Matina: A Culturally-Aligned Persian Language Model Using Multiple LoRA Experts

Sara Bourbour Hosseinbeigi[3], **MohammadAli Seif Kashani**[2], **Javad Seraj**[1],
Fatemeh Taherinezhad[1], **Ali Nafisi**[5], **Fatemeh Nadi**[1], **Iman Barati**[4],
Hosein Hasani[2], **Mostafa Amiri**[1], **Mostafa Masoudi**[1],

[1]University of Tehran, [2]Sharif University of Technology,
[3]Tarbiat Modares University, [4]Iran University of Science and Technology
[5]Bu-Ali Sina University,

Correspondence: s.bourbour@modares.ac.ir

Abstract

Large language models (LLMs) are powerful tools for a variety of applications, but to interact effectively with users, they must align with the cultural values and linguistic nuances of their audience. However, existing LLMs often fall short in adequately modeling underrepresented languages and cultures, such as Persian, limiting their applicability and acceptance. To address this, we construct diverse, high-quality datasets specifically tailored to Persian linguistic and cultural contexts, ensuring a more authentic and context-aware training process. Using these datasets, we develop Matina, a Persian-focused multi-expert model designed to embody Iranian cultural values and linguistic structures. Matina is trained by fine-tuning LLaMA3.1 8B-Instruct models across five domains: culinary, tourism, socio-culture, translation, and summarization. These experts are combined using a classifier to create a unified multi-expert system. By leveraging culturally aligned datasets, Matina outperforms baseline models in both task performance and user satisfaction, demonstrating the importance of data-driven cultural adaptation in LLM development.

1 Introduction

Large language models (LLMs) play a vital role in everyday life, assisting users by understanding their intentions and aligned with their demographics, beliefs, and culture. However, studies have shown that models like GPT and LLaMA often reflect Western (Ramezani and Xu, 2023; AlKhamissi et al., 2024) and American values (Johnson et al., 2022; Cao et al., 2023), largely due to the predominance of English in their training data (Navigli et al., 2023; Tao et al., 2024) and development in English-speaking regions (Masoud et al., 2023).

We define cultural alignment as ensuring AI systems reflect user values and norms, drawing from sociological and anthropological perspectives (Kroeber, 1952). This alignment helps prevent biases that lead to misrepresentation or cultural insensitivity, particularly in moral values, legal frameworks, and traditions. Despite advancements in multilingual LLMs, Persian remains underrepresented, even though it is spoken by over 110 million people[1]. More than just a language, Persian carries a rich cultural heritage shaped by centuries of literature and traditions. However, the lack of high-quality, culturally relevant data limits LLMs' ability to capture its nuances, reducing their effectiveness in Persian-speaking contexts.

To bridge the gap between LLMs and Persian culture and language, we introduce Matina, a Persian-focused language model specifically designed to reflect Iranian values and linguistic nuances. Matina is trained on a large dataset of 264,980,480 tokens and over 42,097 cultural documents for pretraining, supervised fine-tuning, and human alignment, ensuring that Persian cultural and linguistic structures are deeply embedded in its foundation.

Our work focuses on three representative cultural areas—-culinary domain, tourism, and socio-cultural contexts-—while also targeting essential language tasks for Persian users, including English-Persian translation and text summarization. To build our datasets, we employ a multi-faceted approach, including translating existing high-quality resources, collecting domain-specific Persian documents, generating culturally relevant content using GPT, and applying data augmentation techniques to ensure diversity and inclusively.

We fine-tune the LLaMA3.1 8B-Instruct model on these datasets in various experimental configurations and evaluate performance using LLM-based evaluations and human assessments. Our results show significant improvements in cultural under-

---

[1] `https://worldpopulationreview.com/country-rankings/what-countries-speak-farsi`

Findings of the Association for Computational Linguistics: ACL 2025, pages 20874–20889
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

standing, task-specific performance, and consistency in Persian-language outputs compared to baseline models, confirming the efficacy of our strategy.

The rest of the paper is organized as follows: Section 2 summarizes previous research on cultural and linguistic alignment approaches and datasets. Section 3 describes the dataset preparation and training strategies used. Section 4 presents evaluation results, and Section 5 discusses limitations and future work. Additional details are in the appendix.

2 Related Work

Large Language Models (LLMs) have progressed from monolingual (Radford et al., 2019; Raffel et al., 2020) to multilingual systems (Xue, 2020; Bai et al., 2023; Touvron et al., 2023), meeting a wide range of linguistic needs. However, cultural knowledge frequently falls behind linguistic proficiency, with models reflecting biases and stereotypes in training data (Johnson et al., 2022; Ramezani and Xu, 2023; Navigli et al., 2023). This is particularly problematic for conversational agents, which may produce culturally insensitive responses in multilingual regions.In this section, we look at alignment techniques, dataset construction, and advancements in the focal domains.

2.1 Alignment Methods

Aligning LLMs with cultural values involves training-free methods and fine-tuning. **Training-free** methods rely on well designed prompts to lead culturally appropriate replies. (Arora et al., 2022; AlKhamissi et al., 2024). Anthropological and cultural prompting techniques (Tao et al., 2024; Hwang et al., 2023) can help solve cultural differences within places that speak the same language. However, their success is heavily dependent on the model's size, architecture, and underlying dataset. **Training-based** methods use culturally relevant data for direct model training, leveraging pre-training or fine-tuning. Pre-training from scratch (Abbasi et al., 2023; Yoo et al., 2024) or continual training leverages large datasets and has only been adopted by a few researches. A more common approach, fine-tuning, adapts pre-trained models using labeled datasets to enhance cultural alignment for tasks like conversational reasoning (Wang et al., 2024a; Shi et al., 2024; Li et al., 2024b) or domain-specific applications like hate speech detection (Dehghan and Yanıkoğlu, 2024) and emotion recognition (Kim et al., 2024). These strategies underscore the importance of culturally relevant training data in boosting model performance and alignment.

2.2 Dataset Creation

There are three approaches for creating culturally aligned datasets: automatic, semi-automatic, and manual. Automatic methods leverage large-scale multilingual datasets like Wikipedia and CulturaX (Conneau, 2019; Nguyen et al., 2023), processing them to represent specific cultures (Lin and Chen, 2023; Abbasi et al., 2023). Advanced techniques use model-in-the-loop approaches to mine and structure cultural knowledge (Deshpande et al., 2022) or generate data with LLMs seeded with cultural contexts (Nguyen et al., 2024). For instruction datasets, LLMs adapt existing data to specific cultures, such as Putri et al. (2024), and Li et al. (2024a).

Semi-automatic methods combine human expertise with machine processing, enhancing quality and scalability. Techniques include manual curation refined by automation (Bai et al., 2024), cultural localization by native speakers (Alyafeai et al., 2024), and frameworks like STREAM that combine moral values and human oversight (Wang et al., 2024c).

Manually created datasets, though resource-intensive, are vital for aligning models with human values especially in low-resource contexts. Examples include HyperClovaX (Yoo et al., 2024) using Korean data and AceGPT and Jais (Sengupta et al., 2023; Huang et al., 2023) incorporating Arabic instructions.

2.3 Alignment Domain

Advances in culinary AI include food classification, recipe retrieval, and segmentation (Chen et al., 2017b; Kiourt et al., 2020; Chen et al., 2017a), supported by datasets like Food-101 and Recipe1M (Bossard et al., 2014; Salvador et al., 2017). Only a few studies focus on aligning models with gastronomy (Cao et al., 2024).

AI breakthroughs in tourism include personalized recommendations, itinerary planning, and sentiment analysis, supported by domain-specific datasets (Wei et al., 2024; Qi et al., 2024). Efforts focus on optimizing generative AI, multi-modal guidance, and knowledge graphs for better services (Hsu et al., 2024; Mo et al., 2023; Cadeddu et al.,

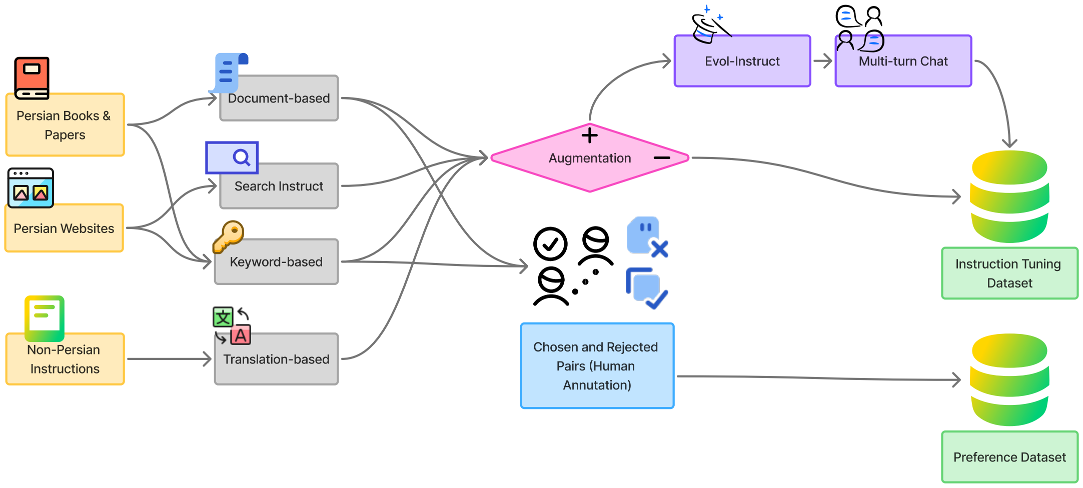


Figure 1: A comprehensive overview of methods, processes, and augmentation strategies for constructing culturally aligned instruction tuning and preference optimization datasets.

2023; Wang et al., 2024d). While datasets like Cultour aid cultural tourism fine-tuning (Wei et al., 2024), aligning LLMs with tourism knowledge still requires further specialization.

Biases in LLMs concerning social norms and values raise ethical challenges (Wang et al., 2023; Cahyawijaya et al., 2024). To address these biases, solutions include methods like few-shot prompting (Ornstein et al., 2022), context aggregation (Dognin et al., 2024), and ideological dataset tuning (Agiza et al., 2024).

LLMs have advanced machine translation beyond traditional machine translation models, succeeding in multilingual tasks but facing challenges in domain adaptation, linguistic nuance, and low-resource languages (Zhu et al., 2024; Zhang et al., 2024). Several ways have been proposed to overcome these challenges, including interactive machine translation (Wang et al., 2024b), retrieval-augmented translation (Zhu et al., 2024), and reinforcement learning-based fine-tuning.

To summarize, Persian attempts rely mainly on static models (Heidary et al., 2021; Farahani et al., 2021), with no interaction ability. LLMs have enhanced summarization and allow user interaction, but they still struggle with low-resource languages like as Persian due to their limited efficiency. To the best of our knowledge, no research has aligned open-source LLMs to enhance Persian summarization.

3 Methodology

As mentioned earlier, our work focuses on three culture-representative domains and two task-based domains that could benefit the Persian-speaking community. We first define each domain in Section 3.1, then detail the data creation methods employed for each area, followed by an overview of the training methodologies applied.

3.1 Areas of Focus

Our model focuses on several key areas to align with Persian language and culture. The culinary culture area includes recipes, ingredients, techniques, tools, cultural influences, nutrition, presentation, and the science of food and beverage preparation, representing Persian gastronomy's long traditions. The tourism area explores travel-related content, such as cultural attractions, itineraries, accommodations, local cuisine, and recommendations for unique experiences. The socio-cultural domain prioritizes balanced representation, cultural sensitivity, and the elimination of biases, including political and religious prejudices, while reflecting varied cultural values and conventions.

Aside from cultural considerations, the translation task focuses on accurately translating content from one language to another, paying attention to idiomatic idioms. Finally, the summary assignment reduces extensive content into brief, coherent, and contextually appropriate summaries while retaining the substance of the original material.

3.2 Data Curation

In this section, we discuss the strategies and techniques that we adopt or develop for creating instruction-following and alignment data, as well as approaches for augmenting the data sets. Our primary goal is to leverage the originality and style notation of proprietary LLMs and connect them with cultural materials to ensure factual consistency in the resulting data. The majority of our data production procedures relied on reference materials rather than being entirely reference-free. The number of data samples generated using each method across the five domains is shown in Table 1, which also reflects the total volume of data created for training and Figure 1 illustrates our data creation pipelines.

Keyword-based data creation concentrated on the socio-cultural sphere, which included norms, rules, values, and beliefs. We crawled data from political and social websites and extract relevant keywords using Llama3.1-70B-Instruct. These keywords were then utilized to generate culturally relevant questions reflecting Iranian societal and cultural concerns.

For effective model alignment using reinforcement learning-like techniques, each question required at least two responses: a preferred "gold" answer to reinforce desirable behavior and a rejected answer to guide the model away from undesirable outputs.In contrast, rejected responses were produced using LLaMA3.1-70B with role prompts biased toward non-Iranian perspectives. This process resulted in a dataset of 35,000 instances. Human evaluators reviewed and refined the examples, ensuring linguistic and cultural accuracy by removing non-Persian content from the gold responses. Additionally, they rated each QA pair on a scale from 1 (low quality/incorrect) to 5 (high quality/accurate). Only high-quality instances were retained, with 30,660 examples scoring three or higher selected for alignment.

Document-based data curation served as the foundation of our approach, providing a reliable alternative when expert-collected data was unavailable. We selected culturally and domain-relevant books, websites, and annotated documents, which were then segmented into smaller, coherent sections. Utilizing multiple language models with tailored prompts, we generated instruction-response pairs while ensuring the model remained contextually grounded to mitigate hallucination.

Our methodology comprised three LLM-based strategies for dataset construction. First, an LLM extracted question-answer pairs from text while preserving contextual alignment. Second, inspired by the Alpaca dataset, we provided the model with sample instruction-response pairs to guide it in generating diverse, high-quality instructions tailored to Persian culture and language. Third, we employed role-based conditioning, where the model was assigned human-like roles relevant to the domain and tasked with generating both questions and answers from given text chunks. To enhance accuracy and coherence, a separate LLM, acting as a subject-matter expert, reviewed and refined the generated responses.

After generating datasets using all the three methods, we applied an LLM-as-Judge framework to evaluate instructions based on contextual independence, complexity, practicality, informativeness, clarity, and creativity. High-scoring instructions were revised, with 75% of questions adapted to an informal tone for a more natural linguistic distribution, and 25% of responses adjusted for consistency. Answers were also reformatted into bullet points using markdown syntax to improve readability.

For socio-cultural data, sourced from books and oral materials, rejected responses were generated by prompting Llama3.1-70B with an Iranian persona. This method produced a total of 30,720 data instances for the domain.

Multi-turn Dialogue Augmentation plays a crucial role in improving our model's capability to manage complex, dynamic interactions. We created a pipeline in which two GPT-4o-mini models participated in 3-5 dialogue exchanges, imitating user interactions with personas such as "curious" or "skeptical." These characters were created using carefully developed prompts to ensure diversity and realism in dialogue scenarios.

Conversations began with prompts from our single-turn datasets (primarily document-based) to maintain a natural flow. For topics requiring factual accuracy, relevant document context was incorporated to ground the dialogue. This approach improved the model's ability to retrieve information and respond in a human-like, contextually appropriate manner. A total of 24,148 instances with an average of 3.7 turns were created for the culinary domain, while 15,660 instances with an average of 3.8 turns were generated for tourism.

Evol-Instruct-based Data Augmentation is integral to our methodology, leveraging Evol-Instruct (Xu et al., 2023) alongside the capabilities of GPT-4o-mini to enhance and expand our dataset. This approach is systematically applied to most of our generated data in order to introduce greater diversity and complexity, enriching the linguistic and semantic depth of the training material.

For factual topics, we used a document-based strategy, providing structured context to ensure correctness and reduce hallucinations. By anchoring discussions in clear context, we improved the model's reliability, precision, and real-world applicability. Through this technique, we expanded our culinary, tourism, and socio-cultural datasets by 53,322, 49,480 and 29,140 instructions respectively.

SearchInstruct method enhances the previous approaches by combining web search and self-instruction strategies.Unlike the Self-Instruct method, this methodology takes only a few basic questions as initial seeds and does not require responses, making it considerably easier to utilize. Initially, an LLM is used to expand the quantity and diversity of these questions, and relevant material is acquired for each question using a web search. Then, using the in-context learning capability of LLMs, an appropriate answer to each question is generated.

To further align the culinary model's output with structured formats, such as markdown tables for ingredients, we sampled 6,380 instances from the augmented culinary datasets, refined by Llama3.1-70B-Instruct to form preferred responses. Rejected responses, on the other hand were taken from the reference SFT culinary model.

Translation-based Data Generation is a technique that allowed us to utilize parallel instruction-following instances. Initially, we translated publicly available high-quality English instruction-following datasets, such as ORCA (Mukherjee et al., 2023) and Ultra-Chat (Ding et al., 2023), into Persian using GPT-4o and GPT-4o-mini. We then created translation task instructions using pairs of translated instructions.

Publicly Available Datasets are another component of our curation methodology for generating instruction-following instances. We gather a range of publicly available NLP datasets and modify them by adding diverse instructions.

For further details on the data curation methods, please refer to Appendix A.

3.3 Model Training

To address the unique attributes and focal areas of each domain, we trained multiple expert models, with each expert specializing in a specific cultural area. The final model aggregates these experts, and a classifier routes the user's query to the most appropriate expert.

We selected Llama-3.1-Instruct-8B as the baseline model for training, as it includes more Persian tokens in its tokenizer than other models, such as Mixtral and earlier Llama versions. This improves the model's understanding of Persian words and context. For training each expert, we applied LoRA (Hu et al., 2021), a technique from the PEFT family of methods.

While continual pre-training some experts on relevant data, our primary focus is on the post-training process, which includes two stages: (A) supervised fine-tuning and (B) cultural alignment, discussed in section 3.3.2 and section 3.3.3, respectively. Further details on the classifier and its deployment can be found in section 3.4.

3.3.1 Pre-training

Pre-training allows models to acquire knowledge by training them on large, unstructured datasets for the task of next-token prediction. While pre-training typically requires massive amounts of data—often billions or trillions of tokens—continual pre-training can be conducted with smaller, domain-specific datasets.

To transfer political and social, tourism, and culinary knowledge to Llama3.1-8B, we trained the model on domain-specific datasets. The data selection was guided by tags generated using the InsTag method (Lu et al., 2023) on a large Persian corpus (Hosseinbeigi et al., 2025b). The datasets included a mix of web-based content, academic papers, and books.

A total of approximately 1.2 billion tokens was used to train the socio-cultural expert. The tourism model was pre-trained on about 60 million tokens, while the culinary expert was trained on approximately 15 million tokens.

3.3.2 Cultural Supervised Fine-tuning

Supervised Fine-Tuning (SFT) is a crucial process for refining pre-trained models, enabling them to adapt to specific tasks or domains. For each task or domain discussed in Section 3.1, a LoRA module is initialized and fine-tuned using Negative Log-Likelihood (NLL) loss and input masking during

Domain/Task	Number of Instances						
	Evol-Instruct	Multi-turn	SearchInstruct	Document-based	Keyword-based	Translation-based	Public
Culinary Culture	53,322	24,148	8,932	48,322	-	-	-
Tourism	-	15,660	7,560	32,854	-	-	-
Socio-culture	29,140	-	-	30,720	30,660	-	-
Translation	-	-	-	-	-	280,000	63,000
Summarization	-	-	-	10,000	-	-	93,200

Table 1: Overview of methods and the number of instances created for instruction-tuning and preference optimization dataset across various areas of focus.

the SFT phase.

The instruction-following step was applied to four areas of focus: *culinary culture*, *tourism*, *translation*, and *summarization*, respectively, using 93,328, 56,074, 340,000, and 103,200 instructions.

We did not use all the instances generated for the culinary domain. Instead, we conducted a careful quality and diversity assessment with the help of five human annotators. They reviewed samples from each sub-dataset and scored them based on criteria including diversity and the introduction of new information. Using these annotator scores and agreement levels, we selected the sub-datasets that demonstrated higher diversity and better quality, while excluding those that were overly similar or of lower quality. This process ensured that only the most valuable and varied culinary data contributed to the final training set.

Our main instruction-following data instances are derived from document-based instances, augmented using *Evol-Instruct* and *multi-turn dialogue generation* methodologies. For the translation task, we used a combination of publicly available parallel data and the *Translation-based Data Augmentation* methodology.

Training details as well as hyperparameter value for each LoRA module are provided in Appendix B.

3.3.3 Cultural Alignment

Model alignment and RLHF help capture subjective, context-dependent human preferences, ensuring responses are truthful, safe, and aligned with human values. This was particularly important for our experts in socio-cultural domains.

We tested alignment methods, including PPO (Schulman et al., 2017), DPO (Rafailov et al., 2024), and ORPO (Hong et al., 2024), which rely on specialized loss functions and paired data. DPO showed instability, while ORPO proved effective and was selected for alignment.

The socio-cultural model was aligned on a dataset of 65,630 instances comprising document-based and keyword-augmented data, as well as Evol-augmented content. Alignment was critical for this domain due to its focus on human values and morals.

The culinary model was also aligned using ORPO on 6,380 instances. As mentioned earlier, subsets of the gastronomy dataset was collected and used to run ORPO on the culinary SFT model. Though having semantic overlap, the alignment datasets had different preferred responses than to those used in the SFT phase.

3.4 Prompt Classifier

We developed a mixture of experts using a hard classification strategy, selecting a single expert model for each task rather than blending outputs. Instead of storing multiple models, we only store LoRA parameters for each fine-tuned expert, significantly reducing storage requirements while enabling dynamic task adaptation.

To choose the appropriate expert for a given query, we fine-tuned an XLM-RoBERTa classifier to categorize the content based on the prompt context. This classifier determines if the query belongs to a specialized domain (tourism, culinary culture, socio-culture, or translation) or an "other" category. Based on the classification, the corresponding LoRA parameters are loaded and merged with the base LLaMA model. For non-specialized queries, the base model remains unchanged.

This approach enhances computational efficiency and domain adaptability by decoupling task-specific fine-tuning and using a lightweight classifier to select the appropriate expert model for each query.

4 Model Evaluation

Evaluating cultural alignment in language models is challenging due to the lack of statistical benchmarks and established standards. To address this, we employed three evaluation methods: (1) Human as a Judge, (2) LLM as a Judge, and (3) Human Satisfaction in multi-experts Interaction. These methods aim to assess how effectively the models reflect the target culture's knowledge, norms, and values, with human preference serving as a central measure of performance compared to baseline models.

As described in section 3.3, we used Llama3.1-Instruct 8B as the base model for training the experts, making it our primary baseline. To validate our datasets and training procedures, we also fine-tuned Llama3.1-Instruct 70B and incorporated it into the same classifier-based system, establishing it as an additional baseline. Results for fine-tuning Gemma2 (Team et al., 2024) are also provided in Table 9 and Table 10.

4.1 Human as a Judge

As noted in section 1, proprietary LLMs like GPT-4 have limitations in low-resource cultural contexts and often exhibit biases toward specific norms (Johnson et al., 2022; Cao et al., 2023; Ramezani and Xu, 2023; AlKhamissi et al., 2024), making them unreliable as cultural evaluators. To address this, we employed human experts as judges to ensure the models successfully captured cultural nuances.

We constructed question-answer datasets for each domain, ensuring they were distinct from the training data to maintain reliability. Each dataset included gold responses to serve as benchmarks. Questions for each domain were answered by the corresponding expert models as well as the baseline models for evaluation.

Three human experts conducted a blind evaluation, comparing the answers from the models and baselines in a win-lose ranking format without knowing which answers belonged to which models. This ensured unbiased judgments. Rankings were based on domain-specific criteria, including accuracy, relevance, structure, hallucination, output length, factual correctness, language proficiency, and alignment with the gold answers. Each question was evaluated by at least three annotators, and majority voting was used to aggregate their preferences, helping to ensure robustness and reduce individual bias. Prior to the full evaluation, we measured inter-annotator agreement on smaller evaluation subsets, achieving a substantial agreement level ($\kappa \approx 0.72$), which supports the reliability of the human judgments.

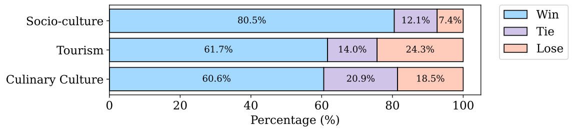


Figure 2: Comparative performance of Matina 8B and Llama3.1-Instruct-8B, based on human judgment. The evaluation categorizes results into three groups: Win (where Matina 8B outperforms), Tie (both models perform equally), and Lose (where Llama3.1-8B-Instruct outperforms).

4.2 LLM as a Judge

For tasks like summarization and translation, where cross-lingual knowledge transfer is less critical, we used GPT-4o mini as a judge after testing its familiarity with Persian linguistic and cultural nuances. Just like section 4.1, this method involved creating task-specific datasets distinct from the training data to ensure unbiased evaluation.

Models, including both experts and baselines, generated answers to these datasets. Unlike the human evaluation, gold responses were withheld, and GPT-4o mini ranked the outputs numerically based on task-specific criteria. For translation, these included accuracy, fluency, cultural relevance and consistency. For summarization, criteria focused on faithfulness, helpfulness, coverage, relevance, and informativeness. This approach ensured consistent and quantifiable evaluations.

4.3 Human Satisfaction in MoE Interaction

In addition to evaluating individual expert models, we conducted a third experiment to assess the effectiveness of the classifier and the system as a whole. This method also allowed us to evaluate domains or contexts potentially overlooked in the datasets, ensuring comprehensive coverage.

We deployed the multi-expert model along with

Areas of Focus	Number of Evaluation Instances
Culinary Culture	300
Tourism	300
Socio-culture	300
Translation	200
Summarization	200

Table 2: Summary of evaluation dataset statistics, categorized by cultural and linguistic focus areas.

Model	Area of Focus				
	Culinary Culture	Tourism	Socio-culture	Translation	Summarization
Llama3.1-Instruct 8B	3.21	3.49	4.14	5.12	4.65
Matina 8B	5.73	5.42	5.78	5.93	5.62
Llama3.1-Instruct 70B	4.18	4.97	4.85	6.72	5.50
Matina 70B	7.58	7.27	7.18	7.11	6.94

Table 3: Average ratings across five areas of focus for different models. Rates are from human participants and range from 1 to 10.

Model	Number of Like	Number of Dislike	Average Rating
Llama3.1-Instruct 8B	126	174	4.61
Matina 8B	261	187	6.01
Llama3.1-Instruct 70B	180	191	5.01
Matina 70B	294	135	6.68

Table 4: Final results from an experiment assessing people's preferences and ratings for model completions.

Model	Summarization Criteria for LLM as a Judge				
	Faithfulness	Helpfulness	Coverage	Relevance	Informativeness
Llama3.1-Instruct 8B	7.47	6.94	6.36	8.11	6.86
Matina 8B	8.50	8.48	8.27	8.72	8.52

Table 5: LLM as Judge Results: Evaluation of Matina 8B on summarization, with criteria including faithfulness, helpfulness, coverage, relevance, and informativeness.

both baselines, providing access to 20 evaluators with diverse educational backgrounds, including computer science, law, and graphic design. Evaluators interacted with the models, posing questions across various domains and tasks. For each query, users were presented with answers from four sources: Matina 8B, Llama3.1-Instruct 8B, Matina 70B, and Llama3.1-Instruct 70B. Evaluators first marked responses as "like" or "dislike" and then rated them on a scale of 1(incorrect answer) to 10(complete and coherent response) based on their accuracy and quality.

Ratings were collected and analyzed to gauge user satisfaction. Importantly, evaluators were not provided with specific ranking criteria, allowing for an assessment based purely on personal satisfaction.

4.4 Evaluation Dataset

To construct the evaluation datasets, we employed two distinct approaches:

1. **Subset Selection:** A subset of the instruction datasets was extracted, ensuring minimal overlap with the training instances. This approach was used to create the summarization dataset and parts of the culinary evaluation dataset.

Model	Translation Criteria for LLM as a Judge			
	Accuracy	Fluency	Cultural Relevance	Consistency
Llama3.1-Instruct 8B	7.53	6.76	6.24	7.14
Matina 8B	8.29	7.12	7.70	7.65

Table 6: LLM-as-Judge evaluation results for translation, comparing Matina 8B and Llama3.1-Instruct 8B across accuracy, fluency, cultural relevance, and consistency

2. **Seed-Based Generation:** Inspired by the self-instruct methodology (Wang et al., 2022), seeds were either sampled from the training set or crafted by evaluators to comprehensively cover all subdomains within each target domain. GPT-4o was then prompted to generate relevant questions based on these seeds.

After constructing the datasets, random samples were carefully reviewed by human evaluators to ensure they captured the cultural concerns and accurately represented the areas to be evaluated. Additionally, any problematic questions reported during the human evaluations were replaced with more relevant alternatives. Dataset statistics are presented in Table 2.

4.5 Evaluation Results

In the domains of culinary culture, tourism, and socio-culture, we employed the Human as a Judge evaluation method, as described in section 4.1. In this setup, we compared the performance of the Matina 8B-parameter model against the baseline. The results, shown in Figure 2, indicate that Llama3.1-Instruct 8B was preferred in only about 30% of cases across all three domains. In contrast, the Matina experts outperformed the baseline, with Matina models being favored in 80.5%, 60.6%, and 61.7% of cases in the socio-culture, culinary, and tourism domains, respectively. While the tourism and culinary experts were able to match the baseline model's responses in 14.0% and 20.9% of the questions, ties occurred only in 12.1% of the socio-culture domain evaluations. These results not only demonstrate the superiority of the Matina models but also highlight the lack of Persian cultural knowledge in the baseline model, particularly in areas such as social norms and politics.

For both summariztion and translation tasks, we applied LLM as a Judge evaluation using GPT-4o mini to evaluate the outputs of the Matina models against the baselines. The results are shown in Table 5 and Table 6. In summarization, as seen in Table 5, Matina's summarization expert significantly outperforms the Llama3.1 baseline across all five evaluation criteria. Although the difference in "relevance" was smallest, Matina provided more comprehensive summaries, making them more informative than those generated by Llama3.1. In translation as shown in Table 6 Matina's translation expert consistently outperformed the baseline across all evaluation criteria. the prompt for *LLM as a Judge* can be find in Appendix D

The results in Table 4 reflect human preferences in a more realistic setup, where participants interacted with the models. These results indicate that the Matina models achieved significantly higher satisfaction ratings from participants, with average scores surpassing those of the baseline models (Llama-3.1-Instruct 8B and Llama-3.1-Instruct 70B). This holds true for both the 8B and 70B parameter models, reinforcing the effectiveness of our training procedures and training datasets.

These findings underscore not only the success of the expert models but also the accuracy of our classifier in directing prompts to the most relevant expert. Notably, the Matina 8B multi-expert outperformed Llama3.1-Instruct 70B, despite having far fewer parameters. This demonstrates the crucial role of culturally aligned data in improving model performance.

The evaluation results on benchmarks for previous Persian language models can be found in the Matina LLM Benchmark (Hosseinbeigi et al., 2025a). This benchmark evaluates LLMs on Persian language across various domains and tasks.

5 Conclusion

In this paper, we introduced Matina, a Persian-focused multi-expert language model designed to bridge the gap between LLMs and Persian linguistic and cultural representation. Our approach involved creating a diverse instruction-tuning and preference optimization dataset across five key areas: culinary, tourism, socio-culture, translation, and summarization. To ensure cultural alignment, our data curation focused on culturally rich documents. We applied this dataset to train models such as LLaMA-3.1-Instruct-8B and Gemma-2-9B through a three-stage process: pretraining, supervised fine-tuning, and preference optimization. Additionally, we leveraged a multi-expert architecture to enhance inference accuracy and memory efficiency. Evaluation results demonstrate that Matina outperforms baseline models in both task performance and user satisfaction, underscoring the effectiveness of our approach in developing culturally aligned AI models.

Limitations

While Matina demonstrates significant advancements in cultural and linguistic alignment for Persian users, there are limitations in the scope and diversity of its datasets. Despite employing diverse data augmentation techniques, the datasets could benefit from broader coverage of queries and subdomains to further enrich their representation. Additionally, our focus on three cultural domains (culinary, tourism, and socio-culture) and two linguistic tasks (translation and summarization) does not fully cover the richness and complexity of Persian culture and language. Persian encompasses a wide variety of cultural expressions and linguistic challenges that remain unexplored in this study.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 2–12.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. 2024. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. 2024. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Andrea Cadeddu, Alessandro Chessa, Vincenzo De Leo, Gianni Fenu, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Salatino, and Luca Secchi. 2023. Leveraging knowledge graphs with large language models for classification tasks in the tourism domain. In *Deep Learning for Knowledge Graphs 2023. CEUR Workshop Proceedings Vol. 3559*, volume 3559.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*.

Y Cao, L Zhou, S Lee, L Cabello, M Chen, and D Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. arxiv. *Preprint posted online on March*, 31.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017a. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1771–1779.

Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. 2017b. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Somaiyeh Dehghan and Berrin Yanıkoğlu. 2024. Multi-domain hate speech detection using dual contrastive learning and paralinguistic features. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11745–11755.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. Stereokg: Data-driven knowledge graph construction for cultural knowledge and stereotypes. *arXiv preprint arXiv:2205.14036*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Pierre Dognin, Jesus Rios, Ronny Luss, Inkit Padhi, Matthew D Riemer, Miao Liu, Prasanna Sattigeri, Manish Nagireddy, Kush R Varshney, and Djallel Bouneffouf. 2024. Contextual moral value alignment through context-based aggregation. *arXiv preprint arXiv:2403.12805*.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.

Ebrahim Heidary, Hamïd Parvïn, Samad Nejatian, Karamollah Bagherifard, and Vahideh Rezaie. 2021. Automatic persian text summarization using linguistic features from text structure analysis. *Computers, Materials & Continua*, 69(3):2845–2861.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025a. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Heshaam Faili, Hamed Baghbani, Fatemeh Nadi, and Mostafa Amiri. 2025b. Matina: A large-scale 73b token persian text corpus. *Preprint*, arXiv:2502.09188.

Cathy H.C. Hsu, Guoxiong Tan, and Bela Stantic. 2024. A fine-tuned tourism-specific generative ai concept. *Annals of Tourism Research*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: Value conflict in gpt-3. arxiv. *arXiv preprint arXiv:2203.07785*.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16274–16289.

Chairi Kiourt, George Pavlidis, and Stella Markantonatou. 2020. Deep learning approaches in food recognition. *Machine learning paradigms: advances in deep learning-based technological applications*, pages 83–108.

Alfred L Kroeber. 1952. Culture: A critical review of concepts and definitions. *Peabody Museum*.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Baichuan Mo, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo, and Jinhua Zhao. 2023. Large language models for travel behavior prediction. *Preprint*, arXiv:2312.00819.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Multi-cultural commonsense knowledge distillation. *arXiv preprint arXiv:2402.10689*.

Joseph T Ornstein, Elise N Blasingame, and Jake S Truscott. 2022. How to train your stochastic parrot: Large language models for political texts. Technical report, Working Paper.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv preprint arXiv:2402.17302*.

Jinhu Qi, Shuai Yan, Wentao Zhang, Yibo Zhang, Zirui Liu, and Ke Wang. 2024. Research on tibetan tourism viewpoints information generation system based on llm. *Preprint*, arXiv:2407.13561.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F Chen. 2024a. Craft: Extracting and tuning cultural instructions from the wild. *arXiv preprint arXiv:2405.03138*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Yanshu Wang, Jinyi Zhang, Tianrong Shi, Dashuai Deng, Ye Tian, and Tadahiro Matsumoto. 2024b. Recent advances in interactive machine translation with large language models. *IEEE Access*, 12:179353–179382.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yuwei Wang, Enmeng Lu, Zizhe Ruan, Yao Liang, and Yi Zeng. 2024c. Stream: social data and knowledge collective intelligence platform for training ethical ai models. *AI & SOCIETY*, pages 1–9.

Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024d. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–20. ACM.

Qikai Wei, Mingzhi Yang, Jinqiang Wang, Wenwei Mao, Jiabo Xu, and Huansheng Ning. 2024. Tourllm: Enhancing llms with tourism knowledge. *Preprint*, arXiv:2407.12791.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How good are llms for literary translation, really? literary translation evaluation with humans and llms. *Preprint*, arXiv:2410.18697.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

Appendix

A Data Creation Details and Samples

Data plays a pivotal role in training and aligning language models, directly influencing how a model learns and responds to various queries. Factors such as data structure, size, diversity, and domain significantly impact the outcomes. To ensure the transfer of cultural and linguistic knowledge aligned with expert needs, we curated multiple datasets, each constructed using distinct methodologies tailored to specific objectives. After training the models on the created data, we carefully examined the shortcomings in the model's responses. In response, we adjusted certain details of the prompts and refined the data creation pipelines to address these issues and improve performance.

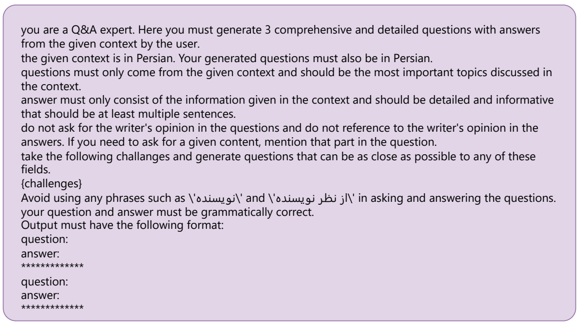
In this section, we outline the prompts used for dataset creation and provide illustrative samples to offer insight into the process.

Document-Based Datasets To develop datasets spanning multiple domains while ensuring diversity and human-like quality, we drew from a range of sources. Various techniques were applied to enhance the naturalness and relevance of the data, including: Question-Answer Pair Generation, Alpaca-Style Instruction Generation, Human-Like Instruction Generation.

Each method employed its own specialized prompting strategy. The Alpaca-style approach consisted of more than 20 independent prompts, each corresponding to a specific task. These prompts were designed to guide the model in generating high-quality instructions. However, each prompt was further optimized and customized based on the specific dataset and domain to ensure relevance and contextual accuracy.

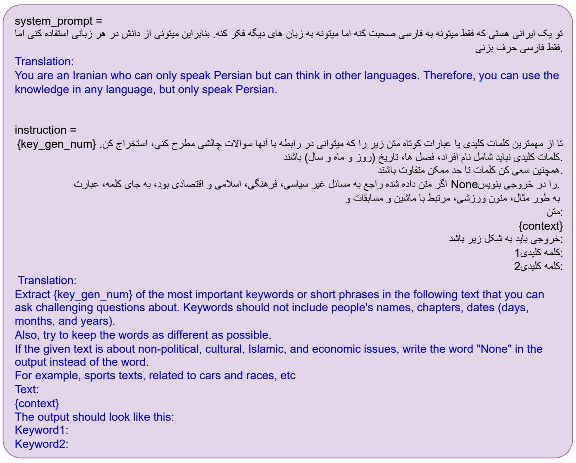
Given the extensive number of prompts involved in the process, presenting all of them in detail would be impractical. Therefore, as a representative example, we provide the base prompt for QA data generation in Figure 2. Additionally, Figure 3 illustrates the base prompt used in the LLM-as-Judge framework for evaluating instructions, and Figure 4 showcases one of the six prompts utilized in the revision process, specifically for cases where the question is informal, the response remains formal, and the answer is structured in a bullet-point format.

Books in the desired domains are a rich source for QA extraction. For socio-culture, books on cultural, social, and religious norms were selected and chunked by subsections. GPT-4o-mini was given a prompt to generate questions in the desired format. To improve results, we instructed the model not to reference specific sentences or use phrases like "according to the writer's opinion." The prompt is shown in fig. 3.



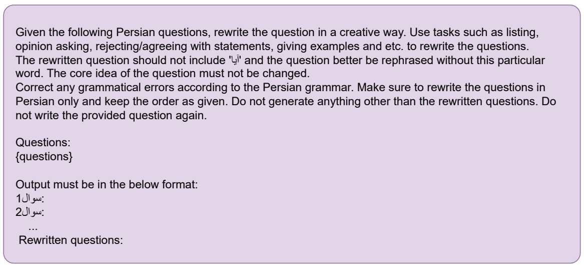
```
you are a Q&A expert. Here you must generate 3 comprehensive and detailed questions with answers
from the given context by the user.
the given context is in Persian. Your generated questions must also be in Persian.
questions must only come from the given context and should be the most important topics discussed in
the context.
answer must only consist of the information given in the context and should be detailed and informative
that should be at least multiple sentences.
do not ask for the writer's opinion in the questions and do not reference to the writer's opinion in the
answers. If you need to ask for a given content, mention that part in the question.
take the following challanges and generate questions that can be as close as possible to any of these
fields.
{challenges}
Avoid using any phrases such as \'نویسنده\' and \'از نظر نویسنده\' in asking and answering the questions.
your question and answer must be grammatically correct.
Output must have the following format:
question:
answer:
*************
question:
answer:
*************
```

Figure 3: Prompt to extract QA from socio-cultural books



```
system_prompt =
تو یک ایرانی هستی که فقط میتونه به فارسی صحبت کنه اما میتونه به زبان های دیگه فکر کنه. بنابراین میتونی از دانش در هر زبانی استفاده کنی اما
فقط فارسی حرف بزنی.
Translation:
You are an Iranian who can only speak Persian but can think in other languages. Therefore, you can use the
knowledge in any language, but only speak Persian.


instruction =
تا از مهمترین کلمات کلیدی یا عبارات کوتاه متن زیر را که میتوانی در رابطه با آنها سوالات چالشی مطرح کنی، استخراج کن. {key_gen_num}
کلمات کلیدی نباید شامل نام افراد، فصل ها، تاریخ (روز و ماه و سال) باشند.
همچنین سعی کن کلمات تا حد ممکن متفاوت باشند.
را در خروجی بنویس. None اگر متن داده شده راجع به مسائل غیر سیاسی، فرهنگی، اسلامی و اقتصادی بود، به جای کلمه، عبارت
به طور مثال، متون ورزشی، مرتبط با ماشین و مسابقات و
متن:
{context}
خروجی باید به شکل زیر باشد:
کلمه کلیدی1:
کلمه کلیدی2:

 Translation:
Extract {key_gen_num} of the most important keywords or short phrases in the following text that you can
ask challenging questions about. Keywords should not include people's names, chapters, dates (days,
months, and years).
Also, try to keep the words as different as possible.
If the given text is about non-political, cultural, Islamic, and economic issues, write the word "None" in the
output instead of the word.
For example, sports texts, related to cars and races, etc
Text:
{context}
The output should look like this:
Keyword1:
Keyword2:
```

Figure 4: Prompt to extract keywords from documents

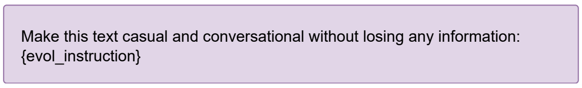


```
Given the following Persian questions, rewrite the question in a creative way. Use tasks such as listing,
opinion asking, rejecting/agreeing with statements, giving examples and etc. to rewrite the questions.
The rewritten question should not include 'آیا' and the question better be rephrased without this particular
word. The core idea of the question must not be changed.
Correct any grammatical errors according to the Persian grammar. Make sure to rewrite the questions in
Persian only and keep the order as given. Do not generate anything other than the rewritten questions. Do
not write the provided question again.

Questions:
{questions}

Output must be in the below format:
سوال1:
سوال2:
   ...
 Rewritten questions:
```

Figure 5: Prompt to diversify question structures

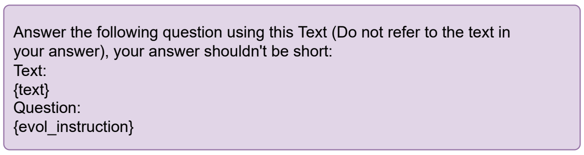
Keyword-Based Dataset To generate a set of keywords that represent the areas and subdomains to be covered in the dataset, and to which the model should be exposed, we extracted keywords from relevant content and created QAs for each keyword. The prompt used for this task is provided in fig. 4, where *key_gen_num* specifies the number of keywords to extract from the given context, and *context* refers to the text segment. We also observed that questions generated for the keywords were biased towards a certain structure and therefore further prompted the model with the questions and asked for revisions as in fig. 5.

Evol-Instruction Augmentation To implement the Evol-Instruct methodology for dataset enhancement, we adopted the prompting framework from (Xu et al., 2023), introducing modifications to ensure all generated content remained strictly within the Persian language domain, thereby avoiding cross-linguistic contamination. Our approach involved a two-step process: first, a specialized prompt (Figure 6a) was applied to transform the generated instructions into a casual, conversational format while retaining their relevance and instructional intent. In the second step, another prompt (Figure 6b) was employed to generate answers for the reformulated questions based on the context provided. This structured adaptation ensured that the Evol-Instruct methodology was optimized for Persian-language dataset augmentation, maintain-



Make this text casual and conversational without losing any information:
{evol_instruction}

(a) Prompt to make instructions informal



Answer the following question using this Text (Do not refer to the text in your answer), your answer shouldn't be short:
Text:
{text}
Question:
{evol_instruction}

(b) Prompt to answer augmented instructions based on content

Figure 6: Prompts to augment dataset by Evol method

ing both linguistic fidelity and contextual relevance.

Multi-turn Dataset Augmentation To create our multi-turn chat dataset in the culinary domain using LLMs, we assigned a specific role to the user-LLM and prompted it to ask questions based on the given characteristics. These questions were designed to reflect the perspective and concerns of someone embodying the assigned role. The roles were originally defined in Persian, with their translations provided in the list below. The agent-LLM, on the other hand, was tasked with taking on the role of an expert chef, offering detailed answers tailored to the given context, if provided. Examples of this dataset for both the culinary and tourism domain are provided in fig. 7.

1. Continue the conversation according to the context. You are in the role of an Iranian user seeking to learn new things about cooking. Ask your questions informally and driven by natural curiosity.

2. Act as someone interested in learning the basics of cooking who is looking for simple techniques. Ask your questions clearly and simply, avoiding complex terms. Use an informal tone for your questions.

3. Take on the role of a dormitory student with limited cooking equipment, seeking recipes that can be prepared using simple pots and minimal facilities. Ask your questions based on these limitations, using informal language.

4. Assume the role of someone with dietary restrictions (like diabetes) looking for low-sugar and low-carbohydrate recipes. Ask your questions about reducing sugar and carbohydrate intake in dishes, using informal language.

5. Act as a busy individual with little time to cook, seeking quick and simple recipes that can be prepared in under 30 minutes. Focus your questions on time efficiency and simplicity, using informal language.

6. Take on the role of an athlete looking for high-protein, energy-boosting meals for post-workout. Ask your questions about nutrients, energy, and protein in recipes, using informal language.

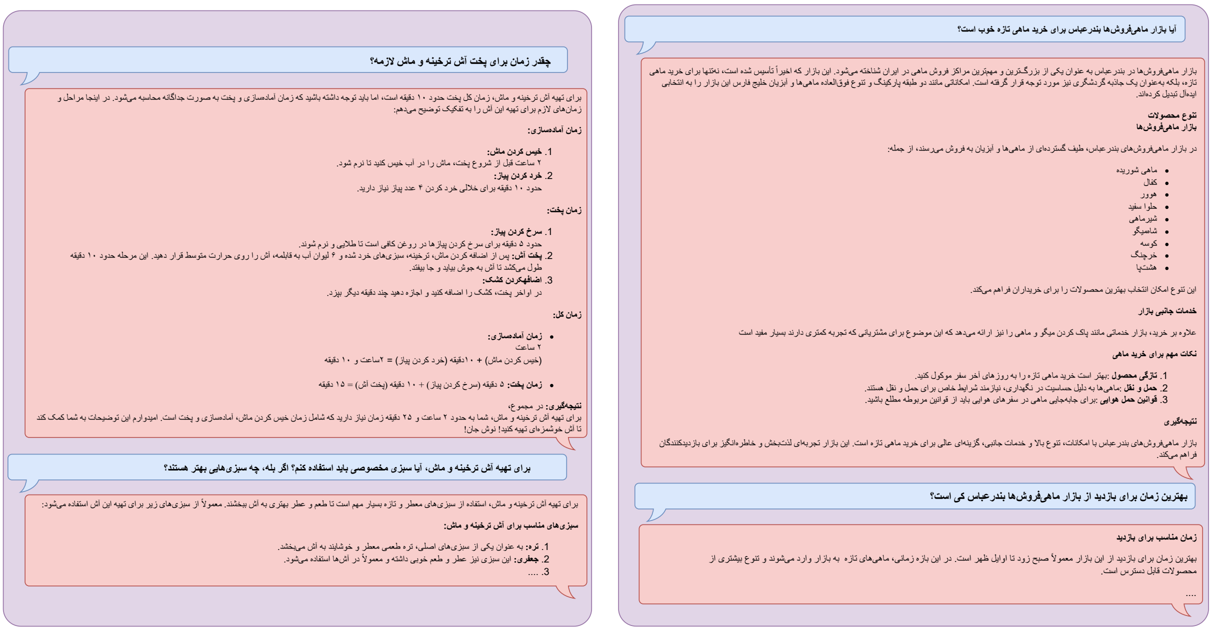
B Training Details

In the three stages of training, we systematically experimented with various hyperparameters and configurations, evaluating the model's performance and refining the settings based on the results. To optimize both resource and time efficiency, some of these experiments were conducted on smaller subsets of the data. As the dataset size increased, configurations were further adjusted accordingly. The parameters explored included learning rate, number of epochs, dataset size, dataset diversity, order of training data, alignment beta, and LoRA rank and alpha. These experiments aimed to achieve a balance where the model retained its prior knowledge while effectively integrating new information. Additionally, we sought to ensure the model could perform reasoning tasks, rather than merely memorizing the newly introduced knowledge.The models were trained using a data and model parallel setup across multiple GPUs—specifically, 4 NVIDIA A100 GPUs were used to train the 8B parameter models, and 6 A100 GPUs were employed to fine-tune LLaMA-70B.

After each training cycle with new configurations, the models were tested with 20 to 30 prompts. If no significant errors were observed in the model's responses, its outputs were then evaluated by human reviewers and compared to previous versions. The best-performing version from each stage was selected for further experiments and subsequent training.

Notably, different expert models were fine-tuned with distinct configurations, highlighting the flexibility and efficacy of our methodology for training multiple specialized models rather than a single aggregated model on the entire dataset. The final hyperparameter settings are detailed in table 7.

The query classifier represents a critical component of our model that requires dedicated training to achieve optimal performance. As mentioned in section 3.4, we used XLM-RoBERTa (Conneau, 2019)



چقدر زمان برای پخت آش ترخینه و ماش لازمه؟

برای تهیه آش ترخینه و ماش، زمان کل پخت حدود ۱۰ دقیقه است، اما باید توجه داشته باشید که زمان آماده‌سازی و پخت به صورت جداگانه محاسبه می‌شود. در اینجا مراحل و زمان‌های لازم برای تهیه این آش را به تفکیک توضیح می‌دهم:

زمان آماده‌سازی:

1. **خیس کردن ماش:**
۲ ساعت قبل از شروع پخت، ماش را در آب خیس کنید تا نرم شود.
2. **خرد کردن پیاز:**
حدود ۱۰ دقیقه برای خلالی خرد کردن ۴ عدد پیاز نیاز دارید.

زمان پخت:

1. **سرخ کردن پیاز:**
حدود ۵ دقیقه برای سرخ کردن پیازها در روغن کافی است تا طلایی و نرم شوند.
2. **پخت آش:** پس از اضافه کردن ماش، ترخینه، سبزی‌های خرد شده و ۶ لیوان آب به قابلمه، آش را روی حرارت متوسط قرار دهید. این مرحله حدود ۱۰ دقیقه طول می‌کشد تا آش به جوش بیاید و جا بیفتد.
3. **اضافهکردن کشک:**
در اواخر پخت، کشک را اضافه کنید و اجازه دهید چند دقیقه دیگر بپزد.

زمان کل:

- **زمان آماده‌سازی:**
۲ ساعت
(خیس کردن ماش) + ۱۰دقیقه (خرد کردن پیاز) = ۲ساعت و ۱۰ دقیقه
- **زمان پخت:** ۵ دقیقه (سرخ کردن پیاز) + ۱۰ دقیقه (پخت آش) = ۱۵ دقیقه

نتیجه‌گیری: در مجموع،
برای تهیه آش ترخینه و ماش، شما به حدود ۲ ساعت و ۲۵ دقیقه زمان نیاز دارید که شامل زمان خیس کردن ماش، آماده‌سازی و پخت است. امیدوارم این توضیحات به شما کمک کند تا آش خوشمزه‌ای تهیه کنید! نوش جان!

برای تهیه آش ترخینه و ماش، آیا سبزی مخصوصی باید استفاده کنم؟ اگر بله، چه سبزی‌هایی بهتر هستند؟

برای تهیه آش ترخینه و ماش، استفاده از سبزی‌های معطر و تازه بسیار مهم است تا طعم و عطر بهتری به آش ببخشند. معمولاً از سبزی‌های زیر برای تهیه این آش استفاده می‌شود:

سبزی‌های مناسب برای آش ترخینه و ماش:

1. **تره:** به عنوان یکی از سبزی‌های اصلی، تره طعمی معطر و خوشایند به آش می‌بخشد.
2. **جعفری:** این سبزی نیز عطر و طعم خوبی داشته و معمولاً در آش‌ها استفاده می‌شود.
3. ....

آیا بازار ماهی‌فروش‌ها بندرعباس برای خرید ماهی تازه خوب است؟

بازار ماهی‌فروش‌ها در بندرعباس به عنوان یکی از بزرگ‌ترین و مهم‌ترین مراکز فروش ماهی در ایران شناخته می‌شود. این بازار که اخیراً تأسیس شده است، نه‌تنها برای خرید ماهی تازه، بلکه به‌عنوان یک جاذبه گردشگری نیز مورد توجه قرار گرفته است. امکاناتی مانند دو طبقه پارکینگ و تنوع فوق‌العاده ماهی‌ها و آبزیان خلیج فارس این بازار را به انتخابی ایده‌آل تبدیل کرده‌اند.

تنوع محصولات
بازار ماهی‌فروش‌ها

در بازار ماهی‌فروش‌های بندرعباس، طیف گسترده‌ای از ماهی‌ها و آبزیان به فروش می‌رسند، از جمله:

- ماهی شوریده
- کفال
- هوور
- حلوا سفید
- شیرماهی
- شاه‌میگو
- کوسه
- خرچنگ
- هشت‌پا

این تنوع امکان انتخاب بهترین محصولات را برای خریداران فراهم می‌کند.

خدمات جانبی بازار

علاوه بر خرید، بازار خدماتی مانند پاک کردن میگو و ماهی را نیز ارائه می‌دهد که این موضوع برای مشتریانی که تجربه کمتری دارند بسیار مفید است

نکات مهم برای خرید ماهی

1. **تازگی محصول** :بهتر است خرید ماهی تازه را به روزهای آخر سفر موکول کنید.
2. **حمل و نقل** :ماهی‌ها به دلیل حساسیت در نگهداری، نیازمند شرایط خاص برای حمل و نقل هستند.
3. **قوانین حمل هوایی** :برای جابه‌جایی ماهی در سفرهای هوایی باید از قوانین مربوطه مطلع باشید.

نتیجه‌گیری

بازار ماهی‌فروش‌های بندرعباس با امکانات، تنوع بالا و خدمات جانبی، گزینه‌ای عالی برای خرید ماهی تازه است. این بازار تجربه‌ای لذت‌بخش و خاطره‌انگیز برای بازدیدکنندگان فراهم می‌کند.

بهترین زمان برای بازدید از بازار ماهی‌فروش‌ها بندرعباس کی است؟

زمان مناسب برای بازدید

بهترین زمان برای بازدید از این بازار معمولاً صبح زود تا اوایل ظهر است. در این بازه زمانی، ماهی‌های تازه به بازار وارد می‌شوند و تنوع بیشتری از محصولات قابل دسترس است.

....

(a) Culinary Multi-turn dataset - User role: Beginner Chef

(b) Tourism Multi-turn dataset

Figure 7: Example of Multi-turn datasets - Data samples have been truncated.

Domain/Task	Training Hyperparameters											
	Lora Rank	Lora Alpha	Pretrain			SFT			ORPO			
			lr	epochs	tokens	lr	epochs	samples	lr	beta	epochs	samples
Culinary	128	256	1e-4	3	15M	1e-4	3	93,328	5e-6	0.1	1	10,900
Tourism	256	512	1e-4	4	60M	1e-4	6	56,074	-	-	-	-
Socio-culture	128	256	1e-4	1	1.2B	-	-	-	5e-6	0.07	1	90,520
Translation	32	64	-	-	-	1e-5	1	340,000	-	-	-	-
Summarization	32	64	-	-	-	1e-5	4	103,200	-	-	-	-

Table 7: Training hyperparameters for different expert.

for its strong performance with Persian-language text. The training dataset included 87,634 "Other", 85,872 "culinary", 57,415 "tourism", and 37,079 "socio-cultural" samples. To improve handling of out-of-distribution and task-specific cases, 20% of domain-specific data was relabeled as "Other." A class-weighted cross-entropy loss addressed class imbalance, enhancing the classifier's ability to generalize across varied prompts. The model was then trained with a learning rate of **0.00002** for **1** epoch. The performance evaluation on a specially curated test set achieved an impressive accuracy of **93.8%**.

C Training Gemma-9B

We train Gemma-9B in three stages. First, we pretrain it on 264,980,480 tokens. Next, we apply supervised fine-tuning, followed by preference optimization using the ORPO method. Unlike Matina, Gemma does not utilize a multi-expert model; instead, we fine-tune it using a single LoRA module. The evaluation results are presented in Appendix C.1 and Appendix C.2

Model	Area of Focus				
	Culinary Culture	Tourism	Socio-culture	Translation	Summarization
Llama3.1-Instruct 8B	3.21	3.49	4.14	5.12	4.65
Matina-Gemma 8B	5.73	5.42	5.78	5.93	5.62

Table 8: Average ratings across five areas of focus for different models. Rates are from human participants and range from 1 to 10.

Outcome	Percentage
Win (Matina-Gemma)	67%
Tie	27%
Lose (Matina-Gemma)	6%

Table 9: Human evaluation results comparing Matina-Gemma with the baseline model.

C.1 Human Evaluation

To assess the cultural alignment of the model, we conducted a human evaluation with native Persian speakers. The evaluators compared the responses generated by our model, **Matina-Gemma**, against those from the baseline model. The results are present in Table 8 and Table 9.

As can be seen in Table 8, Matina-Gemma significantly outperformed the baseline model, Llama3.1-Instruct 8B, across all five evaluated areas, with particularly strong improvements in culturally relevant categories such as Culinary Culture and Socio-culture. Furthermore, as shown in Table 9, human evaluators preferred Matina-Gemma's responses in 67% of cases, with only 6% favoring Llama3.1-Instruct. These results demonstrate the effectiveness of our fine-tuning approach in enhancing cultural alignment and contextual understanding for Persian-speaking users.

C.2 LLM-as-a-Judge Evaluation

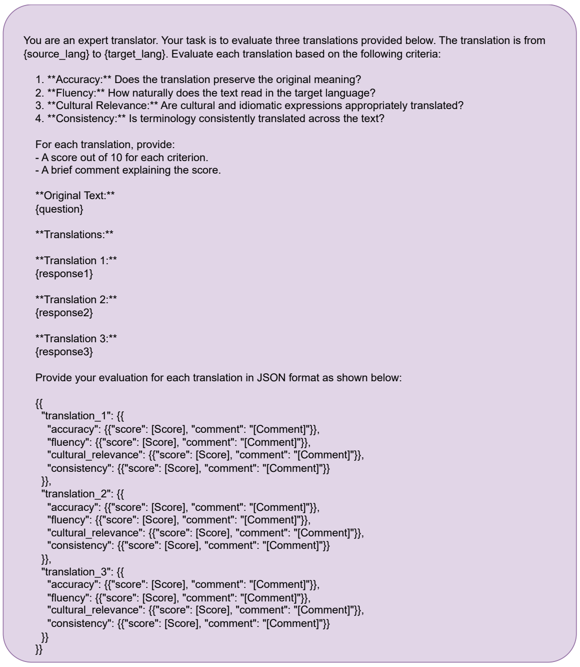
We also employed **GPT-4O-mini** as an automated judge for further evaluation, providing an additional layer of assessment beyond human judgment. This LLM-as-a-Judge setup offered a consistent and scalable means of comparison. Each evaluation criterion was scored as either 0 or 1, with final scores reflecting the overall performance. As shown in Table 10, Matina-Gemma achieved a slightly higher score (7.0) compared to the base model Gemma-9B-IT (6.8). Although the margin is modest, this result reinforces the human evaluation findings by confirming that our fine-tuning improves model performance even under automated, objective evaluation.

Model	Final Score
Gemma-9B-IT	6.8
Matina-Gemma	7.0

Table 10: LLM-as-a-Judge evaluation scores.

D LLM as a Judge Prompts

As described in Section 4.2, prompts were given to GTP4o-mini, along with the answer of multiple models to evaluate which model has superiority over the others. LLM as a judge was used for the domains of translation and summarization. For each task being assessed, a different criteria is outlined and explained for the model in the prompt and model responses are assessed based on these criteria. The prompt used for translation evaluation is depicted in Figure 8.



```
You are an expert translator. Your task is to evaluate three translations provided below. The translation is from
{source_lang} to {target_lang}. Evaluate each translation based on the following criteria:

  1. **Accuracy:** Does the translation preserve the original meaning?
  2. **Fluency:** How naturally does the text read in the target language?
  3. **Cultural Relevance:** Are cultural and idiomatic expressions appropriately translated?
  4. **Consistency:** Is terminology consistently translated across the text?

  For each translation, provide:
  - A score out of 10 for each criterion.
  - A brief comment explaining the score.

  **Original Text:**
  {question}

  **Translations:**

  **Translation 1:**
  {response1}

  **Translation 2:**
  {response2}

  **Translation 3:**
  {response3}

  Provide your evaluation for each translation in JSON format as shown below:

  {{
   "translation_1": {{
    "accuracy": {{"score": [Score], "comment": "[Comment]"}},
    "fluency": {{"score": [Score], "comment": "[Comment]"}},
    "cultural_relevance": {{"score": [Score], "comment": "[Comment]"}},
    "consistency": {{"score": [Score], "comment": "[Comment]"}}
   }},
   "translation_2": {{
    "accuracy": {{"score": [Score], "comment": "[Comment]"}},
    "fluency": {{"score": [Score], "comment": "[Comment]"}},
    "cultural_relevance": {{"score": [Score], "comment": "[Comment]"}},
    "consistency": {{"score": [Score], "comment": "[Comment]"}}
   }},
   "translation_3": {{
    "accuracy": {{"score": [Score], "comment": "[Comment]"}},
    "fluency": {{"score": [Score], "comment": "[Comment]"}},
    "cultural_relevance": {{"score": [Score], "comment": "[Comment]"}},
    "consistency": {{"score": [Score], "comment": "[Comment]"}}
   }}
  }}
```

Figure 8: Prompt to evaluate translation results with LLM (LLM as a judge).