FSTs vs ICL: Generalisation in LLMs for an under-resourced language

Anonymous ACL submission

Abstract

The effectiveness of LLMs remains uncertain in scenarios where pre-trained models have limited prior knowledge of a language. In this work, we examine LLMs' generalization in under-resourced settings through the task of orthographic normalization across Otomi language variants. We develop two approaches: a rule-based method using a finite-state transducer (FST) and an in-context learning (ICL) method that provides the model with string transduction examples. We compare the performance of FSTs and the neural approach in low-resource scenarios, providing insights into their potential and limitations. Our results show that while FSTs outperform LLMs in zero-shot settings, ICL enables LLMs to surpass FSTs, stressing the importance of combining linguistic expertise with machine learning in current approaches for low-resource scenarios.[1]

1 Introduction

Large Language Models (LLMs) have been widely adopted to tackle many traditional NLP tasks. Part of their success is attributed to their extensive pretraining, which enables them to generalize well across different domains and diverse linguistic structures. The rapid advancement of the field has led to techniques like in-context learning (ICL), which have further showcased the impressive generalization capabilities of LLMs, allowing them to adapt to new tasks and domains with minimal training data, often requiring only a few examples.

However, the effectiveness of these approaches remains uncertain in scenarios where pretrained models have limited prior knowledge of a language. This is particularly relevant for under-resourced languages, where training data is scarce or highly non-homogeneous. The performance of LLMs in these situations is not yet well understood, and it is unclear whether their impressive generalization capabilities can be replicated in all type of scenarios. This raises critical questions about the extent to which in-context learning and other recent innovations can bridge the gap in multilingual coverage.

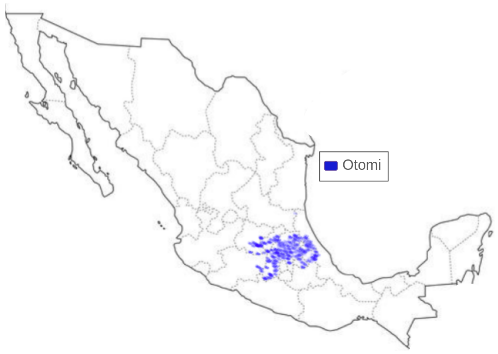


Figure 1: Geographical distribution of Otomi

In this work, we examine LLMs' generalization in under-resourced settings through the task of orthographic normalization across Otomi language variants. We develop two approaches: a rule-based method using a finite-state transducer (FST) and an in-context learning (ICL) approach that provides the model with string transduction examples. We focus on the scenarios that are particularly difficult for a transducer-based approach.

2 Background

In-context learning. A paradigm where language models learn to perform tasks by recognizing patterns from a few provided examples (demonstrations). Unlike supervised learning, which requires explicit parameter updates or a separate training stage. ICL enables models to make predictions by drawing analogies from the given context (Brown et al., 2020).

Although the mechanisms that underlie and influence this type of learning are not fully understood (Dong et al., 2024), it has proved successful in several domains. ICL techniques can be exploited

[1] The code and data will be available at

in domains with minimal training data, requiring only a few examples. For example, multilingual tasks involving under-resourced languages sucha as machine translation and interlinear glossing (Coleman et al., 2024; Clarke et al., 2024; Aycock et al., 2025).

Therefore, ICL can be considered a form of generalization, but it has nuances compared to traditional ML.

Orthographic normalization of Otomi. Orthographic normalization is the process of converting written text into a standardized form within a language. For many dominant languages with long-established writing conventions, this task is either well-solved or of limited concern. However, for numerous other languages, particularly those lacking a long tradition of standardization due to sociopolitical factors, normalization remains a significant challenge. These languages often exhibit high internal diversity, making the task even more complex.

Otomi is a group of languages spoken in central Mexico (Fig. 1) that are part of the Oto-Manguean language family. They exhibit around nine dialectal variants (Lewis, 2009; INALI, 2008). This is an endangered language (round 300,000 speakers) that faces a scarcity of NLP tools and digital resources, plus there are several orthographic standards that speakers can use when writing their language.

Automatically converting text across the different standards is a crucial upstream task for developing more advanced language technologies. To our knowledge, no normalizers exist for Otomi. However, in low-resource settings, common approaches include building FSTs based on linguistic expertise (Johny et al., 2021; O'neil et al., 2023) and applying neural models like seq2seq (Lusetti et al., 2018; Lutgen et al., 2025).

3 Data and Methods

3.1 Orthographic norms

Documents written in Otomi exhibit variability, with multiple orthographic standards in use. This study focuses on the most common norms (Table 1).

Rule-based normalizer. We developed finite-state transducers (FST)[2] to convert text between different Otomi orthographic standards (norms), using a two-step process: first, mapping source text to a phonetic alphabet (IPA), and then generating the target orthography. The transduction rules were informed by a linguist's expertise and existing documentation (Hernández-Green, 2016). The system was implemented using the HFST toolkit[3]. The FST converts text across standards without requiring a specified source norm.

Norm	Description	Ref
INALI	Norm designed by the National Institute of Indigenous Laguages of Mexico	(Inali, 2014)
OTS	Standard used in some texts from variants in the State of Mexico	(De la Vega, 2017)
OTQ	Standard proposed mainly for Querétaro variants.	(Hekking and de Jesús, 1989)
Norm	**Example sentence**	
INALI	[...]bijúgígó escuela pero ndichichithóhó	
OTQ	[...]bijúgígó escuela pero nditxitxithóhó	
OTS	[...]bikjúgígó escuela pero ndichichitjójó	

Table 1: Otomi orthographic standards

Neural approaches If we already have linguistic rules for converting text across orthographic norms, costly neural network-based methods may be unnecessary. However, transducers have limitations, as they lack flexibility and struggle to adapt to speakers' linguistic realities. We identify two key challenges where this approach falls short:

- Code-switching: Texts often include words from other languages, mainly Spanish and Nahuatl (e.g., *escuela* and *pero* in Table 1). Since FSTs apply rules that were specific for otomi, handling these cases is challenging. Additionally, language identification tools for many indigenous languages are limited.

- Ambiguity: Instances where the same input can be mapped to multiple outputs. A transduction rule may be favored over others, sometimes leading to incorrect mappings. In Otomi, we observed that the same grapheme can be transcribed to different phonemes based on the orthographic norm of the source text. This phonological ambiguity can lead to errors, which can propagate to the target norm realization.

With these challenges in mind, we tried to leverage the generalization and adaptability of neural approaches to tackle these complex cases. Specifically, we focus on using LLMs with ICL.

3.2 Few shot examples and test set

To compare FSTs and neural approaches, we built two test datasets: (a) OTS→INALI : OTS as

[2] The tool will be available as an open source normalizer

[3] https://github.com/hfst/hfst/wiki

	OTS→INALI	OTS+OTQ→INALI
# sentences (test)	191	191
# sentences (few-shot)	10	10

Table 2: Datasets size for each normalization setting

the source norm and INALI as the target, and (b) OTS+OTQ→INALI: sentences in OTS or OTQ with INALI as the target[4].

These sentences[5] were selected in all cases to ensure coverage of code-switching, ambiguity phenomena, and the typical transformations across the written standards. The initial transduction for obtaining the different normalizations of the test sentences was performed using FSTs, followed by manual correction of errors. The target norm is always INALI . We made this decision as it represents an institutional effort to unify standards; converting text to this standard is a typical case of use for speakers of this language.

Additionally, for each case, we manually selected 10 representative sentences to use as demonstrations for an ICL approach (Table 2).

3.3 LLMs

For the ICL neural approaches, we used the following base models: *GPT-3.5 Turbo, GPT-4o, LLaMA 3.1, and LLaMA 3.3.*

Zero-shot setting: We prompt the system to generate normalized test sentences without prior examples. For OTS→INALI , we explicitly specify the source and target norms. For OTS+OTQ→INALI , we do not specify the source norm, only that input sentences may come from either norm and should be transduced into INALI .

Few-shot setting: We provide models with 10-shot examples of orthographic transductions. For OTS→INALI we show 10 examples and specify that test sentences are in OTS , requiring INALI orthographic normalization. For OTS+OTQ→INALI , examples include transductions across INALI , OTS , and OTQ . During testing, the source norm is unspecified (either OTS or OTQ), and the model must generate the INALI normalization.

In all settings, we state the language and alert the system that there might be cases of code-switching and challenging transductions. See Appendix A for example prompts and details about the models.

4 Findings and Interpretation

To compare model performance, we measure the error rate between predicted orthographic normalizations and gold standards using Word Error Rate (WER) and Character Error Rate (CER), the latter being particularly useful for morphologically rich languages (James et al., 2024). Figure 2 shows the overall results.

Both settings, OTS→INALI and OTS+OTQ→INALI, exhibit a similar trend: the rule-based FST approach outperforms state-of-the-art LLMs in orthographic normalization of Otomi text when they are prompted in a zero-shot setting. This is notable since FSTs are computationally lightweight compared to the extensive resources (data and infrastructure) required for training large neural models. Despite their robustness across many tasks, these LLMs struggle to generalize to orthographic variations in an under-resourced language like Otomi

Surprisingly, providing neural models with few-shot examples drastically improves their performance. Models like GPT-4o show some of the worst performances in a zero-shot setting. Still, with just 10 examples provided, the error rate decreases, outperforming FSTs and becoming the best model for orthographic normalization. See for example, GPT-4o_zero (*WER: 31.5% CER: 10.1%*) vs. GPT-4o_few (*WER: 11.2% CER: 2.6%*) in OTS→INALI .

The plots show a clear trend: LLMs make more errors than FSTs but surpass them in a few-shot setting, highlighting the effectiveness of ICL. However, further exploration is needed, as some studies suggest LLMs' sensitivity to input and target probabilities may affect their emergent capabilities, especially in rare languages and text sequences (McCoy et al., 2024).

Interestingly, the most recent models perform worst in zero-shot, contrary to expectations given their sophistication. We conjecture that they may prioritize specialized reasoning capabilities over multilingual flexibility. However, their ICL capability remains remarkable.

Despite expectations, the OTS+OTQ→INALI setting yielded lower error rates, suggesting it was easier. We anticipated greater difficulty since neural models lacked source orthography information at test time. A possible reason is that OTQ and INALI are similar, with few character transformations, making sentences in this norm relatively easy to normalize, even without source orthography information at test time.

[4] We omitted the OTQ→INALI case since both standards are similar. Instead, we created a more challenging setup

[5] Our dataset is built from a small online Otomi corpus that gathers different sources and varieties https://tsunkua.elotl.mx/

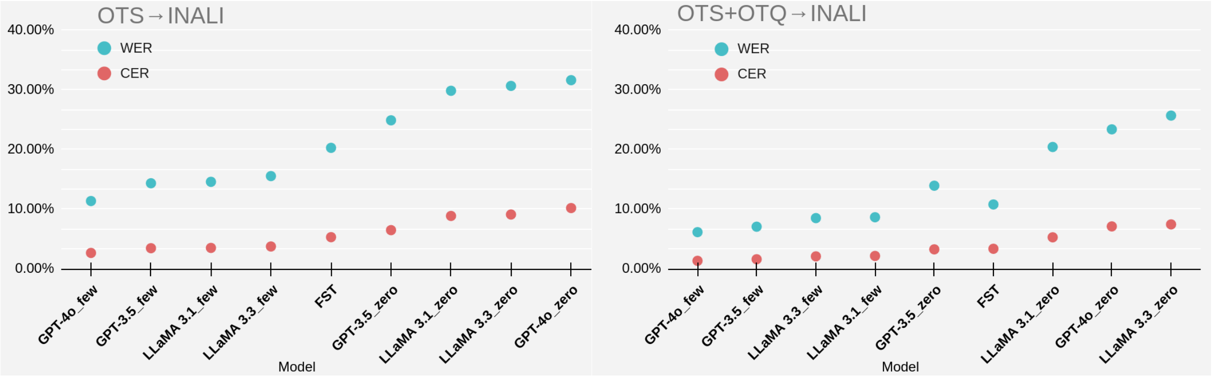


Figure 2: Performance of different models for Otomi orthographic normalization

4.1 Error analysis

We know that FSTs make mistakes when trying to normalize cases of code-switching and rule ambiguities since our approach has limited strategies to deal with that. But what errors are neural models most prone to? We analyzed test sentences with the highest CER and WER to answer this.

A key finding is that both zero-shot and few-shot neural models handle code-switching well, easily recognizing non-Otomi words and avoiding unnecessary transformations. This good handling performed by LLMs is expected for dominant languages like Spanish and English but also extends to proper names, place names, and loanwords of Nahuatl, an under-resourced language.

Errors in the few-shot setting mainly stemmed from failing to infer transduction rules (especially the ambiguous ones) and mixing up norms when trying to generate the target norm. In zero-shot, there were additional errors that introduced noise,e.g., difficulty in handling the graphemes that correspond to vowels and tones of Otomi, the systems often modified them or added accent marks even though this was not required for the normalization (e.g., ä→aa, o̲→o, umbabihe →ūmbábihé), and hypercorrections that treated Otomi words as Spanish.

Original sentences vs gold standard: As a sanity check, we calculated error rates (WER, CER) between the source text and its INALI-standardized form (gold standard). These rates are typically high due to differences in written forms and should decrease when a normalization tool is applied. However, our case is unique since many test sentences contain code-switching, where several words should remain unchanged. This results in lower-than-expected error rates, making it unsuitable as a baseline. Still, few-shot models outperformed this measure. See Appendix B for details.

5 Conclusions

Our work investigates the generalization limits of LLMs in an orthographic normalization task where the models likely have little prior knowledge of the language. To do this, we developed the first rule-based system for converting Otomi text across different norms and compared its performance against neural approaches, particularly in zero-shot and few-shot settings.

The test set was designed to assess the model's capacity to normalize cases that are difficult for a rule-based approach, i.e., code-switching and ambiguous orthographic rules.

One of the main takeaways is that when working in a limited resource scenario, one can leverage knowledge of the language to build an FST, and this can be more effective than simply doing zero-shot prompting with sophisticated LLMs. However, once you have a FST you can use it to generate demonstrations of orthographic transductions across orthographic norms and use them to improve a neural model. Our results showed that LLMs surpass FSTs with just 10 examples in few-shot settings, they were particularly good in code-switching cases.

This highlights the potential of ICL to generalize from limited data, reducing the reliance on extensive labeled datasets while reaffirming the value of linguistically informed data. This could be promising for many practical applications, including developing technologies for under-resourced languages.

6 Limitations

In this work, we examine the limits of LLMs' generalization through an orthographic normalization task using ICL approaches. While our conclusions are based on experimental results, a more comprehensive understanding of these limits may require testing across additional languages and tasks.

Although we cover the main orthographic norms used for this language, we excluded some lesser-used variants and phonological transcriptions. Phonological transcriptions were used for building the FSTs but not for the LLM approach. Additionally, our experiments kept the target norm constant, potentially simplifying the task for LLMs. Future work could explore different normalization directions across multiple norms.

Finally, while we demonstrated that combining LLMs with ICL and linguistic knowledge is a promising approach for orthographic normalization in under-resourced languages, practical considerations remain. The cost-benefit of using large models for relatively simple tasks should be evaluated, especially regarding accessibility for speakers and researchers working with these languages. Additionally, concerns about data handling in commercial systems must be addressed to ensure ethical and practical deployment.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can llms really learn to translate a low-resource language from one grammar book? *ICLR 2025*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Clarke, Roland Daynauth, Jason Mars, Charlene Wilkinson, and Hubert Devonish. 2024. GuyLingo: The Republic of Guyana creole corpora. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 792–798, Mexico City, Mexico. Association for Computational Linguistics.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Lázaro Margarita De la Vega. 2017. Aprendiendo otomí (hñähñu). *Ciudad de México, Comisión Nacional para el Desarrollo de los Pueblos Indígenas*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Ewald Hekking and Severiano Andrés de Jesús. 1989. *Diccionario español-otomí de Santiago Mexquititlán*, volume 22. Universidad Autónoma de Querétaro.

Nestor Hernández-Green. 2016. Misteriosas figurillas de barro de san jerónimo acazulco. *Tlalocan*, 21:19–48.

INALI. 2008. Catálogo de las lenguas indígenas nacionales: Variantes ling uísticas de méxico con sus atodenominaciones y referencias geoestadísticas. https://www.inali.gob.mx/clin-inali/.

Inali. 2014. *Njaua nt'ot'i ra hñähñu. Norma de escritura de la lengua hñähñu (otomí) de los estados de Guanajuato, Hidalgo, Estado de México, Puebla, Querétaro, Tlaxcala, Michoacán y Veracruz*. Instituto Nacional de Lenguas Indígenas (inaLi), SEP, Mexico.

Jesin James, Deepa P Gopinath, et al. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint arXiv:2410.07400*.

Cibu Johny, Lawrence Wolf-Sonkin, Alexander Gutkin, and Brian Roark. 2021. Finite-state script normalization and processing utilities: The Nisaba Brahmic library. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 14–23, Online. Association for Computational Linguistics.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Massimo Lusetti, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić, and Elisabeth Stark. 2018. Encoder-decoder methods for text normalization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 18–28, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar*

Languages, Varieties and Dialects, pages 115–127, Abu Dhabi, UAE. Association for Computational Linguistics.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Alexandra O'neil, Daniel Swanson, Robert Pugh, Francis Tyers, and Emmanuel Ngue Um. 2023. Comparing methods of orthographic conversion for bàsàá, a language of Cameroon. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 97–105, Dubrovnik, Croatia. Association for Computational Linguistics.

A Models and prompts

GPT-3.5 Turbo, GPT-4o, LLaMA 3.1, and *LLaMA 3.3* were used with default hyperparameters, setting a temperature of 0.2 via the API.

The LLaMA models were trained with 70B parameters, while the exact parameter count for GPT models is not publicly available.

- Prompt for OTS→INALI (few-shot)

```
Below are examples of orthographic conversions
    of strings from the OTS standard (State of
    Mexico Otomi) to the INALI standard for the
    Otomi language. Note that some loanwords
    retain their original orthography, and
    certain linguistic phenomena may affect the
    transformations.

Examples:

1. OTS: r'atsa noya ra sahagún: ndxkjua k'oi
    florentino. je̲m'i xiii, xe̲ni xiii (versión
    del náhuatl por ángel ma. garibay k.).
 INALI: r'atsa noya ra sahagún: ndxjua k'oi
    florentino. he̲m'i xiii, xe̲ni xiii (versión
    del náhuatl por ángel ma. garibay k.).

2. [...]

Task:

Using these examples as a guide, predict the
    INALI orthographic standardization for the
    following sentence. Return only the
    standardized sentence without any
    explanation.

OTS: ( test sentence )

INALI: [Your prediction here]
```

- Prompt for OTS+OTQ→INALI (few-shot)

```
Below are examples of orthographic conversions
    of strings from the OTS standard (State of
    Mexico Otomi) and the OTQ standard (Otomi of
     Queretaro) to the INALI standard for the
    Otomi language. Note that some loanwords
    retain their original orthography, and
    certain linguistic phenomena may affect the
    transformations.

Examples:

1. OTS: r'atsa noya ra sahagún: ndxkjua k'oi
    florentino. jem'i xiii, xeni xiii (versión
    del náhuatl por ángel ma. garibay k.).
   OTQ: r'atsa noya ra sahagún: ndxjua k'oi
    florentino. hem'i xiii, xeni xiii (versión
    del náhuatl por ángel ma. garibay k.).
 INALI: r'atsa noya ra sahagún: ndxjua k'oi
    florentino. hem'i xiii, xeni xiii (versión
    del náhuatl por ángel ma. garibay k.).

 2. [...]

 Task:

Using these examples as a guide, predict the
    INALI orthographic standardization for the
    following sentence. The sentence originates
    from either the OTS or OTQ standards, but
    its source is unspecified. Return only the
    standardized sentence without any
    explanation.

Sentence: ( test sentence )

INALI: [Your prediction here]
```

- Prompt for OTS→INALI (zero-shot)

```
Predict the INALI orthographic standardization
    for the following Otomi sentence written in
    the OTS standard (State of Mexico Otomi) (
    please return only the normalized sentences,
     no explanations). Note that some loanwords
    retain their original orthography, and
    certain linguistic phenomena may affect the
    transformations.

OTS: ( test sentence )

INALI: [Your prediction here]
```

- Prompt for OTS+OTQ→INALI (zero-shot)

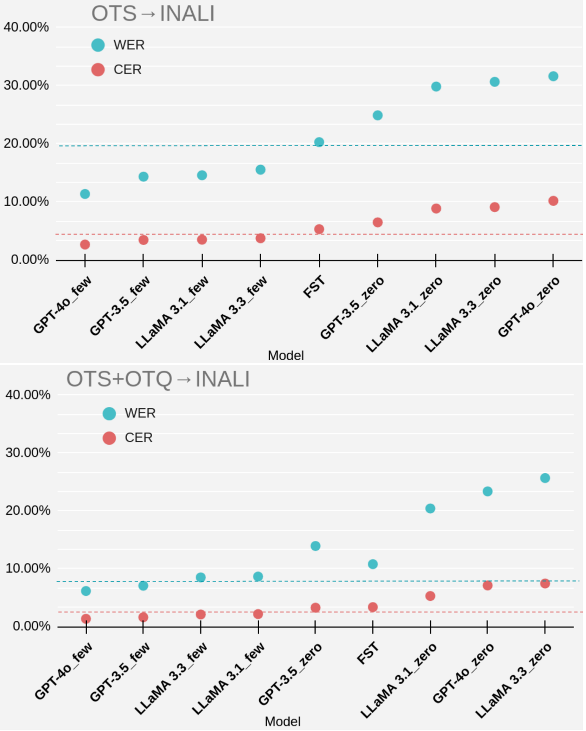
```
predict the INALI orthographic standardization
    for the following sentence. The sentence
    originates from either the OTS (State of
    Mexico Otomi) or OTQ (Queretaro Otomi)
    standards, but its source is unspecified.
    Return only the standardized sentence
    without any explanation.

Sentence: ( test sentence )

INALI: [Your prediction here]
```

B Original sentences vs gold standard

The following plots show the WER and CER for the different models. We have added a dotted line that indicates the error rate when comparing the source text and its INALI standardized form (gold standard), i.e., how dissimilar the source and target sentence are when no normalizer has been applied.


OTS→INALI
WER
CER
GPT-4o_few
GPT-3.5_few
LLaMA 3.1_few
LLaMA 3.3_few
FST
GPT-3.5_zero
LLaMA 3.1_zero
LLaMA 3.3_zero
GPT-4o_zero
Model
OTS+OTQ→INALI
WER
CER
GPT-4o_few
GPT-3.5_few
LLaMA 3.3_few
LLaMA 3.1_few
GPT-3.5_zero
FST
LLaMA 3.1_zero
GPT-4o_zero
LLaMA 3.3_zero
Model