5	67.0	68.1	75.9	78.1	29.1	82.6	80.2	73.8	57.0	65.3	62.7	68.5	53.3	82.0	80.8	78.7	78.3	81.9	67.4	78.2	78.4	73.0	44.6	80.9	61.1	57.3	79.4	68.2	56.1	75.8	64.2	99.4	70.9
gen_em	70.3	44.0	64.3	48.5	48.6	25.7	73.9	81.6	73.9	69.3	78.3	80.4	44.4	71.2	70.4	65.6	38.9	72.4	70.8	66.0	68.1	38.6	48.5	69.3	68.7	21.1	77.6	73.1	55.1	21.6	49.8	54.3	59.0	41.1	74.9	75.7	71.5	71.5	73.0	53.4	72.5	73.5	60.5	37.1	74.7	21.0	48.5	71.8	54.7	35.9	71.3	48.8	99.1	59.5
loc_f1	13.6	7.3	12.3	6.6	8.5	20.3	14.8	12.8	14.0	10.8	14.6	14.3	8.5	15.3	11.6	11.3	5.2	10.8	15.7	8.7	13.5	16.2	9.5	11.9	13.6	20.3	15.3	15.5	10.4	22.5	7.2	9.3	8.4	6.7	13.8	15.8	12.9	14.8	15.2	10.1	12.1	11.4	11.2	6.0	15.5	21.5	8.9	14.0	8.4	7.2	12.7	7.7	20.6	12.2
loc_em	9.7	3.0	8.6	2.2	4.2	1.6	10.1	9.0	9.8	7.4	10.0	9.7	4.9	10.5	7.9	7.9	2.2	6.5	10.6	6.1	9.6	2.6	4.3	7.8	8.0	0.9	10.7	10.8	2.7	1.9	3.0	6.9	3.8	2.8	9.2	11.6	8.6	9.8	11.2	5.9	8.4	7.8	6.6	3.1	11.5	1.2	4.9	9.4	4.7	2.0	9.0	1.8	14.3	6.6
port_f1	27.8	12.9	22.2	18.8	22.7	27.9	33.1	18.2	31.7	13.8	30.8	30.5	14.8	35.4	22.1	12.6	17.1	24.7	35.0	10.7	30.3	22.3	18.5	23.2	20.7	25.9	29.4	32.0	19.6	31.3	12.5	16.5	16.6	12.5	22.4	33.1	24.6	32.8	32.2	26.4	23.7	23.3	15.6	9.3	27.7	28.4	14.4	25.5	23.1	13.1	24.3	20.8	50.8	23.0
port_em	19.7	6.7	14.8	10.1	13.1	3.2	25.2	13.1	22.9	8.1	23.0	21.5	8.2	27.3	14.7	6.9	8.6	17.4	25.8	5.9	21.4	5.1	8.1	15.9	12.8	3.0	22.4	23.6	7.9	2.8	4.7	10.9	10.5	6.4	15.9	24.9	17.8	24.6	23.6	17.9	16.3	16.2	9.3	5.0	19.9	1.4	6.9	17.6	15.2	3.8	18.2	11.5	41.2	13.8

Llama3.1-8B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	76.0	51.6	70.7	50.6	57.3	41.7	68.6	86.1	77.2	70.3	81.0	75.6	47.4	73.6	71.9	65.3	39.8	79.0	76.4	67.8	65.9	62.1	63.9	74.3	72.8	43.7	86.4	74.2	50.0	47.1	50.5	65.7	71.0	68.4	77.8	79.7	74.1	77.8	77.0	62.0	74.5	70.4	71.9	57.5	81.7	53.3	47.3	66.1	56.7	43.4	69.6	43.0	96.9	65.5
rel_em	70.5	41.9	65.0	38.0	45.8	17.1	60.2	80.8	71.3	64.5	75.4	68.5	40.2	65.8	65.1	56.8	29.5	73.1	68.2	63.7	56.8	35.9	49.2	68.2	64.8	23.0	84.2	68.0	37.7	20.5	40.4	57.4	64.2	61.0	73.7	74.7	69.0	72.3	69.8	52.5	69.1	64.8	66.1	47.7	77.3	22.8	38.2	59.4	48.9	32.0	63.5	35.3	96.1	56.3
gen_f1	72.4	49.4	69.0	50.5	55.6	41.5	66.2	84.2	74.8	67.7	79.1	73.5	45.8	72.0	70.1	65.3	38.1	76.5	74.3	66.7	64.3	61.9	63.0	71.8	70.7	43.8	84.6	72.7	48.3	47.7	49.9	62.7	68.9	67.1	77.0	77.3	73.3	75.4	76.4	60.3	73.0	68.7	71.3	56.9	80.0	53.8	47.1	65.1	56.2	41.9	67.2	41.5	87.7	64.1
gen_em	66.9	39.6	63.4	38.2	43.9	16.4	56.9	78.1	68.4	61.1	73.5	66.1	38.9	63.9	63.3	56.4	28.1	70.5	66.9	63.3	55.6	35.2	47.8	65.4	62.5	23.3	81.7	66.5	35.3	20.2	39.8	54.4	62.2	59.5	72.5	71.7	68.3	69.3	69.1	50.7	67.4	64.0	65.3	46.8	75.7	22.9	38.2	58.4	47.8	30.9	60.7	33.2	85.3	54.7
loc_f1	9.6	4.7	10.2	4.9	6.8	17.8	11.3	7.2	10.3	6.7	13.0	12.3	5.9	11.8	6.4	6.3	4.9	8.5	11.5	4.3	9.7	14.3	7.2	10.0	10.3	19.1	12.3	11.8	7.3	25.4	7.5	6.1	6.5	6.5	10.3	12.9	10.4	12.9	11.1	7.4	9.6	8.9	5.5	7.0	12.7	20.0	8.4	11.7	7.2	5.9	10.5	6.5	19.7	9.8
loc_em	5.3	1.1	6.6	1.6	3.0	0.9	7.4	3.9	6.2	3.9	8.9	8.2	2.0	7.4	3.6	3.6	2.0	5.5	6.7	2.4	6.2	1.4	2.7	6.2	5.7	1.2	8.0	7.4	1.9	1.8	1.8	3.8	2.7	3.0	6.5	8.5	6.8	8.5	7.2	2.7	5.9	6.1	2.8	4.0	8.8	1.4	4.7	7.2	3.3	2.2	5.8	1.6	12.7	4.6
port_f1	10.2	6.7	7.4	8.4	9.4	21.3	9.6	8.2	11.5	6.1	10.4	12.0	5.8	12.2	5.3	5.0	7.2	8.8	12.3	5.2	9.7	20.1	10.0	9.8	9.2	18.8	8.8	13.7	9.2	25.4	6.7	6.5	5.2	6.0	7.7	13.9	9.5	13.4	9.2	13.2	8.5	7.5	6.7	8.1	10.9	23.0	5.5	9.3	10.7	5.9	11.1	6.9	25.8	10.1
port_em	5.1	3.1	4.0	4.2	5.1	2.6	4.9	4.2	7.1	2.8	5.7	6.5	1.1	6.6	1.9	2.0	2.8	4.7	6.6	2.6	4.7	4.3	4.5	5.4	4.9	1.6	5.3	8.6	3.6	0.7	2.7	2.2	2.3	2.6	4.3	8.6	5.4	8.3	4.7	7.4	4.7	3.8	3.5	4.3	5.7	0.8	1.9	5.3	6.1	1.6	7.3	3.0	18.2	4.3

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	81.3	66.8	42.1	66.5	72.7	63.3	80.4	86.3	85.7	79.4	83.2	86.3	59.5	82.4	81.2	77.1	60.4	81.8	84.1	76.9	76.6	74.7	76.4	79.7	81.1	59.0	87.0	81.6	81.2	58.7	72.8	64.1	71.1	73.7	80.7	84.9	80.6	82.7	83.8	74.8	81.7	81.6	77.4	67.4	84.8	60.5	65.3	81.5	73.0	62.7	78.8	77.2	98.3	75.3
rel_em	75.9	52.6	32.7	50.9	57.2	27.9	73.1	80.0	80.8	75.0	77.8	82.1	49.0	75.8	74.6	68.2	46.2	76.5	75.9	73.0	68.6	46.2	58.0	73.4	73.5	28.3	84.2	75.5	69.6	24.3	58.4	56.7	61.0	66.2	74.7	80.7	75.4	76.7	76.8	63.3	76.6	76.3	67.2	54.9	79.9	23.1	53.0	73.9	59.8	43.6	72.6	62.3	98.0	64.2
gen_f1	80.7	66.9	39.6	66.3	72.6	62.6	80.2	86.5	85.4	79.4	82.6	86.5	58.8	82.2	80.9	77.2	59.3	81.2	83.8	76.7	75.9	74.2	75.8	79.4	80.8	58.5	86.0	80.8	81.0	58.7	72.4	63.8	71.5	73.4	80.4	84.7	80.8	81.6	83.7	74.4	81.1	81.5	77.1	66.6	83.8	60.2	65.8	81.4	73.4	61.9	79.3	75.9	97.9	74.9
gen_em	75.0	52.4	29.9	51.3	56.7	27.5	73.0	80.2	80.1	74.7	76.7	82.5	48.9	75.6	74.6	68.0	45.2	75.8	75.6	72.5	67.7	46.2	57.6	73.1	73.2	27.9	82.8	74.7	68.8	23.9	57.8	56.3	61.8	65.5	74.4	80.2	75.4	75.2	76.9	62.7	76.0	76.3	66.9	53.9	78.4	22.2	53.6	73.6	60.6	42.5	73.1	61.1	97.4	63.8
loc_f1	16.1	6.8	9.5	8.4	9.9	21.0	15.4	12.8	15.2	11.3	17.1	18.3	8.4	16.2	12.6	10.3	6.9	13.4	17.7	8.3	14.7	15.6	8.6	14.7	14.8	21.8	15.8	15.8	11.8	26.4	9.4	6.3	10.7	10.6	15.6	19.7	13.9	17.0	16.5	10.9	12.8	13.3	11.9	9.9	16.5	20.8	9.7	12.4	9.8	7.3	12.7	11.7	25.1	13.4
loc_em	10.2	2.6	5.3	3.8	5.0	1.9	10.4	8.3	10.8	6.7	12.1	13.1	3.9	10.8	8.2	6.9	2.8	9.0	12.1	5.1	10.1	1.9	3.0	10.1	9.0	1.9	10.7	10.5	3.5	2.3	3.1	4.6	5.7	6.4	10.5	13.6	9.5	11.0	11.9	5.9	9.2	8.9	7.2	5.5	11.5	1.1	5.1	7.9	5.1	2.2	8.1	3.6	18.0	7.1
port_f1	25.1	14.6	7.3	17.9	20.5	25.3	25.4	17.6	24.9	16.0	27.6	28.4	15.5	25.2	17.5	10.3	16.8	21.3	26.2	15.3	22.3	28.2	19.9	23.3	19.2	27.2	21.9	29.1	21.6	29.7	14.9	10.2	16.2	13.2	20.5	29.2	24.3	28.4	24.5	22.1	19.6	19.3	18.6	17.1	26.2	26.0	11.4	22.3	21.8	14.5	19.1	21.9	38.9	20.8
port_em	17.0	7.4	3.4	9.3	10.6	3.8	18.2	12.5	17.4	10.9	20.1	20.2	6.7	18.2	11.6	5.3	8.9	14.8	18.4	10.9	15.8	7.0	8.8	16.7	11.6	3.1	15.2	20.9	12.3	1.6	7.2	5.7	9.3	7.0	14.2	20.6	18.5	21.4	17.7	13.3	13.8	12.5	12.7	10.8	18.4	1.5	5.4	15.7	13.5	5.3	13.8	11.6	29.7	12.1

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	78.5	65.3	70.0	66.9	69.9	57.9	82.1	84.3	81.5	78.8	83.2	84.6	55.4	77.9	82.3	78.1	56.7	79.3	80.3	76.7	80.9	72.4	73.7	80.5	80.8	56.3	83.3	78.7	80.8	55.2	64.5	69.0	73.2	71.7	84.3	82.2	81.1	79.1	81.4	75.6	80.6	80.3	74.2	60.1	82.5	63.1	60.0	81.5	70.2	60.7	79.2	76.8	98.7	74.3
rel_em	72.3	51.4	59.0	52.4	56.1	20.3	73.5	77.9	76.9	73.6	77.1	80.9	47.9	70.0	76.6	66.9	43.3	73.8	70.9	72.4	73.5	45.7	56.0	74.6	72.2	25.6	78.0	70.8	65.8	23.9	54.5	60.9	64.0	64.1	77.9	76.2	75.8	72.0	73.3	63.7	74.1	75.5	62.5	48.8	76.6	24.7	52.4	73.7	60.2	40.8	72.0	62.4	98.3	63.2
gen_f1	79.0	64.3	69.7	66.0	69.5	57.5	81.8	84.7	81.5	78.7	82.4	84.4	55.5	76.9	81.2	78.2	57.2	79.1	78.9	76.1	80.2	72.4	73.6	79.8	80.5	56.5	82.6	78.7	80.6	54.7	63.5	68.8	72.0	72.1	83.4	82.0	80.6	78.5	81.4	75.1	80.5	80.4	74.8	58.9	82.8	63.2	59.1	81.9	69.9	61.5	79.3	76.3	98.5	74.0
gen_em	72.7	49.9	58.6	52.1	55.5	19.9	73.4	78.6	76.3	73.5	76.3	80.5	48.5	69.0	75.5	66.9	43.9	73.4	69.9	71.7	72.1	46.3	55.4	73.6	71.6	25.6	77.4	71.2	65.8	23.6	53.6	60.6	62.5	64.7	77.1	76.2	75.3	71.6	73.1	62.9	74.0	75.6	62.9	48.0	76.9	24.9	51.8	74.1	60.0	41.6	72.0	62.4	98.0	62.9
loc_f1	17.3	7.8	14.0	9.9	10.0	21.9	18.0	15.5	15.9	14.2	19.2	20.3	10.3	18.4	14.6	15.3	8.3	15.2	19.2	10.2	18.6	17.5	10.8	16.8	17.7	22.3	19.1	18.4	15.3	26.7	10.7	10.2	12.7	12.0	18.3	20.7	15.1	19.3	19.0	13.0	17.1	14.9	14.6	9.1	18.3	21.1	12.0	16.9	10.7	7.9	17.4	14.5	25.7	15.5
loc_em	11.3	2.4	8.8	4.2	4.9	2.0	11.8	10.8	10.1	9.6	13.6	14.3	5.3	11.8	9.0	10.2	4.0	10.6	12.7	6.9	12.3	2.7	4.7	11.0	11.1	2.3	12.8	12.5	6.3	2.3	4.5	7.2	6.8	6.5	12.5	14.2	10.1	12.1	13.1	7.3	12.0	10.0	9.5	5.0	11.9	0.8	6.6	11.3	5.1	2.4	11.2	5.4	18.4	8.4
port_f1	28.2	18.6	19.8	22.6	23.5	29.6	35.3	18.9	29.9	20.3	30.1	27.6	16.9	36.5	22.1	15.5	20.6	23.5	38.2	16.4	33.8	33.2	24.1	23.1	27.4	27.5	30.6	34.7	28.6	30.8	15.9	12.4	21.3	17.3	26.9	31.3	23.9	35.1	31.0	28.9	22.6	22.5	21.8	18.2	27.8	27.3	16.7	26.6	23.7	16.1	26.4	27.5	44.9	25.2
port_em	19.5	9.3	13.2	12.3	13.3	4.9	25.8	13.6	22.3	14.0	23.0	19.1	8.9	26.5	14.3	8.6	11.0	15.5	28.8	11.4	24.8	9.1	11.2	16.3	18.5	4.0	22.9	25.4	15.1	3.1	8.6	6.8	13.5	9.3	19.3	22.8	17.0	26.0	22.5	19.1	15.9	15.6	14.0	10.9	20.0	1.8	8.6	19.1	14.5	5.3	18.2	16.3	36.7	15.2

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	79.0	66.3	76.0	67.2	70.1	63.7	83.6	85.6	82.8	80.4	83.7	84.2	60.6	80.1	83.2	76.0	58.3	80.1	82.1	79.0	81.7	74.1	75.8	80.5	79.4	42.0	84.0	80.8	81.2	59.7	70.5	65.4	71.8	60.9	85.5	83.4	79.7	79.3	82.3	72.7	81.8	81.8	74.7	65.7	82.1	64.7	60.7	82.6	72.3	61.6	78.5	74.2	99.2	74.9
rel_em	71.9	52.5	65.7	52.4	54.1	25.4	75.5	78.9	78.2	75.2	77.5	80.4	50.5	72.4	77.5	65.6	44.0	75.0	73.1	74.2	73.6	46.3	57.4	74.4	70.4	25.6	78.7	73.0	69.9	24.7	55.7	57.9	61.3	50.1	79.3	77.6	74.1	71.9	74.6	59.1	77.4	77.0	62.5	55.5	75.4	26.7	52.9	75.2	59.5	41.2	72.2	60.0	98.8	63.7
gen_f1	78.8	66.9	76.2	67.2	70.0	63.6	83.4	85.4	82.3	80.5	83.9	84.8	60.0	80.3	83.0	76.7	57.7	79.5	82.2	78.3	81.2	73.6	76.4	80.1	79.2	42.2	83.6	80.6	81.3	59.1	70.2	65.3	73.0	61.1	85.7	83.1	79.3	79.2	82.4	72.4	82.2	81.4	75.0	64.7	81.9	64.8	60.2	82.6	72.5	61.6	78.8	74.1	99.2	74.8
gen_em	72.1	53.3	65.8	52.5	54.1	25.2	75.1	78.9	77.0	75.2	77.8	80.8	50.7	72.8	77.3	66.5	43.9	74.2	73.0	73.5	72.8	45.5	58.1	73.9	70.2	25.4	78.0	73.1	70.1	24.0	56.0	57.6	62.6	50.1	79.3	77.0	73.3	72.1	74.4	58.8	77.5	76.4	62.9	54.2	75.7	27.2	52.5	75.3	60.0	40.9	72.6	60.2	98.8	63.6
loc_f1	19.3	8.7	13.3	8.3	11.1	20.7	19.2	17.7	17.5	15.7	20.3	21.3	10.8	20.0	15.8	16.6	7.0	16.6	20.3	12.6	18.9	18.4	13.0	16.7	18.4	23.1	21.0	19.7	14.2	24.9	7.3	11.6	14.3	11.6	19.3	22.5	16.3	20.0	18.1	14.0	16.6	17.6	16.3	11.6	19.6	22.4	12.3	17.8	12.2	9.0	18.8	10.1	27.0	16.2
loc_em	13.6	3.8	9.7	2.4	5.4	2.4	13.5	12.4	11.4	10.8	14.7	15.5	5.8	13.9	10.9	11.6	3.8	11.3	14.1	8.9	12.9	4.4	6.8	11.3	12.3	3.1	15.4	13.3	5.4	3.1	4.3	8.2	8.4	5.8	13.6	16.3	10.9	13.2	12.6	8.1	11.5	12.3	10.4	7.0	13.4	1.4	7.7	12.1	6.9	3.4	12.8	3.6	19.7	9.4
port_f1	38.7	21.4	27.5	28.0	29.3	33.6	40.2	30.2	38.8	26.7	41.7	44.2	22.3	41.2	33.9	23.4	23.7	37.0	42.6	24.8	39.4	33.9	28.8	36.1	36.0	30.9	41.8	41.6	34.0	33.8	25.5	21.4	26.7	24.3	37.4	43.4	37.9	42.3	39.7	32.7	36.2	34.5	27.7	24.6	39.2	31.0	22.2	37.6	33.1	20.4	32.0	33.5	55.1	32.9
port_em	28.4	12.8	18.6	16.0	18.3	5.7	30.2	23.4	28.8	20.2	31.2	33.8	13.7	30.2	25.2	15.5	12.8	28.5	31.8	17.6	29.6	11.2	13.9	26.4	26.2	5.7	32.8	31.6	19.0	3.5	13.8	15.3	17.5	15.2	29.1	33.3	28.7	32.6	29.3	23.7	27.4	25.7	18.4	16.0	29.5	2.3	12.5	28.5	22.9	8.4	23.7	21.0	43.6	21.5

Table 9: Full Results on zsRE.

CounterFact

Llama3.2-3B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	27.4	30.5	56.1	45.5	42.9	66.8	26.9	57.7	45.1	33.0	47.4	28.0	25.9	28.4	44.5	45.2	39.6	53.6	33.4	25.5	24.2	62.0	44.6	49.7	58.9	54.9	58.9	30.8	44.9	67.2	37.0	42.1	44.0	43.7	52.8	32.4	42.4	30.4	52.3	54.7	40.2	37.5	40.8	45.2	58.4	59.6	38.0	48.2	54.9	31.7	44.5	32.8	93.7	43.5
rel_em	25.0	21.6	52.5	34.3	32.7	40.8	22.9	55.5	41.7	30.8	45.2	26.2	17.5	25.6	39.8	38.3	35.5	45.6	30.5	23.1	21.2	31.8	37.3	46.2	51.4	26.8	56.0	27.5	39.3	33.5	32.3	35.9	37.4	34.4	49.9	30.6	38.7	27.1	48.9	50.1	36.3	34.0	36.6	40.6	56.2	26.9	29.1	43.5	48.6	23.5	43.5	28.8	92.9	36.3
gen_f1	26.5	26.5	52.5	44.7	42.2	66.1	25.3	55.3	36.9	34.1	45.5	24.7	25.0	17.9	39.4	45.4	39.4	48.2	34.9	25.6	20.9	58.9	44.4	44.8	51.8	54.2	57.5	33.2	43.6	66.1	34.3	43.0	38.4	38.6	48.6	28.3	35.8	24.7	46.4	53.1	37.7	34.5	40.0	39.8	58.4	58.6	34.2	43.4	49.3	30.3	43.3	31.9	79.9	40.8
gen_em	24.8	18.6	49.1	34.6	32.3	38.8	22.0	53.2	33.8	31.7	43.7	23.4	16.3	15.5	35.1	38.6	36.1	41.2	32.4	23.3	18.4	30.2	36.8	41.3	45.3	25.8	54.6	30.6	38.8	29.6	30.2	37.9	31.2	30.2	45.4	26.6	32.4	22.2	43.7	48.7	34.5	31.7	36.1	35.5	56.6	27.1	25.2	38.5	43.1	22.4	42.5	27.2	78.2	33.9
loc_f1	5.6	3.9	7.8	6.1	6.0	21.8	4.4	6.0	6.0	3.8	7.8	5.8	4.2	3.0	4.5	5.5	5.3	6.5	5.8	4.2	6.5	16.4	6.3	4.5	8.4	24.5	8.5	4.9	6.9	37.0	4.8	2.5	4.3	4.3	6.6	7.1	6.7	4.3	4.5	7.8	5.0	3.0	3.8	3.0	10.0	26.3	6.2	6.4	7.8	4.8	8.5	6.3	32.9	7.5
loc_em	4.7	1.5	6.5	3.3	3.1	6.1	2.9	5.3	4.3	3.0	6.6	4.5	1.2	2.0	2.9	3.6	4.2	4.9	4.3	3.7	4.7	3.0	3.5	3.2	6.4	1.4	6.8	3.2	3.3	3.1	2.6	1.7	2.6	2.2	5.0	6.2	5.1	3.0	3.1	6.0	3.1	2.0	2.5	1.7	8.2	2.6	2.9	5.0	5.2	2.4	7.2	2.9	31.8	3.8
port_f1	2.3	4.0	6.0	4.7	4.2	20.9	1.9	3.9	4.1	1.6	3.3	1.4	3.6	2.7	3.3	4.1	4.4	3.5	3.1	4.9	3.3	15.6	7.2	3.4	5.6	21.6	6.1	2.4	5.9	28.1	3.2	2.0	3.3	3.4	4.7	2.7	4.1	2.2	3.0	7.7	2.6	2.3	2.6	2.2	5.1	24.7	4.2	4.3	5.5	4.9	4.8	5.2	24.8	5.6
port_em	0.7	0.5	2.4	1.4	1.2	2.0	0.5	1.8	1.9	0.5	1.4	0.5	0.3	1.1	0.9	0.5	1.1	1.2	1.5	2.0	1.1	0.7	2.3	1.5	1.8	0.8	3.0	1.3	1.8	0.9	1.7	0.2	0.5	0.3	2.1	1.5	1.5	0.9	1.6	4.2	1.0	0.6	0.5	0.3	2.4	1.5	1.7	1.7	2.5	1.1	2.5	1.8	17.7	1.3

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	71.5	72.5	62.6	66.2	64.2	82.0	62.7	93.2	59.2	76.6	84.1	71.2	66.9	72.7	55.7	51.8	69.3	56.9	77.1	58.4	66.6	89.0	77.5	62.1	60.4	76.8	83.3	78.0	87.7	77.6	25.3	37.2	55.8	57.5	79.0	66.2	74.1	74.7	80.7	87.0	46.6	71.7	70.3	41.2	72.7	73.2	74.1	63.4	68.1	56.9	58.8	85.6	84.6	68.3
rel_em	68.4	63.3	54.2	54.8	52.7	56.5	59.0	92.9	54.7	74.8	82.1	69.1	56.5	69.2	39.7	31.5	64.7	40.3	74.6	55.1	61.0	79.4	69.2	58.1	44.8	43.6	80.2	75.3	85.2	31.4	24.4	30.0	46.0	49.7	76.1	64.1	71.0	71.1	77.8	84.4	41.4	68.4	64.9	35.3	70.5	37.5	68.4	49.5	61.3	43.3	58.3	82.0	84.5	59.9
gen_f1	70.4	71.8	61.1	65.9	67.5	81.3	60.8	93.3	59.8	76.6	82.1	71.9	66.2	71.3	55.6	49.0	69.0	57.7	76.2	57.1	64.8	88.2	76.5	61.6	61.2	76.6	81.8	77.0	87.6	77.6	24.3	39.3	55.7	57.3	78.3	64.6	72.4	74.2	80.1	86.5	48.1	71.4	69.5	44.8	71.9	71.9	73.6	64.4	71.3	56.7	58.4	84.9	89.2	68.0
gen_em	67.4	63.2	52.6	54.0	57.3	56.5	57.0	93.0	55.1	74.9	79.8	69.9	55.6	67.7	39.8	26.7	64.3	43.7	73.5	53.8	59.8	79.1	68.1	57.6	46.9	43.8	78.8	74.4	85.3	33.5	23.8	32.6	45.2	49.7	75.4	62.7	69.1	70.5	77.4	83.8	43.8	68.1	64.6	38.3	69.5	35.3	68.0	52.2	63.5	43.5	58.0	81.1	88.8	59.8
loc_f1	7.9	2.5	2.2	4.1	3.6	16.2	6.5	6.2	5.5	2.5	11.4	10.2	3.1	9.0	2.0	1.0	2.6	3.1	8.2	2.4	7.9	13.8	5.1	5.0	4.2	21.9	4.2	11.8	3.0	32.8	0.9	0.7	3.1	4.8	2.3	13.4	6.4	7.0	6.1	5.7	3.8	7.3	5.8	1.5	10.5	24.1	2.6	2.1	4.2	2.2	3.1	5.3	24.5	6.6
loc_em	7.3	1.2	1.8	1.9	1.4	1.8	6.2	5.8	4.6	2.3	10.9	9.4	1.7	8.5	1.1	0.6	1.8	2.7	7.7	1.9	7.2	2.9	3.0	4.0	3.5	0.5	3.6	11.2	1.0	0.3	0.4	0.3	1.1	2.7	1.9	12.9	5.6	6.1	5.5	3.8	2.9	6.4	5.1	0.5	9.8	0.5	0.8	1.6	2.5	0.2	2.7	1.8	24.0	3.7
port_f1	17.5	14.9	14.9	17.1	17.9	26.5	13.6	13.3	18.5	9.8	17.8	21.0	14.2	21.1	9.2	5.1	12.8	12.7	20.9	9.5	16.6	27.4	17.2	15.3	12.7	26.6	14.6	20.3	20.8	32.4	2.7	3.9	8.2	10.8	11.8	23.6	16.7	19.2	18.2	27.1	13.5	12.4	9.2	5.0	24.6	29.5	12.9	18.5	22.3	11.1	12.2	22.3	39.1	16.3
port_em	12.6	9.2	9.0	9.6	8.9	6.1	8.2	8.9	12.6	6.2	12.7	15.9	7.9	15.0	4.2	0.8	7.3	7.0	15.0	4.9	11.9	8.8	8.3	9.2	5.7	2.9	9.6	14.7	10.3	1.8	1.9	0.8	2.3	5.2	8.2	18.5	12.6	14.3	13.8	19.6	8.0	7.3	5.2	1.7	19.1	4.7	7.3	10.8	15.3	4.2	9.0	14.2	30.6	9.0

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	73.2	73.4	58.6	73.7	80.5	82.3	83.5	83.1	75.7	72.2	81.9	90.3	71.2	77.0	56.9	60.8	68.8	52.1	79.4	56.3	72.1	89.4	79.3	58.4	54.7	76.1	75.6	80.0	87.5	75.7	79.9	56.2	58.2	53.9	72.7	82.6	77.4	78.9	77.1	59.4	59.1	58.2	66.9	47.0	81.3	70.9	73.4	69.2	57.0	56.7	59.7	86.5	99.1	70.8
rel_em	69.5	67.3	51.0	65.3	74.6	56.7	81.3	81.0	72.5	69.5	77.6	89.4	64.6	72.9	41.0	41.6	63.7	35.0	76.1	50.2	67.0	81.4	70.2	46.5	37.9	47.2	72.0	76.6	84.3	32.4	75.6	48.5	46.7	43.6	70.6	81.0	72.4	76.0	72.7	51.1	52.9	54.8	61.1	43.2	77.7	36.7	69.1	57.4	48.1	43.3	58.7	81.4	99.0	62.3
gen_f1	73.5	73.1	58.2	72.9	79.8	81.2	83.1	82.2	73.5	71.7	81.5	89.7	71.0	76.1	56.2	60.6	68.7	53.1	78.7	53.7	70.9	89.3	78.7	59.5	55.4	75.2	75.7	79.3	87.3	76.5	79.2	56.0	55.9	53.3	72.5	81.4	76.6	77.3	76.0	59.5	58.1	58.7	66.0	47.5	80.3	70.3	72.0	68.9	59.0	56.7	59.7	86.2	97.9	70.3
gen_em	69.9	67.0	51.1	64.7	73.6	55.1	81.0	80.0	70.2	69.1	77.3	88.8	64.9	71.9	39.5	41.9	64.0	38.5	75.6	48.0	65.8	80.9	69.6	49.0	39.1	46.5	72.2	75.9	84.2	34.9	74.9	48.5	43.7	42.6	70.3	79.7	71.7	74.2	71.6	51.3	52.5	55.0	60.1	43.4	76.8	35.4	67.6	57.1	49.5	43.1	58.7	81.2	97.9	61.9
loc_f1	2.3	2.1	0.3	2.7	6.3	15.5	6.0	2.7	4.1	1.1	10.7	9.5	4.9	6.9	1.9	1.0	2.3	2.7	2.4	0.5	5.8	11.6	3.4	2.7	1.5	21.2	2.6	6.0	4.9	32.0	2.1	0.8	3.8	1.9	3.7	5.2	2.0	4.5	5.5	2.5	2.1	2.7	2.2	1.3	8.4	23.5	2.5	1.3	2.5	2.2	1.7	6.4	5.6	5.1
loc_em	1.8	0.6	0.1	0.7	4.1	1.3	5.2	2.4	3.2	0.9	9.6	8.2	3.0	6.3	1.2	0.6	1.7	2.3	1.9	0.1	5.0	0.6	1.3	1.4	1.0	0.0	1.7	5.1	3.2	0.1	0.7	0.6	1.2	0.3	2.6	4.8	1.1	3.8	3.9	0.8	1.2	1.6	1.7	0.6	6.8	0.2	0.8	0.9	1.1	0.1	0.9	3.0	5.0	2.2
port_f1	13.4	13.8	9.3	14.7	19.6	25.8	16.0	9.9	18.7	6.7	16.6	18.1	17.4	12.9	10.3	5.8	11.1	8.8	11.7	6.6	13.3	23.0	13.5	11.3	10.0	26.8	9.3	13.2	20.0	31.3	12.6	3.6	7.8	8.4	6.9	14.5	9.4	15.3	16.0	16.4	14.2	9.0	6.9	8.1	17.6	29.0	11.9	13.8	16.3	10.8	7.3	20.6	25.1	13.7
port_em	9.1	7.9	4.2	8.6	12.7	5.5	11.7	5.1	12.6	3.3	11.2	12.9	10.7	8.6	4.6	0.9	6.3	4.3	8.3	2.2	8.2	6.7	5.9	5.7	3.0	4.2	4.9	10.4	9.9	2.1	6.2	1.2	1.8	2.5	3.9	11.1	6.0	11.2	11.6	9.8	7.7	4.5	2.2	4.0	12.9	4.3	6.1	8.2	10.2	4.4	5.0	12.2	19.9	6.9

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	68.0	59.5	62.8	65.3	64.4	80.7	79.1	65.0	56.7	70.1	79.4	88.1	79.1	78.3	58.8	59.1	66.1	53.1	72.0	59.0	66.7	89.0	77.9	55.3	67.5	74.0	80.2	67.3	83.9	76.5	77.3	49.8	51.0	51.9	72.5	81.9	52.3	45.7	78.4	61.6	55.1	58.0	65.3	44.7	78.0	73.3	76.5	77.2	66.1	59.2	69.1	65.8	99.3	67.6
rel_em	63.3	50.9	54.9	53.4	53.5	51.3	76.0	63.8	51.9	66.5	75.4	85.0	73.7	74.4	41.1	42.5	62.0	36.0	66.5	54.1	60.6	72.8	68.1	48.2	62.7	44.0	76.0	62.0	82.0	39.2	74.5	44.3	35.0	39.0	69.9	78.1	45.9	39.0	73.5	52.9	49.9	53.0	58.8	39.1	74.6	37.4	71.8	71.5	56.7	45.1	65.2	60.3	99.2	58.6
gen_f1	66.2	57.5	62.2	66.9	65.8	79.5	79.1	62.8	57.0	69.6	79.4	86.5	78.7	77.9	59.4	58.3	65.0	52.9	71.8	57.7	64.6	88.7	77.6	55.1	69.6	73.5	78.7	67.2	83.6	76.7	76.4	49.1	48.2	51.3	72.4	79.6	52.1	46.1	77.3	63.4	54.9	57.9	64.8	47.1	77.5	73.1	75.9	76.8	66.2	59.8	68.5	63.6	95.9	67.1
gen_em	61.9	49.3	53.9	55.1	55.8	48.8	75.9	61.8	52.2	66.2	75.3	83.3	73.6	73.8	42.0	40.7	60.7	37.7	66.4	52.8	58.5	73.9	67.6	48.0	64.7	43.7	74.7	61.9	81.5	40.1	73.9	43.7	31.5	38.1	70.0	76.0	45.9	39.8	72.4	55.1	50.6	53.0	58.4	40.2	73.7	36.8	71.4	71.7	55.6	46.1	64.5	58.6	95.7	58.2
loc_f1	41.8	16.1	28.0	18.8	24.0	35.8	39.1	33.9	36.3	23.6	46.6	49.9	29.5	49.3	29.1	23.6	14.4	29.5	42.4	11.5	43.5	28.0	23.0	34.5	30.5	29.1	40.4	44.9	34.1	43.0	10.4	30.5	20.4	16.4	35.5	49.6	36.6	43.1	42.9	28.6	29.3	31.8	33.9	16.3	45.3	29.4	31.0	34.1	31.2	7.1	35.1	31.4	55.7	31.6
loc_em	40.0	14.0	25.5	15.4	20.2	21.3	37.5	32.7	33.9	22.6	44.9	48.6	26.7	48.1	22.5	20.3	13.3	22.5	40.6	9.2	40.5	15.7	20.4	30.7	24.1	6.3	37.0	43.3	31.9	9.0	8.6	28.0	13.3	9.8	33.5	48.7	32.9	40.1	39.6	26.8	24.9	28.1	30.5	14.4	41.7	4.5	28.0	33.0	27.9	4.7	33.4	27.3	54.6	26.9
port_f1	32.8	24.2	18.2	28.4	33.6	34.7	38.8	30.3	33.4	22.6	41.3	44.5	27.9	42.7	28.8	21.6	20.3	32.4	42.4	17.6	40.4	34.1	26.4	32.2	32.3	34.4	39.2	39.8	35.8	38.5	24.2	21.3	21.8	15.5	31.4	43.0	33.8	40.3	36.8	38.3	29.6	24.6	26.0	13.1	44.3	33.5	19.7	36.8	33.7	17.1	31.4	36.9	60.2	31.2
port_em	24.1	16.9	10.7	18.6	23.1	9.0	29.9	23.8	26.3	16.1	34.0	35.1	19.9	35.2	20.8	12.6	10.9	22.4	33.6	11.1	32.1	11.2	11.5	24.3	23.6	7.4	31.6	32.3	21.1	6.4	12.3	14.2	12.4	8.0	23.9	34.8	27.9	31.8	28.1	30.1	20.8	17.1	17.9	8.5	36.5	4.9	12.6	28.3	23.8	8.2	24.8	23.9	50.7	20.9

Llama3.1-8B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	62.9	62.8	65.9	59.0	67.9	68.8	52.9	87.6	64.6	60.7	75.6	68.5	56.8	60.1	63.5	56.3	61.3	62.4	63.9	55.3	51.6	79.3	66.9	67.4	68.2	57.7	75.3	63.9	55.9	65.9	65.0	52.9	58.4	62.0	73.2	71.0	60.4	64.9	72.3	74.4	58.5	59.7	62.6	57.2	75.6	69.1	40.9	58.7	64.0	43.8	58.5	44.5	94.4	63.0
rel_em	60.8	57.4	61.8	53.3	65.1	49.4	48.4	86.7	61.3	58.0	73.1	66.4	53.1	56.1	58.1	49.9	58.2	52.9	59.8	51.4	47.7	67.2	62.6	64.6	61.3	37.7	72.5	60.8	41.1	35.5	61.6	45.2	50.4	56.1	70.0	69.5	56.4	61.5	69.7	72.9	55.2	55.3	58.3	51.3	73.7	40.1	36.7	53.8	62.1	36.4	57.7	42.7	94.0	57.1
gen_f1	58.9	58.3	63.0	55.6	65.2	68.5	52.5	85.1	61.8	57.3	72.7	65.3	53.7	57.1	60.4	54.7	57.9	59.5	62.1	52.2	50.3	78.5	64.1	63.0	65.4	60.8	72.0	63.5	50.5	68.2	61.4	53.4	57.0	59.9	69.5	67.9	56.2	62.6	68.0	71.3	56.5	56.7	61.1	54.6	72.9	68.8	37.1	56.7	61.4	41.7	56.2	41.5	78.9	60.6
gen_em	56.5	52.7	59.1	50.2	62.1	47.3	48.0	84.3	58.1	54.9	70.5	62.7	50.2	53.1	55.2	49.0	55.0	51.0	58.4	48.5	46.4	65.9	59.6	60.1	58.5	40.7	69.3	60.1	37.1	36.8	58.5	46.3	48.7	53.0	66.6	66.3	53.0	59.3	65.4	69.7	53.3	52.6	57.2	49.5	71.0	40.0	31.7	52.3	58.9	34.8	55.4	38.6	77.6	54.7
loc_f1	15.3	10.1	17.1	17.0	18.1	27.4	17.6	21.1	18.7	11.7	28.3	20.1	12.8	24.6	15.4	10.2	10.2	21.0	21.1	12.8	17.3	22.6	13.8	18.4	22.9	31.0	26.6	17.8	14.0	40.6	13.5	7.7	14.7	14.8	21.4	23.1	18.9	19.4	23.3	15.9	15.6	14.2	16.9	15.7	30.5	30.2	11.0	20.7	19.3	9.2	25.9	14.0	39.4	18.7
loc_em	13.4	6.3	14.1	14.4	14.7	11.8	15.1	20.1	17.6	10.6	26.7	18.6	7.6	21.9	12.6	8.4	8.9	17.7	18.1	11.4	14.7	7.7	11.2	16.3	21.6	9.1	24.0	15.8	10.2	6.9	11.5	6.3	11.5	11.2	19.3	21.5	16.4	17.0	21.3	13.9	13.8	12.4	15.0	13.1	28.3	7.4	7.6	19.2	17.1	6.8	24.4	11.0	38.7	14.5
port_f1	10.3	8.8	7.4	11.0	10.0	23.4	9.3	8.1	12.3	5.9	13.4	13.5	7.1	13.6	5.7	4.6	9.1	9.5	13.9	7.7	10.2	24.0	12.8	10.6	9.2	22.0	11.8	12.5	8.2	29.8	7.8	3.9	6.0	7.1	10.2	16.0	10.6	11.4	11.3	16.2	8.8	6.7	6.5	8.6	15.7	25.6	7.3	9.1	12.9	5.5	10.9	6.5	25.6	11.2
port_em	5.7	4.0	4.0	6.9	6.2	4.3	4.1	4.0	7.2	2.6	8.6	7.8	2.9	7.4	2.0	0.9	4.7	5.0	7.9	3.5	4.5	4.5	6.9	5.9	3.9	2.3	7.1	7.0	1.9	1.9	4.2	1.2	1.8	2.8	6.1	10.0	6.2	5.7	7.7	11.1	4.8	2.3	2.9	4.6	11.1	1.7	4.8	4.6	9.0	1.5	6.2	3.7	18.5	5.0

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	75.3	77.8	59.3	70.3	67.7	85.2	69.6	93.6	50.9	77.5	87.2	77.4	85.0	81.9	57.7	43.4	80.7	52.5	81.7	62.0	67.3	91.9	84.9	52.3	54.9	73.9	83.6	83.4	95.6	79.6	40.0	33.5	72.7	66.9	81.4	77.6	74.1	76.7	81.0	91.6	40.4	74.5	69.1	50.3	86.7	78.0	80.2	68.8	77.6	65.6	56.7	92.3	94.8	71.9
rel_em	73.2	71.0	48.7	55.2	56.0	63.7	66.9	93.2	46.4	75.5	85.1	75.8	80.8	79.3	39.6	18.1	77.8	31.7	79.3	58.0	63.4	81.8	80.3	48.5	31.5	35.4	81.3	81.5	94.5	39.3	39.5	26.3	66.2	60.9	78.4	76.3	71.0	73.6	78.8	89.7	34.8	71.5	65.7	43.8	85.3	45.8	75.4	50.6	71.1	55.4	56.4	90.0	94.8	63.8
gen_f1	73.3	77.4	56.8	71.1	70.6	85.0	67.5	93.4	50.0	77.1	85.1	76.3	83.8	79.2	58.8	44.6	79.7	50.9	78.7	61.7	64.4	90.7	85.0	52.8	52.6	75.1	81.6	81.5	95.2	79.6	38.1	36.1	70.4	65.8	80.1	77.2	72.9	75.9	78.9	90.9	43.3	72.3	68.7	51.4	85.1	77.3	80.6	71.7	78.3	65.1	56.0	91.5	93.5	71.3
gen_em	71.2	71.0	46.3	55.5	59.9	63.4	64.8	93.0	45.4	74.9	82.9	74.7	78.9	76.3	40.6	20.7	76.5	31.3	76.2	58.1	60.0	80.1	80.7	49.5	30.5	38.8	79.4	79.5	94.1	40.7	37.7	29.7	62.6	59.1	77.1	76.0	70.1	72.9	76.5	88.8	38.5	69.0	65.3	45.4	83.5	45.3	75.3	58.0	70.9	53.6	55.7	88.9	93.3	63.4
loc_f1	17.8	7.6	7.2	8.6	11.8	21.7	17.2	10.4	12.6	7.3	16.5	14.6	11.4	12.2	10.4	5.4	10.8	9.4	12.1	11.7	9.8	25.9	10.3	8.1	9.9	26.2	12.7	17.9	11.2	36.9	2.0	5.3	9.6	6.2	10.0	16.8	16.4	13.1	10.7	10.6	6.6	14.0	10.9	8.5	13.7	29.5	16.2	9.8	12.2	6.2	10.7	13.9	29.8	12.7
loc_em	17.0	6.2	5.7	5.6	8.4	7.3	16.4	10.1	11.3	7.0	15.7	13.8	9.9	11.5	8.0	4.0	9.7	6.4	10.9	10.8	8.9	14.6	8.1	7.0	7.8	2.8	11.5	17.4	9.4	3.3	1.8	4.0	6.6	3.9	9.4	16.2	15.4	11.6	10.1	8.7	5.2	13.1	10.4	6.8	13.0	5.7	13.5	8.7	9.7	3.6	10.5	10.0	29.4	9.3
port_f1	24.6	22.1	17.2	21.1	28.6	34.0	20.1	16.0	24.7	12.0	28.5	28.0	22.1	25.4	15.9	8.7	21.7	17.4	20.0	13.8	20.5	34.8	26.5	21.0	15.8	30.5	23.4	26.1	28.8	36.3	3.1	8.7	17.4	14.5	18.2	29.3	21.3	23.6	22.2	31.1	17.4	18.5	15.6	13.1	28.7	32.6	21.3	20.6	30.5	18.3	19.0	30.1	44.3	21.9
port_em	17.4	14.9	9.3	10.8	16.8	9.7	13.8	11.1	19.4	7.8	22.0	21.3	14.7	18.4	8.7	2.6	13.7	8.7	14.3	8.6	13.5	11.5	15.5	14.9	6.7	3.4	17.3	19.6	17.2	3.5	1.8	4.1	8.3	7.9	12.8	23.1	17.1	17.4	16.2	23.1	12.2	13.3	10.6	7.5	22.5	5.5	13.7	12.8	21.6	7.8	14.1	20.5	34.6	13.1

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	75.9	79.2	59.7	78.0	82.9	85.0	84.3	83.1	76.2	71.3	79.4	88.5	78.7	78.8	60.2	70.8	79.4	0.0	85.2	61.6	76.7	92.7	83.9	69.4	67.3	72.7	78.1	80.9	95.5	79.4	88.2	60.4	68.2	61.8	76.6	88.5	78.5	78.5	74.0	76.0	65.7	70.4	65.6	68.9	83.8	79.0	84.2	73.9	80.5	68.4	69.8	92.0	96.4	75.1
rel_em	72.6	74.0	52.5	69.8	76.1	64.7	81.5	81.6	72.3	67.8	75.9	87.7	74.0	74.3	46.1	63.0	76.6	0.0	82.9	55.8	72.1	85.3	79.8	65.2	54.3	42.7	74.2	76.6	94.6	41.9	85.2	52.7	59.6	52.1	73.0	87.8	73.8	75.0	70.2	69.8	60.6	66.1	58.5	64.2	80.1	47.7	79.8	65.7	74.0	58.8	68.2	89.6	96.0	68.2
gen_f1	75.8	78.9	58.4	77.2	80.8	84.3	83.7	81.7	74.9	71.0	78.0	87.7	78.6	78.0	59.8	70.4	78.3	0.0	83.9	60.5	75.8	92.5	82.7	68.6	67.0	72.5	77.7	81.0	95.0	78.9	87.6	59.0	66.1	61.5	77.2	88.7	77.8	77.5	73.9	74.8	63.9	70.5	64.8	66.5	83.2	77.9	83.7	74.4	78.6	68.3	69.7	91.3	94.9	74.4
gen_em	72.6	73.2	50.9	68.6	73.4	63.8	81.0	80.1	70.5	67.1	74.4	86.8	74.0	73.4	45.2	63.1	75.8	0.0	81.5	54.6	70.6	85.2	78.5	64.3	54.4	42.5	73.9	76.6	94.0	41.0	84.7	51.6	57.3	51.8	73.8	88.1	73.0	74.1	70.0	68.3	58.8	65.9	58.0	61.8	79.6	47.3	79.3	67.3	71.8	58.1	68.2	89.1	94.2	67.5
loc_f1	9.3	5.6	6.5	8.6	13.8	20.5	7.7	4.5	11.4	5.8	23.0	12.3	5.4	7.9	13.9	8.6	8.8	0.0	8.1	3.8	15.9	14.9	14.0	15.8	14.2	33.6	5.3	9.2	15.6	35.6	5.5	2.0	14.3	7.1	2.1	13.2	10.9	7.8	15.6	6.1	12.9	14.4	6.8	20.3	9.1	30.1	7.3	12.0	6.3	7.9	9.0	12.3	4.2	11.4
loc_em	8.3	3.7	4.7	6.2	10.5	5.8	6.4	4.0	10.1	5.3	21.1	11.0	4.0	6.8	10.9	6.6	7.8	0.0	7.2	2.6	14.5	3.9	11.6	13.6	12.3	10.6	3.8	8.2	14.2	2.3	3.9	1.7	9.9	4.2	1.2	12.7	9.1	6.0	13.5	3.6	11.0	12.6	5.5	18.1	7.1	5.9	5.2	11.1	4.2	5.7	7.2	8.9	3.4	7.8
port_f1	25.3	18.4	14.5	28.7	32.2	32.5	30.5	16.5	26.9	14.9	37.5	30.7	25.8	28.2	17.7	13.4	20.4	0.0	21.1	12.2	30.0	33.1	20.5	25.7	22.0	31.0	21.9	29.1	29.0	35.1	15.1	5.9	20.0	13.7	15.1	29.4	26.4	30.2	30.2	33.3	25.0	21.0	16.6	22.1	31.9	31.0	17.9	27.3	29.7	17.5	20.2	30.6	27.5	23.7
port_em	17.0	11.6	7.5	19.8	22.4	8.7	22.6	10.9	20.2	10.0	29.7	23.8	17.7	20.4	9.5	6.5	12.4	0.0	14.6	7.2	21.3	13.3	11.0	17.6	11.5	5.5	15.6	21.4	16.5	3.7	8.3	2.9	10.4	5.2	10.7	22.4	20.4	22.7	21.2	23.2	17.4	13.9	9.7	15.9	24.4	5.0	10.1	19.2	20.2	8.1	14.8	19.7	20.2	14.5

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	70.8	71.3	61.0	75.9	74.2	83.5	78.6	82.8	63.2	69.3	82.6	90.8	85.5	80.6	55.6	61.3	77.3	56.4	77.9	55.5	80.8	93.9	86.1	56.0	80.3	77.2	78.6	73.7	92.7	79.1	82.2	56.2	53.4	48.9	81.8	84.8	66.5	74.0	82.8	73.2	61.8	63.0	67.8	69.3	85.7	79.2	84.3	82.7	63.5	69.7	74.1	84.7	99.6	73.9
rel_em	67.1	64.5	51.5	65.0	66.3	53.3	74.5	81.8	59.5	65.2	79.2	88.5	81.4	76.4	34.8	45.2	74.3	37.3	72.8	49.9	78.4	85.2	81.9	49.5	75.2	50.3	74.3	68.3	91.8	43.2	80.5	49.9	33.1	30.5	78.6	81.5	60.8	70.2	78.4	66.2	58.0	57.7	60.8	64.1	82.0	45.8	79.7	77.5	53.6	59.9	71.0	82.6	99.5	65.5
gen_f1	70.9	71.1	60.3	75.7	74.4	83.0	78.1	81.8	60.9	67.9	81.8	90.1	84.4	79.2	55.9	61.7	77.5	54.9	74.8	54.5	79.4	93.3	85.3	55.3	80.3	76.5	77.9	72.4	91.9	78.5	81.3	54.8	51.9	47.4	81.0	84.0	66.3	72.2	81.6	75.3	62.7	63.9	67.1	66.4	83.9	78.9	84.0	82.7	64.1	69.7	73.0	84.7	94.0	73.2
gen_em	67.3	65.1	50.9	64.9	66.7	52.4	74.2	80.8	56.9	63.4	78.3	87.7	79.9	75.1	35.8	46.9	74.6	37.8	69.7	48.7	76.6	84.1	81.1	49.2	75.3	50.0	73.8	67.0	91.1	42.6	79.6	49.1	30.8	29.4	77.8	80.7	60.5	67.8	77.3	68.5	58.5	58.4	60.3	61.2	80.1	46.1	79.4	77.8	53.2	59.8	69.6	82.1	93.7	64.9
loc_f1	54.1	33.4	45.0	39.5	40.7	46.1	47.6	43.8	45.3	47.1	57.4	60.6	46.9	59.3	50.1	45.4	37.7	37.3	56.6	34.7	49.6	44.3	44.4	52.4	46.5	44.1	51.2	58.4	49.4	52.5	47.8	41.5	37.8	35.3	53.2	61.3	52.2	53.5	56.6	49.5	43.9	48.9	46.4	39.4	59.2	42.1	45.3	54.7	47.6	34.0	48.2	53.3	62.1	47.6
loc_em	51.7	30.2	40.5	34.0	34.9	30.0	45.8	41.5	43.0	45.0	54.8	58.7	45.2	57.0	47.5	40.7	36.4	27.2	54.4	30.8	45.8	30.2	41.2	48.1	38.9	20.8	46.6	56.0	47.3	18.2	45.8	38.5	27.4	24.3	50.5	59.9	47.7	48.7	52.8	47.1	39.6	45.6	42.4	36.2	55.4	18.6	42.0	51.9	43.0	29.9	46.2	50.5	60.4	42.0
port_f1	45.0	36.1	26.1	38.7	41.6	41.1	48.2	43.7	45.2	35.2	49.8	54.6	37.3	53.0	41.7	33.4	31.9	44.1	52.0	30.1	52.7	47.9	37.7	40.7	42.0	38.4	44.8	52.8	47.0	41.4	32.4	26.0	34.4	29.9	40.8	52.2	47.4	50.7	49.8	48.9	40.1	36.8	35.9	26.1	51.7	37.9	27.6	44.5	42.2	28.9	40.5	42.5	63.3	41.2
port_em	31.4	25.5	15.3	26.4	28.6	11.3	37.8	37.6	37.3	26.1	39.3	43.2	27.1	42.4	31.7	22.5	18.8	31.8	40.8	21.1	43.1	20.6	20.2	32.1	29.7	8.4	37.2	41.9	28.2	6.2	20.3	18.5	23.4	19.9	30.7	40.7	39.9	39.6	37.7	37.6	31.0	28.0	27.4	18.4	40.4	6.9	18.5	33.5	30.5	14.8	31.6	29.0	53.9	28.5

Table 10: Full Results on CounterFact.

WikiFactDiff

Llama3.2-3B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	34.5	27.0	83.6	38.7	43.5	45.5	65.8	84.7	79.2	62.8	49.4	54.3	35.5	51.9	80.0	78.4	26.3	83.3	49.0	71.8	41.9	41.4	47.5	77.7	86.3	37.6	69.2	77.5	25.6	33.0	33.2	70.7	81.3	73.7	79.7	68.0	77.8	56.6	81.9	52.6	76.0	61.1	66.5	68.9	75.7	41.1	53.6	82.0	44.2	30.8	74.7	33.6	100.0	59.0
rel_em	21.1	14.8	79.6	17.1	18.0	8.4	53.8	76.4	70.3	54.9	23.3	29.9	28.5	20.5	73.3	73.7	16.6	76.0	19.6	67.1	19.8	8.4	28.7	64.4	77.9	7.8	58.7	69.9	14.9	7.4	21.2	63.7	73.3	64.8	73.6	56.5	70.2	33.0	75.6	34.6	67.2	37.4	53.6	52.6	66.7	10.6	46.4	76.1	21.4	18.6	68.6	26.7	100.0	44.5
gen_f1	50.8	26.7	81.4	34.5	44.0	43.2	61.4	81.4	79.1	48.8	52.9	53.7	34.2	44.0	77.3	78.0	25.4	81.8	47.2	68.9	46.7	40.9	48.7	76.3	84.9	37.6	71.1	71.3	24.9	34.6	31.4	74.6	78.8	69.9	80.5	71.1	76.9	48.2	80.8	48.0	75.7	64.9	67.7	70.2	78.0	40.1	52.5	79.8	42.4	30.5	72.6	33.2	97.5	58.1
gen_em	32.4	14.8	77.0	13.7	20.9	7.4	51.0	72.2	70.4	40.3	33.6	35.3	27.9	22.8	70.7	72.8	15.9	74.4	29.0	63.9	31.4	8.5	31.3	64.8	77.1	7.7	62.0	63.0	14.2	7.1	21.0	67.7	70.7	60.8	75.0	62.4	69.1	35.5	75.3	29.1	67.9	46.0	57.4	58.0	70.4	10.1	45.8	73.7	20.8	18.8	68.6	26.3	97.1	45.1
loc_f1	3.9	4.5	6.7	5.0	3.8	19.5	6.4	7.1	6.8	4.6	3.9	6.4	3.5	5.5	5.1	5.4	3.3	5.4	4.7	4.7	5.0	16.7	6.9	4.9	6.4	22.7	6.9	6.0	5.0	23.2	4.2	4.6	7.4	6.3	7.3	4.2	6.5	4.8	6.6	4.9	5.5	4.9	4.5	4.6	5.6	21.9	4.5	6.3	5.4	4.9	6.7	4.3	16.3	6.8
loc_em	0.5	1.0	3.1	1.5	0.6	0.8	2.2	3.1	3.1	2.3	1.3	1.9	1.1	1.5	1.8	3.2	1.3	2.8	1.7	2.4	1.0	0.4	2.3	2.3	3.1	0.8	4.0	3.2	1.5	0.5	1.5	2.6	3.3	3.3	4.1	1.2	4.3	1.4	3.4	2.2	2.9	2.0	2.0	1.5	3.4	1.0	2.3	2.9	1.8	0.9	3.4	0.8	11.2	2.1
port_f1	1.0	0.9	2.2	2.1	0.9	17.3	0.7	1.2	1.4	0.8	0.7	1.0	2.1	1.5	1.4	1.5	1.6	1.9	1.2	1.0	0.9	15.3	2.4	1.3	1.3	19.8	1.7	1.4	1.0	25.8	1.0	1.3	1.6	1.2	2.5	1.3	1.4	0.9	2.0	1.4	1.1	1.0	1.2	1.1	1.1	22.5	0.8	2.0	2.3	1.7	3.0	2.4	5.1	3.2
port_em	0.5	0.0	1.0	0.6	0.0	0.1	0.1	0.3	0.6	0.3	0.1	0.1	0.8	0.1	0.3	0.4	0.1	0.5	0.3	0.1	0.1	0.0	0.3	0.4	0.1	0.3	0.6	0.8	0.0	0.0	0.0	0.1	0.3	0.1	0.5	0.5	0.5	0.3	0.3	0.3	0.3	0.1	0.4	0.1	0.1	0.0	0.0	0.5	0.9	0.3	1.8	0.1	2.9	0.3

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	79.6	10.4	86.9	44.1	61.4	53.7	81.9	91.6	81.7	82.2	85.9	89.1	23.5	85.9	86.8	80.7	12.2	83.2	88.2	81.9	83.1	33.8	35.0	81.9	70.1	10.3	87.4	84.2	6.6	40.0	35.7	81.7	87.1	79.9	84.2	89.5	83.0	85.9	85.5	60.2	81.0	79.3	82.0	67.4	90.0	23.6	55.2	85.8	56.7	38.8	81.7	35.2	99.9	66.8
rel_em	66.7	7.9	83.0	19.5	35.0	11.2	78.2	85.7	72.7	77.2	77.3	82.3	3.7	80.6	79.9	75.8	10.7	75.4	81.9	78.1	76.2	12.0	30.7	72.8	58.5	8.2	79.8	77.0	2.7	4.6	25.8	75.5	79.0	70.9	80.1	82.9	75.0	78.2	80.9	39.6	72.6	69.6	76.3	57.5	83.9	10.6	46.9	79.8	30.4	22.1	78.6	29.9	99.6	56.8
gen_f1	82.8	11.1	86.0	42.5	60.0	52.5	83.2	91.2	83.2	81.5	86.2	88.6	23.0	85.3	86.0	80.2	11.8	83.4	86.4	81.0	82.1	33.8	34.6	82.3	69.8	10.6	86.9	83.9	6.9	40.1	34.6	81.7	85.1	80.2	83.4	88.6	82.1	85.8	85.5	58.9	79.2	78.2	82.3	69.4	88.8	22.7	53.0	84.9	55.6	38.3	81.4	34.8	99.1	66.4
gen_em	74.0	8.3	82.1	18.8	34.4	11.5	79.7	85.3	75.0	76.0	78.1	81.6	2.6	80.4	78.8	75.4	10.5	76.0	79.6	77.3	75.3	12.8	30.7	73.9	59.4	8.3	79.2	76.3	3.1	4.5	25.3	75.5	77.1	71.2	79.8	82.3	74.2	78.6	80.9	38.3	71.6	69.0	77.0	59.6	82.7	10.6	45.4	78.9	32.1	21.1	78.4	29.2	98.9	56.7
loc_f1	5.6	1.0	6.6	5.3	6.1	22.6	6.7	7.4	5.9	5.6	6.0	7.8	2.5	6.3	5.6	4.5	0.8	5.0	7.0	5.6	6.4	12.0	3.1	6.0	4.7	2.9	6.7	6.3	1.2	25.5	3.9	5.0	7.0	5.6	5.6	7.6	5.6	7.9	6.9	6.3	5.1	5.5	5.3	5.0	6.2	9.6	4.5	6.3	6.1	7.6	5.2	4.4	11.6	6.4
loc_em	2.7	0.3	2.9	1.2	1.5	0.9	3.2	3.4	2.9	3.4	3.2	4.0	0.1	3.2	2.8	2.6	0.4	2.7	4.0	3.4	3.1	0.5	1.3	2.6	1.8	0.6	3.7	3.7	0.4	0.3	1.2	2.7	2.4	2.3	3.5	3.7	3.3	3.2	3.3	2.0	2.4	2.7	2.8	1.9	3.3	0.6	1.9	3.2	1.5	0.9	3.1	1.7	6.4	2.3
port_f1	2.0	0.1	2.3	2.4	2.2	18.1	2.4	1.9	1.7	1.5	1.8	3.3	2.3	2.8	1.9	2.0	0.0	1.8	4.9	1.4	1.7	10.5	0.7	2.4	1.4	2.0	2.3	2.9	0.0	26.8	1.0	1.3	2.3	1.6	1.8	3.9	2.2	2.5	5.0	2.7	1.7	2.0	1.4	3.8	2.0	6.3	1.3	2.2	3.6	2.3	1.9	1.5	8.5	3.1
port_em	0.4	0.0	1.5	0.1	0.6	0.3	0.3	0.5	0.8	0.4	0.8	1.4	0.4	0.6	0.9	1.2	0.0	0.5	1.3	0.3	0.5	0.3	0.4	0.9	0.3	0.0	1.5	1.3	0.0	0.4	0.1	0.1	0.9	0.5	0.6	1.5	0.9	0.6	0.9	1.2	0.5	0.9	0.6	1.0	0.5	0.0	0.1	0.9	1.4	0.0	0.9	0.3	4.5	0.6

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	90.1	17.3	86.5	38.4	32.0	57.3	81.8	93.7	81.2	81.3	85.2	82.7	40.5	79.4	85.6	83.8	11.5	80.8	82.0	77.8	77.5	26.8	49.4	82.7	89.0	17.5	84.4	78.5	14.5	39.9	16.1	80.3	85.3	81.6	85.2	86.4	80.3	77.2	81.9	49.4	82.8	80.6	78.1	42.7	85.6	34.1	47.9	86.6	38.9	37.3	79.5	32.3	99.9	65.0
rel_em	83.7	13.8	79.1	23.6	18.6	10.6	79.1	88.3	74.4	76.4	76.5	74.9	32.3	75.4	77.8	76.9	10.2	70.0	75.9	74.7	72.7	16.2	27.4	75.9	81.6	10.0	77.2	72.2	9.8	4.5	12.5	74.2	76.2	75.3	82.4	80.7	73.9	70.8	77.6	33.7	75.9	73.2	75.0	35.7	77.9	10.2	46.9	81.2	27.8	11.7	77.3	25.1	99.6	56.1
gen_f1	89.1	17.1	86.4	39.2	32.4	56.0	82.3	93.3	80.9	80.5	84.4	82.4	37.4	79.1	84.2	83.2	11.4	81.0	81.6	75.6	77.2	26.2	48.5	82.0	88.1	17.3	83.8	78.0	14.6	40.1	16.1	79.6	84.7	81.4	85.0	85.9	80.5	77.0	81.3	51.3	82.4	80.3	77.9	42.1	85.2	33.9	47.6	86.0	40.4	36.1	79.6	32.0	99.3	64.6
gen_em	82.0	13.8	79.9	23.6	19.4	10.6	79.3	87.9	74.1	75.6	75.5	74.6	26.0	75.1	76.2	75.9	10.1	70.0	75.5	72.7	72.5	15.7	27.2	75.1	80.8	10.1	76.7	71.7	9.6	3.8	12.5	73.5	75.8	75.0	82.3	80.2	73.9	70.7	76.8	34.6	75.5	72.8	74.6	35.0	77.6	11.0	46.7	80.6	28.3	11.5	77.3	25.1	99.1	55.6
loc_f1	8.2	1.9	6.0	5.3	3.9	22.4	6.4	7.3	6.0	5.9	5.7	6.4	2.6	5.3	6.5	5.3	0.8	5.3	6.0	4.9	5.0	7.7	7.3	5.3	6.6	6.2	6.0	5.8	2.4	24.9	1.5	4.8	6.9	6.0	5.7	6.9	5.4	5.0	5.9	4.6	5.6	5.3	4.8	4.3	5.5	17.9	2.4	5.6	5.5	8.1	5.1	2.8	10.8	6.2
loc_em	2.9	0.9	2.7	1.9	1.7	0.8	3.8	3.6	3.2	3.8	3.2	3.3	1.5	3.2	3.2	2.4	0.4	2.6	4.1	3.1	2.8	0.9	1.7	2.7	2.8	0.8	3.1	3.7	0.9	0.1	0.5	2.7	2.4	2.8	3.6	3.2	3.4	2.9	3.7	1.7	2.7	2.6	2.9	1.2	3.3	0.5	1.5	2.9	1.9	0.8	3.1	1.0	5.6	2.4
port_f1	4.6	0.4	2.8	2.3	1.8	19.7	4.4	1.8	4.1	1.8	2.4	4.7	1.6	5.7	2.3	2.4	0.2	5.0	5.8	1.2	2.5	6.7	2.9	4.5	2.4	6.8	2.7	6.6	0.2	27.0	0.3	1.3	2.0	1.1	2.5	6.0	3.5	6.5	4.5	3.8	2.6	2.9	1.6	1.5	3.9	13.0	0.5	2.9	3.7	2.9	2.7	0.9	10.5	4.0
port_em	1.7	0.1	1.7	0.4	0.6	0.5	2.6	0.5	1.7	0.8	1.2	2.7	0.5	2.7	0.9	1.7	0.1	2.0	2.8	0.3	1.2	0.7	0.8	2.3	0.6	0.0	1.3	3.4	0.0	0.4	0.1	0.3	0.9	0.1	1.3	3.1	1.2	2.7	0.6	1.3	1.4	1.7	0.8	0.4	2.7	0.0	0.0	1.3	1.3	0.0	1.5	0.1	6.1	1.1

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	90.9	21.2	87.8	43.0	27.5	57.3	89.2	93.2	86.1	81.1	90.6	90.5	33.4	85.8	84.2	79.1	27.1	82.0	85.0	81.5	80.4	50.4	52.2	81.4	86.4	28.3	87.1	87.6	28.9	40.9	27.4	78.9	83.0	72.1	90.3	92.1	82.6	67.1	84.5	50.8	84.2	82.0	78.9	14.0	90.2	38.4	55.6	93.5	51.2	45.1	83.7	38.3	99.9	67.8
rel_em	84.4	15.2	82.7	19.5	18.5	9.2	85.0	87.2	77.2	74.5	83.0	83.4	11.9	80.1	75.8	73.6	21.7	72.8	76.0	77.8	74.7	20.4	38.7	70.7	79.7	9.8	78.8	79.1	18.2	2.3	20.3	73.3	71.7	62.0	86.2	84.7	76.0	43.1	77.8	33.3	76.8	73.3	66.5	9.6	83.3	5.4	52.9	90.0	30.2	18.0	80.4	30.9	99.6	56.3
gen_f1	90.0	20.8	86.6	43.1	28.1	56.1	88.4	92.9	85.8	81.0	89.7	89.3	32.7	85.2	83.4	78.1	26.2	81.5	85.4	81.1	80.1	49.4	51.4	81.5	85.5	28.2	86.3	86.2	27.9	41.3	26.3	78.4	83.4	71.9	89.7	91.5	82.1	70.8	84.4	51.6	83.6	81.8	79.3	16.4	90.1	38.8	55.4	93.1	52.4	44.4	83.3	37.3	99.3	67.5
gen_em	82.5	14.9	81.4	20.4	18.6	8.9	84.2	87.0	77.0	74.7	82.0	81.8	11.1	79.3	74.9	71.9	21.1	72.5	76.4	77.2	74.5	19.9	38.0	71.8	79.1	10.2	78.0	77.6	17.6	2.3	19.4	72.6	73.2	62.1	86.1	83.9	75.5	50.1	78.3	34.6	76.2	73.2	68.0	12.1	83.3	5.7	53.1	89.7	33.0	17.7	79.9	30.5	99.1	56.2
loc_f1	8.7	3.7	8.6	5.6	6.5	22.6	11.7	12.8	7.6	6.1	8.3	9.2	3.6	11.5	7.3	6.5	2.5	7.1	10.3	7.1	9.8	17.9	7.9	7.6	8.6	19.9	11.6	10.4	6.0	21.4	3.4	8.7	10.3	9.2	8.0	10.5	6.6	9.6	10.2	7.0	8.0	7.7	8.7	2.1	8.4	23.7	5.2	9.7	6.5	7.8	7.5	7.8	15.9	9.1
loc_em	5.0	2.4	5.7	2.4	2.9	1.0	7.3	8.2	4.3	3.8	5.4	5.6	1.9	7.3	4.2	4.0	1.7	3.7	5.7	5.0	4.9	0.9	1.2	4.3	4.2	0.8	6.9	7.9	2.4	0.3	1.8	6.0	4.6	4.4	6.0	6.3	4.5	5.0	5.7	2.6	5.2	4.9	5.6	0.4	4.6	0.4	2.7	6.0	2.3	1.0	5.0	2.7	10.5	4.0
port_f1	11.8	2.4	8.4	6.4	6.5	25.2	12.8	5.2	12.0	8.5	11.1	9.8	4.3	13.6	8.8	8.5	7.1	12.0	13.6	5.0	14.3	19.3	6.2	13.0	12.9	25.4	11.5	12.2	5.3	30.0	2.4	3.3	5.1	3.7	12.2	12.5	13.2	13.4	13.6	6.9	11.0	12.1	11.6	1.8	12.1	25.8	5.9	13.1	8.7	5.2	9.4	11.9	15.6	10.7
port_em	8.4	0.9	5.7	1.0	2.6	0.9	9.2	3.2	8.2	5.6	8.3	6.6	2.2	10.2	6.4	6.4	3.3	8.6	10.3	2.7	10.5	2.7	1.3	10.1	9.5	0.6	7.8	9.1	1.7	0.0	1.3	1.0	1.7	1.9	9.1	8.9	9.7	10.6	9.4	3.6	7.9	9.7	8.8	0.5	8.9	0.0	4.2	9.8	3.8	1.8	8.2	2.8	11.0	5.5

Llama3.1-8B

0-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	87.3	31.2	85.2	31.8	38.6	30.3	82.3	93.1	85.1	77.0	85.7	86.6	44.4	80.5	86.3	81.3	32.6	86.5	82.0	83.0	71.1	44.7	57.6	84.7	85.5	20.6	85.9	86.3	30.3	27.6	47.5	82.9	84.4	82.1	85.1	90.7	85.0	85.4	85.4	35.4	83.2	81.9	85.6	75.7	90.9	29.6	59.0	83.4	33.7	34.5	82.1	39.2	99.6	67.8
rel_em	79.0	20.7	82.1	21.7	23.9	9.2	78.6	87.5	77.9	70.8	76.0	79.0	35.6	70.4	80.4	76.2	24.0	80.5	71.4	79.3	53.4	18.8	42.2	77.6	79.1	10.1	78.4	78.3	15.4	10.6	33.4	77.7	78.4	74.1	81.4	85.2	79.2	78.6	81.5	24.4	75.8	72.8	80.1	67.7	85.6	13.8	53.7	78.7	24.7	23.9	80.1	32.7	99.5	59.1
gen_f1	85.7	30.1	84.0	29.9	34.7	29.8	81.2	92.2	84.0	71.1	84.3	85.6	42.8	77.3	84.9	81.0	31.0	85.3	78.4	80.6	74.8	45.0	55.6	83.7	83.4	22.0	84.2	83.4	26.3	28.2	44.1	82.7	85.0	81.8	82.9	88.9	83.1	85.8	83.9	33.5	83.9	80.5	84.4	75.6	88.7	30.5	57.4	80.8	31.3	32.9	79.5	35.3	93.7	66.4
gen_em	78.2	19.9	81.3	20.4	21.9	9.1	77.2	86.9	76.8	63.8	75.6	77.7	34.8	66.3	78.7	75.6	23.6	79.2	67.9	76.9	62.8	18.1	41.1	75.6	75.7	10.2	77.2	75.0	13.5	10.2	30.3	77.8	78.4	74.3	79.6	83.0	77.4	79.0	80.1	22.5	76.5	71.4	78.6	66.3	83.4	12.9	52.3	75.9	23.4	22.3	77.2	30.6	92.7	57.8
loc_f1	10.5	6.9	10.0	4.6	6.1	16.0	11.4	13.2	11.5	6.6	9.0	12.9	4.4	9.8	10.3	7.9	4.5	10.7	11.6	9.0	10.3	17.5	7.8	10.6	11.7	19.3	11.6	11.4	5.0	23.9	7.6	8.7	9.7	9.2	12.1	11.2	10.1	11.7	12.1	5.4	8.8	9.4	10.6	9.0	10.8	17.1	6.1	9.6	5.0	5.1	10.5	6.2	17.8	10.0
loc_em	5.4	3.2	5.4	1.8	2.6	1.3	5.7	8.4	7.5	4.6	4.9	8.0	1.4	5.4	6.1	4.7	2.7	7.1	7.3	6.0	5.0	0.8	2.8	6.5	7.4	2.2	7.0	7.3	2.2	1.7	3.5	5.2	5.0	5.4	7.3	7.8	6.3	6.4	7.8	2.6	4.9	5.6	5.0	5.1	7.0	0.9	2.4	4.7	2.4	1.7	6.8	2.3	12.5	4.8
port_f1	2.3	2.4	2.9	3.4	3.2	14.9	2.8	2.9	2.3	1.4	2.9	4.4	2.6	4.4	2.1	2.4	2.4	2.9	5.1	1.9	2.9	16.8	6.2	3.1	2.3	11.8	2.1	3.7	2.5	20.0	1.7	1.3	1.9	1.8	1.7	5.2	3.2	4.1	4.9	2.9	1.8	3.0	1.9	2.7	3.0	18.7	1.3	3.4	3.1	3.3	2.2	1.9	8.5	4.2
port_em	0.6	0.1	1.4	1.3	1.5	0.9	0.5	0.9	0.9	0.5	0.9	2.6	1.0	0.9	1.2	1.3	0.4	1.3	1.5	0.4	1.2	0.5	2.6	1.2	0.6	0.0	1.2	1.9	0.3	0.4	0.4	0.1	0.6	0.9	0.4	1.7	1.0	1.7	1.8	1.3	0.6	0.8	1.0	1.3	1.5	0.0	0.4	1.7	1.7	2.2	1.3	0.0	5.4	1.0

1-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	91.8	15.6	86.7	41.1	70.7	60.9	86.1	94.1	86.8	82.0	89.8	87.9	29.4	88.6	87.8	84.4	16.0	90.2	88.6	83.6	83.7	39.3	40.7	87.7	85.3	8.8	87.8	91.3	11.4	43.2	47.9	83.4	82.7	86.6	84.6	92.6	81.6	85.9	85.7	58.6	84.9	86.0	84.9	81.1	91.0	31.6	60.5	90.6	59.3	44.8	81.2	41.0	100.0	70.5
rel_em	85.6	12.6	83.6	25.5	46.9	14.8	82.7	89.0	79.3	77.2	82.8	80.0	4.6	83.6	82.4	79.3	14.5	84.3	82.7	80.4	76.7	19.9	34.7	80.1	79.6	8.2	81.1	84.6	7.4	8.3	35.2	77.8	71.5	80.0	82.1	87.1	75.4	78.2	81.5	41.9	77.3	78.3	79.5	70.9	85.1	13.4	55.0	85.8	34.4	23.2	78.6	33.9	100.0	61.1
gen_f1	90.2	15.3	86.5	40.8	69.3	59.1	85.6	93.0	88.4	82.2	88.2	87.1	28.8	88.0	87.7	83.8	15.2	89.5	86.8	82.5	82.8	38.5	40.3	86.3	83.8	9.1	87.7	88.2	12.0	43.6	47.0	82.9	82.8	84.8	83.3	91.3	82.2	85.8	84.0	57.5	85.2	85.4	84.9	81.1	90.4	31.2	59.7	89.7	58.6	43.5	80.3	39.9	99.5	69.8
gen_em	82.7	12.4	83.3	25.3	45.4	15.1	82.3	87.9	80.7	76.9	80.4	79.3	3.2	83.2	82.0	78.4	13.7	83.7	80.5	79.1	76.3	20.4	34.4	79.3	77.7	8.0	81.2	81.4	7.8	7.7	34.0	77.6	71.9	77.0	80.8	85.6	76.3	78.2	80.0	40.7	77.7	77.7	79.3	71.7	84.3	13.5	55.1	84.9	34.4	23.5	77.9	33.3	99.5	60.5
loc_f1	10.1	1.1	6.8	5.0	7.5	24.2	9.7	8.1	8.5	6.1	7.2	8.3	3.1	8.0	7.8	5.7	1.2	7.9	10.3	5.8	7.9	14.8	4.5	8.3	6.4	4.6	7.1	8.8	1.6	26.2	6.4	6.2	6.9	8.2	5.7	8.6	7.0	8.6	7.6	6.4	8.2	7.6	7.5	5.9	7.1	15.0	4.3	8.8	6.5	9.3	6.9	4.5	13.5	7.8
loc_em	5.1	0.4	3.1	1.7	2.6	0.9	5.9	4.0	4.9	3.8	3.8	4.7	0.1	4.5	4.0	2.9	0.8	4.5	6.0	3.6	4.1	1.3	1.7	4.7	3.5	1.0	4.1	4.7	0.8	0.3	2.2	3.6	2.6	3.5	3.8	4.9	4.5	4.5	4.7	2.0	4.3	3.6	3.8	2.7	4.5	1.2	2.2	4.9	2.1	2.3	4.3	1.5	8.0	3.2
port_f1	4.0	0.1	3.0	3.3	4.2	19.5	3.2	2.3	3.2	2.2	2.8	5.0	2.8	4.8	2.9	2.2	0.4	2.8	6.8	2.2	2.4	13.4	3.2	4.6	1.7	2.6	1.9	4.8	0.1	27.8	1.5	1.8	2.2	2.0	1.4	6.3	2.7	5.2	6.3	3.3	1.9	3.2	1.7	3.8	3.5	10.4	1.3	3.0	4.7	5.0	2.3	2.3	13.4	4.2
port_em	1.8	0.0	1.8	0.9	1.7	0.5	0.6	0.9	1.8	1.2	0.9	2.0	0.5	2.2	1.9	1.0	0.3	1.2	2.8	0.8	0.9	0.4	2.0	1.9	0.4	0.0	0.5	2.6	0.0	0.5	0.5	0.6	0.8	0.9	0.3	3.1	1.3	2.4	3.3	1.7	0.8	1.7	1.0	1.5	2.3	0.0	0.3	1.5	2.1	2.2	1.3	0.3	7.7	1.2

8-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	91.9	25.0	0.0	39.4	45.6	60.8	83.3	93.5	85.0	85.6	88.8	85.2	47.0	85.2	87.7	85.7	22.3	85.1	84.9	82.0	84.3	52.8	51.7	86.2	89.6	23.1	87.9	82.3	26.4	43.7	28.4	81.7	87.0	87.7	92.7	91.0	82.1	84.0	87.0	37.3	84.4	84.6	85.1	71.4	88.1	42.9	56.3	88.4	42.2	46.6	83.4	41.6	100.0	68.6
rel_em	84.4	19.9	0.0	23.5	29.2	14.2	80.4	88.1	78.4	79.6	81.1	77.6	37.8	81.5	80.6	76.9	20.0	68.9	78.6	78.7	79.6	24.1	29.5	78.7	82.9	11.1	81.1	76.0	18.1	4.7	21.9	75.6	78.3	81.4	90.9	85.2	75.4	77.3	81.8	26.4	77.8	77.0	79.0	61.5	81.1	10.3	53.3	83.4	27.8	15.9	80.2	32.4	100.0	58.4
gen_f1	91.5	24.0	0.0	39.2	45.0	58.7	83.4	92.8	85.5	85.2	86.9	84.9	45.8	84.2	87.1	85.5	22.0	85.4	84.3	81.2	83.5	53.9	50.7	85.5	89.3	23.8	87.3	81.6	25.7	43.0	27.8	81.3	86.4	87.6	90.8	90.5	81.6	82.2	85.1	36.8	83.9	84.0	84.5	65.6	87.1	42.2	55.0	88.1	42.3	45.6	82.9	40.7	98.9	67.9
gen_em	84.2	19.4	0.0	23.2	28.8	14.0	80.4	87.4	79.0	79.3	78.4	77.4	36.7	80.0	80.0	76.4	19.8	74.0	78.3	77.9	78.6	24.1	29.0	77.8	82.5	11.1	80.6	75.1	17.0	4.2	21.4	75.4	77.8	82.1	88.9	84.7	75.0	75.5	80.1	25.5	77.3	76.3	78.8	55.9	79.9	10.2	52.6	83.0	27.9	15.4	80.2	31.6	98.9	57.9
loc_f1	11.5	4.9	0.0	6.4	6.4	24.4	7.3	9.9	8.6	6.7	7.3	8.0	4.8	7.8	8.3	7.8	2.4	10.3	9.5	5.9	9.1	18.4	7.3	7.2	7.8	11.5	7.5	8.6	3.6	26.4	3.1	6.0	7.7	8.2	10.1	8.7	5.9	8.2	8.0	4.0	6.5	6.0	7.1	4.9	7.3	21.7	4.0	8.4	5.2	10.3	7.7	6.1	16.8	8.3
loc_em	5.6	2.3	0.0	2.0	2.9	1.2	4.7	5.4	5.4	4.5	4.6	4.5	2.8	5.2	4.6	4.0	1.8	5.9	6.6	3.7	5.6	1.2	2.4	3.7	4.3	0.9	4.5	5.5	1.2	0.6	0.9	3.8	2.9	3.5	6.3	4.7	4.0	5.0	4.7	2.2	3.2	2.7	3.8	1.4	4.7	0.5	2.3	5.1	1.8	1.7	5.0	2.6	12.0	3.5
port_f1	11.5	3.8	0.0	6.1	5.7	24.6	11.7	4.9	10.4	5.0	8.3	10.7	3.2	11.0	3.8	6.7	1.0	9.9	10.4	4.3	10.9	21.7	6.8	8.8	9.0	14.6	11.0	13.2	1.5	29.0	2.3	2.5	3.4	2.5	13.1	12.9	10.4	12.6	12.6	9.2	9.5	9.0	3.8	2.0	11.6	18.5	2.5	12.3	7.2	8.1	9.9	6.3	18.2	8.9
port_em	8.4	0.8	0.0	1.4	1.7	1.2	8.3	2.9	6.1	4.0	6.0	7.3	1.2	7.7	2.3	4.7	0.3	5.7	6.4	2.4	7.0	3.1	2.3	6.3	5.2	0.0	6.4	9.2	0.4	0.3	1.3	1.4	1.7	1.0	8.7	9.1	6.8	8.4	8.4	4.0	6.1	6.1	2.9	0.4	8.6	0.3	1.2	9.1	4.0	1.9	8.2	1.3	12.0	4.2

8a-shot

	af	ar	az	be	bg	bn	ca	ce	cs	cy	da	de	el	es	et	eu	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ka	ko	la	lt	lv	ms	nl	pl	pt	ro	ru	sk	sl	sq	sr	sv	ta	th	tr	uk	ur	vi	zh	en	Avg.
rel_f1	90.5	28.6	84.9	48.1	29.3	60.3	91.8	88.6	87.9	85.4	91.4	89.8	46.5	91.5	88.0	83.5	30.6	86.2	90.8	86.1	83.6	50.6	50.9	87.9	89.3	34.5	89.5	92.2	36.2	43.6	29.3	83.1	80.3	76.5	93.8	93.4	86.9	87.6	92.2	43.1	86.2	84.6	85.5	69.3	90.6	43.3	59.4	95.2	39.5	47.6	91.0	46.8	100.0	72.0
rel_em	83.2	21.8	76.4	26.5	20.5	13.5	85.1	80.2	79.1	76.7	84.7	81.4	36.4	85.3	80.5	78.4	24.2	76.0	82.7	81.5	76.8	20.2	31.1	78.7	81.0	10.1	82.0	85.0	23.3	4.2	22.8	77.4	66.6	62.1	90.7	87.0	79.3	77.3	87.4	30.0	78.2	75.5	75.3	59.2	84.1	4.9	56.0	92.5	21.5	15.7	86.6	34.7	100.0	60.2
gen_f1	89.9	28.9	84.4	47.3	28.8	58.4	90.5	89.0	88.0	84.8	90.7	89.8	45.3	90.1	87.5	82.5	30.0	86.5	90.0	84.9	84.0	50.6	50.2	87.5	89.1	34.7	88.8	91.8	35.5	44.0	28.7	82.5	78.7	76.9	93.6	92.7	86.1	84.1	91.5	44.3	85.9	83.9	85.5	62.2	89.2	42.9	58.8	94.3	41.4	47.5	89.7	45.8	99.6	71.3
gen_em	82.5	22.2	76.4	26.2	19.4	13.4	82.9	81.0	79.2	76.3	83.9	81.6	35.0	83.9	80.0	76.5	24.2	76.4	81.9	80.4	77.3	19.9	31.1	78.6	80.8	10.2	81.2	84.7	23.1	4.6	22.4	76.7	65.2	63.1	90.4	86.1	78.7	71.4	86.4	31.4	77.7	74.5	76.3	52.7	82.3	5.0	56.0	91.7	24.7	16.2	85.3	34.2	99.6	59.7
loc_f1	14.5	9.6	12.4	8.4	9.7	26.2	18.4	17.4	9.7	13.9	11.6	14.5	7.3	18.1	13.8	10.8	4.8	14.0	18.2	10.4	17.9	22.6	12.6	11.9	14.3	24.7	14.4	17.1	11.5	27.6	4.6	13.0	14.3	14.9	13.8	15.5	9.6	16.7	16.9	7.5	10.8	13.4	13.5	9.7	12.4	24.9	8.8	12.9	9.7	13.3	12.5	13.2	19.5	13.8
loc_em	9.8	5.5	7.5	3.7	6.3	1.5	12.5	12.5	6.3	8.9	7.7	9.7	4.5	12.6	8.8	6.8	3.2	8.3	11.9	7.3	11.7	2.2	5.2	7.8	9.1	1.7	10.0	12.0	5.7	1.9	1.8	8.8	8.3	10.2	9.5	11.2	6.1	10.7	11.7	5.0	7.1	8.9	8.3	5.6	6.0	0.8	5.4	8.3	5.2	6.3	8.2	7.1	13.7	7.4
port_f1	17.1	9.0	15.3	8.9	11.5	25.9	16.1	8.8	15.8	11.1	15.1	13.1	6.2	17.6	13.5	12.1	10.8	14.0	18.4	13.3	16.5	24.8	10.0	14.6	16.7	29.1	15.6	16.9	10.1	30.4	4.1	7.6	10.5	8.7	16.4	18.0	15.2	19.5	18.0	13.6	15.5	16.8	15.5	4.8	16.4	26.2	9.8	16.2	11.4	7.9	14.4	13.4	21.8	14.6
port_em	12.8	5.1	11.2	1.9	5.1	1.8	12.0	5.5	11.0	9.2	12.2	10.7	3.3	13.4	11.1	9.6	4.3	9.7	13.9	9.6	12.4	5.6	3.6	11.2	12.1	1.4	11.1	12.6	2.3	0.5	2.6	5.4	5.9	5.1	12.4	13.7	11.4	14.5	13.5	7.4	11.2	12.5	11.7	2.4	13.0	0.4	6.6	12.5	6.3	2.6	12.2	4.9	15.9	8.4

Table 11: Full Results on WFD.