TalkFa: A Unified Benchmark for Farsi Dialogue Generation and Understanding

Anonymous ACL submission

Abstract

Farsi, spoken by more than 120 million people, still lacks a unified benchmark for training and evaluating dialogue models; existing resources mainly target speech or generic Natural Language Understanding (NLU), leaving open-domain conversation, dialogue acts, emotions, and sentiment underexplored. We introduce TALKFA, a three-part benchmark: (1) WIKI-FADIAL, pairing 4.1K Wikipedia paragraphs with six-turn chats for knowledge-grounded dialogue generation; (2) DAILYDIALOG-FA, containing 6.6K everyday dialogues with act and emotion labels inherited from the original corpus; and (3) PLAYDIAL-FA, providing 2.0K theatrical exchanges labeled for sentiment. Generated or translated dialogues were verified by native speakers against their source texts. Parameter-efficient LORA adaptation of six LLAMA/MISTRAL checkpoints from 1B to 24B parameters improves WIKI-FADIAL performance by +13 ROUGE-L F_1 points and $\approx$ 39% BLEU-4. For classification tasks, FABERT achieves the best act performance (0.75), LORA-MISTRAL-7B the best emotion score (0.38), and XLM-R_{Base} remains a strong multilingual baseline with 0.73/0.34 macro-F_1. For sentiment, MISTRAL-24B surpasses the best encoder baseline (0.62 vs. 0.58 F_1). Using only half of the training data retains more than 90% of these gains. All datasets, annotation guidelines, code, and checkpoints will be released publicly.

1 Introduction

Progress in conversational AI depends on public benchmarks for dialogue understanding and generation. In English, datasets such as DAILYDIALOG (Li et al., 2017), EMPATHETICDIALOGUES (Rashkin et al., 2019), and TOPICALCHAT (Gopalakrishnan et al., 2019) support systematic evaluation and model development. By contrast, **Farsi**[1], spoken by more than 120M people, lacks a standard benchmark for text-based conversation. Existing Farsi resources mainly target speech, sentiment, or general NLU (e.g., PARSINLU (Khashabi et al., 2021)), leaving dialogue modeling underexplored. Although cross-lingual transfer with mBERT/XLM-R can mitigate data scarcity (Ruder et al., 2019), it often misses cultural nuances and informal conversational registers. Recent low-resource efforts such as NUSADIALOGUE (Purwarianti et al., 2025) show that LLM-assisted generation with human verification can bootstrap conversational resources, but no comparable benchmark exists for Farsi.

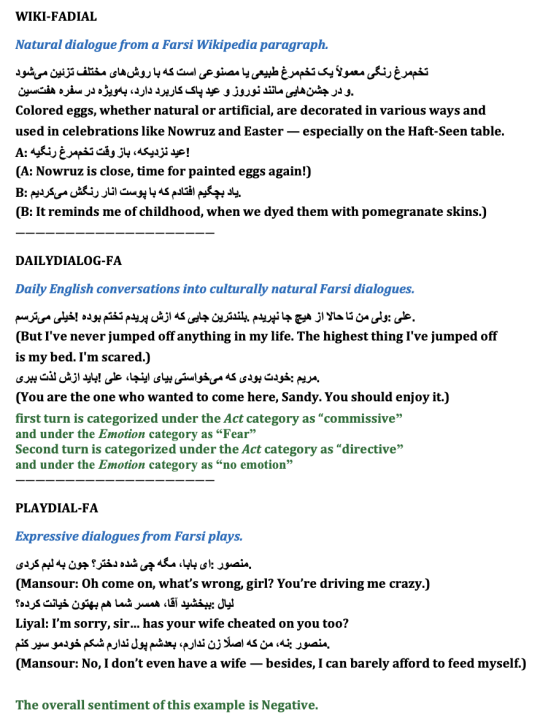

Figure 1: The sample dialogues from three sources in the TalkFa dataset; full dialogues are provided in Appendix A.

[1] Also called Persian; we use the terms interchangeably.

The absence of a *public multi-task benchmark* limits reproducible comparison of modeling techniques, systematic study of open-domain Farsi dialogue generation, and evaluation of turn-level pragmatic phenomena such as dialogue acts, emotions, and sentiment. This motivates our central question:

> Problem Statement
>
> **Can a culturally aware pipeline combining LLM-assisted data creation with native-speaker verification establish the first standard Farsi dialogue benchmark with clear, reproducible baselines?**

We investigate five research questions:

RQ1 (Generation, WIKI-FADIAL) How do instruction-tuned LLaMA/Mistral models perform before and after LoRA fine-tuning on knowledge-grounded Farsi dialogue generation?

RQ2 (Acts/Emotion, DAILYDIALOG-FA) Across ten encoder backbones and three training regimes—full fine-tuning, frozen backbone+MLP, and zero-shot classification—which yields the best macro-F_1 for dialogue-act and emotion tagging?

RQ3 (Sentiment, PLAYDIAL-FA) How do the same encoders compare on theatrical dialogue sentiment classification?

RQ4 (Zero-shot vs. LoRA–fine-tuned LLMs) How do six LLMs (1–24B) perform as zero-shot classifiers compared with their LoRA-fine-tuned counterparts?

RQ5 (Data-Efficiency Ablation) How does dialogue generation quality scale with the fraction of WIKI-FADIAL fine-tuning data (25%, 50%, 75%, 100%)?

Contributions. (i) We introduce *TalkFa*, the first unified benchmark for Farsi dialogue, comprising WIKI-FADIAL, DAILYDIALOG-FA, and PLAYDIAL-FA, covering knowledge-grounded, everyday, and dramatic dialogue (Figure 1). (ii) We present a reproducible pipeline combining LLM-assisted generation or translation with native-speaker verification and cultural adaptation. (iii) We establish baselines using six LoRA-adapted LLaMA/Mistral models (1–24B) and ten transformer encoders (100–600M) for dialogue generation, dialogue acts, emotions, and sentiment. (iv) We release all resources—datasets, annotation guidelines, code, LoRA adapters, and fine-tuned checkpoints—to support reproducible Farsi dialogue research.

Paper Organization. Section 2 reviews related work; Section 3 presents TALKFA; Section 4 describes experiments; Section 5 concludes.

2 Related Work

Dialogue Benchmarks in High-Resource Languages. English dialogue datasets have driven major advances in conversational AI. DAILYDIALOG (Li et al., 2017), EMPATHETICDIALOGUES (Rashkin et al., 2019), and TOPICALCHAT (Gopalakrishnan et al., 2019) cover emotion, grounding, and knowledge integration, while other benchmarks target persona modeling (Zhang et al., 2018), multi-domain interaction (Budzianowski et al., 2018), and commonsense grounding (Zhou et al., 2018).

Low-Resource Dialogue and Cross-Lingual Transfer. Efforts such as NUSADIALOGUE (Purwarianti et al., 2025) extend dialogue resources to underrepresented languages, while multilingual transfer methods (Ruder et al., 2019) with models such as mBERT and XLM-R partially alleviate data scarcity. However, they often miss the cultural and stylistic nuances central to natural Farsi dialogue.

Farsi Datasets and Dialogue Gaps. Existing Farsi resources mainly address sentiment (Shamsfard et al., 2024), question answering (Zeinalipour et al., 2024), and broader NLU tasks such as PARSINLU (Khashabi et al., 2021), with little attention to dialogue. No existing dataset supports end-to-end dialogue modeling across genres and stylistic variation, motivating a unified benchmark for Farsi dialogue generation and classification.

Synthetic Generation, Cultural Adaptation, and Modeling. Recent work uses LLMs for synthetic dialogue generation: Xu et al. (2023) generate Wikipedia-based dialogues with limited post-editing, while DIASYNTH (Suresh et al., 2025) uses chain-of-thought prompting for domain-specific conversations. Translation-based multilingual resources such as mT5 (Xue et al., 2021) and MTOP (Li et al., 2021) often lack cultural adaptation, producing unnatural outputs, whereas culturally grounded work such as ARABCULTURE (El-Haj et al., 2023) highlights the value of native-authored or culturally adapted data. Transformer models such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) have improved dialogue generation, but their effectiveness in low-resource, culturally adaptive settings remains underexplored. Dialogue systems are also increas-

ingly designed to adapt to user traits and personas (Chen et al., 2024), yet such progress remains limited in Farsi due to the lack of foundational dialogue resources.

To address these gaps—limited Farsi dialogue resources, insufficient cultural grounding, and the absence of a unified evaluation suite—we introduce TALKFA, a publicly released multi-task benchmark paired with rigorous LLM and encoder baselines.

3 TalkFa Creation

Built from three complementary sources, TALKFA covers knowledge-grounded dialogue (WIKI-FADIAL, 4.1K six-turn chats), everyday conversation (DAILYDIALOG-FA, 6.6K dialogues with turn-level act and emotion labels), and dramatic dialogue (PLAYDIAL-FA, 2.0K dialogues with dialogue-level sentiment).

3.1 WIKI-FADIAL: Context-Aware Casual Farsi Dialogues

We study *context-aware casual conversation generation in Farsi*: given an expository paragraph, the goal is to produce a short, informal dialogue that remains faithful to the source while sounding natural and idiomatic in everyday Farsi. To our knowledge, WIKI-FADIAL is the first Farsi dataset for this task, pairing encyclopaedic paragraphs with multi-turn chats in a conversational register.

Source discovery. We collect pages from ten curated indices of pedagogically valuable Farsi Wikipedia content, including *Featured Articles*, *Vital Articles*, and *Most-Viewed by Topic* (Wikipedia contributors, 2025b) (Wikipedia contributors, 2025a) (Wikipedia contributors, 2014) (Wikipedia contributors, 2009). This yields **8,894** pages spanning mathematics, history, biology, literature, and other core disciplines.

Content filtering. We keep only the headline paragraph of each article and apply three filters: (i) **length filter**, removing paragraphs with fewer than 100 words; (ii) **policy filter**, excluding sensitive or disallowed content; and (iii) **topicality/clarity filter**, preferring self-contained definitions, key facts, or short narratives. This yields **4,184** high-quality seed paragraphs.

Prompting and generation. Each paragraph is converted into a six-turn dialogue using GPT-4O; the full prompt is provided in Appendix B. Constraints include informal, culturally appropriate Farsi, exactly six turns, 1–20 words per turn, and a `reference` field in every turn to quote or summarize the source paragraph for traceability. The model returns one JSON object per paragraph with the key `"dialogue"`, whose value is an array of six `{speaker, text, reference}` triples.

Human post-editing. Two native Farsi annotators[2] with prior NLP experience post-edited each dialogue. Both were born and raised in Iran and are familiar with Farsi linguistic conventions and cultural context. In the first stage, one annotator corrected grammatical errors, normalized slang, inserted suitable discourse markers, and verified the factual consistency of all `reference` snippets against the source paragraph. In the second stage, the other annotator independently reviewed fluency, internal consistency, and faithfulness to the source text. Because each generated turn includes a `reference` field, both annotators explicitly verified that the dialogue remained grounded in the original paragraph and introduced no unsupported information. Remaining issues were resolved through adjudication. Detailed guidelines and examples are provided in Appendix C and Appendix D.

WIKI-FADIAL features. Table 1 summarizes corpus statistics for the source paragraphs (`Wiki-FaDial-Parag`) and paired dialogues (`Wiki-FaDial`). WIKI-FADIAL contains **4,184** dialogues and **16,499** turns, with a median of **81** tokens per dialogue (mean **81.79**)[3]. Compared with DAILYDIALOG-FA and PLAYDIAL-FA, it has similar length but higher lexical diversity (TTR words **0.11**; lemmas **0.09**) and modest length variance (token std. **8.46**) due to the fixed six-turn design.

Figure 2 shows that WIKI-FADIAL dialogues cluster around 80–90 tokens, while the source paragraphs are substantially longer. Figure 3 shows balanced POS coverage, including nouns, verbs, adpositions, pronouns, auxiliaries, and connectives[4].

Faithfulness without Copying. We compute embedding similarity between each generated dialogue and its paired Wikipedia paragraph using sentence embeddings from `bert-fa`. Average BLEU-4 is 0.005 and ROUGE-L F_1 is 0.10, indicating minimal surface overlap, while the embedding-cosine score of 0.66 confirms semantic faithfulness to the

[2] Both annotators are authors of this paper and carried out the annotation and verification voluntarily without financial compensation.

[3] Using the `bert-fa` tokenizer.

[4] We performed Farsi POS tagging with Hazm and fell back to Stanza if needed.

Metric	Wiki-FaDial-Parag	Wiki-FaDial	DailyDialog-FA	PlayDial-FA
Tokens (total)	1,109,523	342,200	635,942	200,158
Entries (dialogues)	NA	4,184	6,599	2,070
Turns (total)	NA	25,104	52,018	16,499
Mean Dialogue turn	NA	6	7.88	7.97
Mean tokens Dialogue	265.18	81.79	96.37	96.69
Median tokens Dialogue	235	81	95	91
Std. deviation (tokens)	118.64	8.46	26.64	46.53
Unique word types	45,157	28,448	13,624	12,025
Unique lemmas	39,477	24,394	10,664	9,646
TTR (words)	0.05	0.11	0.03	0.08
TTR (lemmas)	0.04	0.09	0.03	0.07

Table 1: Corpus statistics.

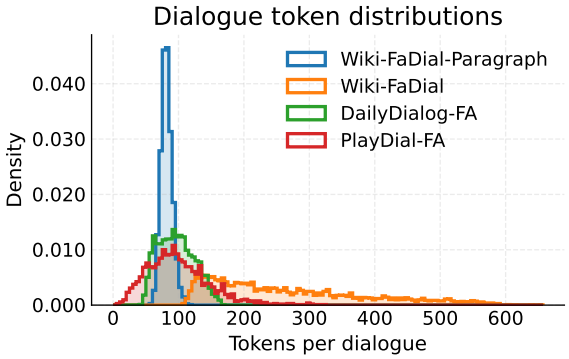

Figure 2: Dialogue token distributions across corpora.

source.

3.2 DAILYDIALOG-FA: Everyday Farsi Dialogues for Turn-Level Act & Emotion Classification

We study *turn-level* classification for two targets: (i) *dialogue acts*, the communicative function of a turn, and (ii) *emotions*, the affective stance conveyed by that turn. As a culturally localized counterpart to English DAILYDIALOG, DAILYDIALOG-FA provides Farsi supervision for these tasks.

Label Inventory. We use four dialogue-act categories—*commissive*, *directive*, *inform*, and *question*—and seven emotion categories—*anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, and *no emotion*. These inventories follow the original DAILYDIALOG taxonomy while preserving its original act and emotion labels in the translated Farsi dialogues.

Source Corpus and Machine Translation. We start from English DAILYDIALOG (Li et al., 2017). Using GPT-4O, each dialogue is translated into Farsi under constraints enforcing natural spoken Farsi, culturally appropriate names and references, preservation of turn order and flow, and alignment with Iranian politeness norms. Each translated dialogue mirrors the original structure. The full translation prompt is provided in Appendix E.

Cultural Localisation and Human Post-Editing. The same two annotators reviewed all translated dialogues in two passes. In the first, one annotator improved fluency and tone, localized cultural references (e.g., currencies, places, and measurement units), assigned culturally appropriate names, corrected grammatical or tense inconsistencies, and verified each dialogue against the original English version to preserve meaning, turn count, and conversational flow. In the second, the other annotator independently audited clarity, semantic fidelity, and coherence. Remaining issues were resolved through adjudication, and all dialogues were checked for punctuation, formatting, and valid JSON. Because annotators verified translations rather than assigning new labels, inter-annotator agreement metrics are not applicable; quality control instead relied on independent comparison with the original English dialogues. Guidelines and examples are provided in Appendix F and Appendix G.

DailyDialog-FA features. As shown in Table 1, DAILYDIALOG-FA contains **6,599** dialogues and **52,018** turns, totaling **635,942** tokens. Dialogues average **7.88** turns and **96.37** tokens (median **95**; std. dev. **26.64**). Lexical diversity reflects everyday conversation (types **13,624**; lemmas **10,664**; TTR by words **0.03**; TTR by lemmas **0.03**).

Token-per-dialogue distributions in Figure 2 show DAILYDIALOG-FA concentrated in short-to-medium lengths—ideal for turn-level supervision and stable batching—broader than the fixed six-turn WIKI-FADIAL design yet tighter than PLAYDIAL-FA. POS profiles in Figure 3 indicate balanced coverage of nouns and verbs with strong occurrence of pronouns, adpositions, auxiliaries,

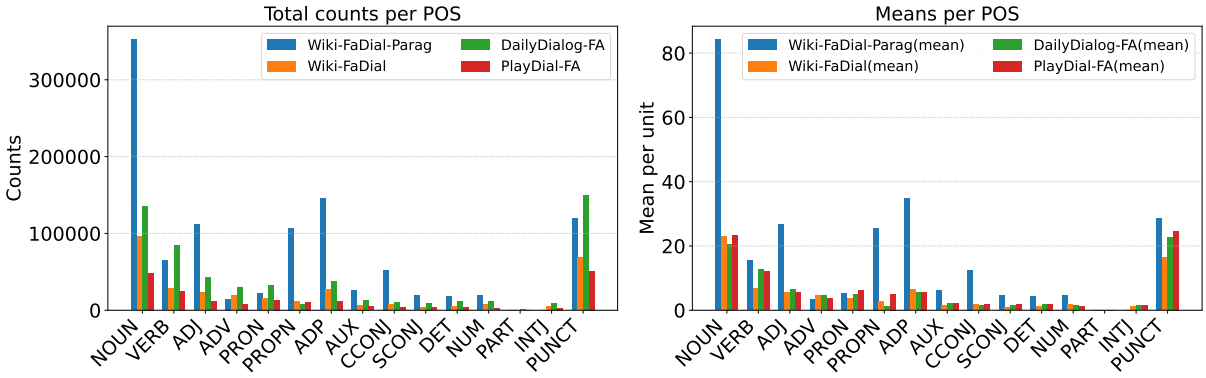

Figure 3: POS counts (bars) and per-unit means (lines) across corpora.

and connectives. This functional balance helps models rely on discourse cues (e.g., particles, pronouns, modal verbs) that are predictive of both *act* (e.g., directives vs. questions) and *emotion* (e.g., markers of surprise or politeness mitigation).

3.3 PlayDial-FA: Dramatic Farsi Dialogues for Turn-Level Sentiment Classification

We target *dialogue sentiment classification* in highly expressive dramatic dialogue. Theatre scripts intensify affect and employ archaic or rhetorical forms, making PlayDial-FA a useful complement to the everyday and knowledge-grounded registers covered by DailyDialog-FA and WikiFaDial.

Corpus Harvesting. We collected 30 public-domain Farsi plays (1940–2020) from the IranNLP Drama Archive (TakBook, 2025) in PDF format and converted them to machine-readable text via OCR (PDF2Go, 2025). Because the original plays already reflect culturally appropriate Farsi dialogue, the goal of cleaning was to extract spoken content without altering the original text. In the first pass, one annotator removed stage directions, formatting noise, and OCR artifacts, retaining only dialogue lines. In the second pass, the cleaned text was rechecked against the original scripts, and discrepancies were resolved by adjudication.

Segmentation and Reformulation. Using a structured GPT-4o prompt, scripts were segmented into multi-turn dialogues based on *scene continuity*; the full prompt is provided in Appendix H. To avoid verbatim reproduction of the original scripts while maintaining stylistic consistency, each segmented scene was then reformulated by GPT-4o. Both annotators reviewed the resulting dialogues in two passes, checking fluency, coherence, and consistency with the original scenes. The detailed editing and verification protocol is provided in Appendix I, and examples after segmentation, reformulation, and human verification are provided in Appendix J.

Sentiment Labels and Human Validation. Each dialogue is assigned one overall sentiment label from *negative*, *neutral*, or *positive*. Initial labels were automatically generated using GPT-4o, then independently reviewed by both annotators in light of the original play context; disagreements with the automatic label were corrected manually. Inter-annotator agreement reached **95.89%** with **Cohen's κ of 0.929**, indicating near-perfect agreement. Remaining disagreements were resolved through adjudication.

PlayDial-FA features. As shown in Table 1, PlayDial-FA comprises **2,070** dialogues and **25,104** turns (**200,158** tokens). Dialogues average **7.97** turns and **96.69** tokens (median **91**; std. dev. **46.53**). Lexical diversity exceeds everyday talk (types **12,025**; lemmas **9,646**; TTR words **0.08**; TTR lemmas **0.07**), reflecting richer style and emotive vocabulary.

Length distributions in Figure 2 show a wider tail than DailyDialog-FA, consistent with monologue stretches and dramatic pacing; the larger token variance (std. dev. **46.53**) indicates higher length variability that is useful for robustness testing. POS profiles in Figure 3 reveal strong coverage of verbs, nouns, and function words (pronouns, auxiliaries, connectives) that signal stance, (dis)agreement, and intensity—key cues for polarity decisions. Compared to DailyDialog-FA, PlayDial-FA exhibits more expressive adjectives/adverbs and interjections, aligning with theatrical affect and providing harder, complementary supervision for sentiment models.

4 Experiments

This section evaluates **TalkFa** along two complementary axes—*dialogue generation* and *utterance-level classification*—addressing the research questions introduced in Section 1.

Model families. We evaluate two groups of publicly available models suited to these tasks. *(A) Instruction-tuned LLMs (generation and zero-shot classification).* We use six open-source checkpoints spanning compact to mid-large scales (1–24B parameters), allowing us to examine how model size influences Farsi generation quality and zero-shot labeling: LLAMA-3.2-1B (Meta AI, 2024e), LLAMA-3.2-3B (Meta AI, 2024f), LLAMA-3.1-8B (Meta AI, 2024d), MISTRAL-7B (Mistral AI, 2024a), MISTRAL-NEMO-12B (Mistral AI, 2024b), and MISTRAL-SMALL-24B (Mistral AI, 2025). These cover three practical tiers: *tiny* (∼1–3B), *base* (∼7–8B), and *mid-large* (∼12–24B). *(B) Sentence encoders (supervised classification).* For DailyDialog-FA turn-level act and emotion prediction (and PlayDial-FA sentiment classification), we benchmark ten transformer encoders: the MULTILINGUAL-E5 family—SMALL (Microsoft, 2024c), BASE (Microsoft, 2024a), and LARGE (Microsoft, 2024b)—two Farsi-specific BERT derivatives (PARSBERT (HooshvareLab, 2020a) and FABERT (Shahid Beheshti University NLP Group, 2024)), and five widely used multilingual backbones (mBERT (Google, 2024), RoBERTa-Base (Meta AI, 2024a), RoBERTa-Large (Meta AI, 2024b), XLM-R-Base (Meta AI, 2019), XLM-R-Large (Meta AI, 2024c)). This setup enables comparison between (i) multilingual and Farsi-specific pretraining and (ii) different model sizes within the same architecture family (e.g., E5 small/base/large; XLM-R base/large).

Training regimes. We evaluate instruction-tuned LLMs and sentence encoders under complementary setups tailored to dialogue generation and classification.

LLMs (generation and zero-shot classification). We consider two modes: 1- ZERO-SHOT (PROMPT-ONLY): the released checkpoints are used without gradient updates, relying on task-specific instructions and label verbalizers at inference time. 2- LORA-ADAPTED: low-rank adapters are fine-tuned while all backbone weights remain frozen (Hu et al., 2021). For dialogue generation, a single adapter is trained on WIKI-FADIAL. For turn-level classification, separate adapters are trained for each corpus—DAILYDIALOG-FA (dialogue acts and emotions) and PLAYDIAL-FA (sentiment). Only the adapter and task-specific heads are updated, while backbone parameters remain fixed.

Encoders (supervised classification). We evaluate two standard regimes: 1- FROZEN-BACKBONE (MLP HEAD): the encoder is kept fixed while a two-layer classifier is trained on pooled embeddings. 2- FULL FINE-TUNING: all encoder parameters are updated jointly with the task head.

Training details. Compute. Experiments are conducted on 3×NVIDIA RTX A6000 GPUs (48,GB each) with mixed-precision (bf16), FlashAttention, and gradient checkpointing.

LLMs (LoRA fine-tuning). Seed 100; 3 epochs; AdamW optimizer; cosine learning rate schedule; base LR 1×10^{-4}; weight decay 1×10^{-4}; warmup ratio 0; max grad-norm 1.0. Per-device batch size is 4 with gradient accumulation 2, yielding an effective batch size of 24 sequences/step across 3 GPUs. LoRA configuration: rank 128, α 256, dropout 0.1.

Encoders (classification). *Frozen encoder + MLP head* (used for "MLP" rows): encoder frozen; 4-layer head 512, 256, 128, 64 → `LogSoftmax`; NLLLoss; 100 epochs; batch size 64; LR 1×10^{-5} (AdamW). *Full fine-tuning* (used for "Full" rows): encoder and head updated jointly for 3 epochs; batch size 32; LR 2×10^{-5} (AdamW). After this stage, the encoder is frozen and the same 4-layer MLP head is optionally re-optimized to stabilize classifier performance.

Evaluation. We evaluate models on dialogue classification and generation using standardized metrics; reporting details and a data-efficiency ablation follow.

Classification. For each encoder and LoRA-adapted LLM we report *accuracy*, *macro-precision*, *macro-recall*, and *macro-F_1*. **Dialogue generation.** We measure quality with four automatic metrics: (i) BLEU-4; (ii) ROUGE-L (F-measure); (iii) BERTSCORE-F (Microsoft, 2024d) as the scorer backbone; and (iv) BERTCOS—cosine similarity of sentence embeddings from `bert-fa` (HooshvareLab, 2020b). **Ablation.** To gauge data efficiency, we repeat LLM fine-tuning on $\{25\%, 50\%, 75\%, 100\%\}$ of WIKI-FADIAL and plot the BERTCos and ROUGE-L (F-measure) against data fraction to reveal saturation trends.

Model	ROUGE-L ↑		BLEU-4 ↑		BERTScore$_F$ ↑		BERTCos ↑	
	Base	+LoRA	Base	+LoRA	Base	+LoRA	Base	+LoRA
Llama-3.2-1B	0.449	0.650	0.322	0.572	0.340	0.623	0.824	0.959
Llama-3.2-3B	0.353	0.660	0.229	0.582	0.333	0.627	0.714	0.959
Mistral-7B	0.601	0.683	0.499	0.614	0.476	0.645	0.895	0.969
Llama-3.1-8B	0.655	0.668	0.578	0.596	0.544	0.640	0.944	0.962
Mistral-Nemo	0.613	**0.688**	0.515	0.614	0.436	0.653	0.893	**0.973**
Mistral-24B	0.546	0.687	0.434	**0.617**	0.453	**0.655**	0.919	0.972

Table 2: Base vs. +LoRA performance on WIKI-FADIAL (*test*). Bold marks the best score per metric across models.

Model	Act		Emotion		Sentiment	
	Base (P | R | F$_1$)	+LoRA (P | R | F$_1$)	Base (P | R | F$_1$)	+LoRA (P | R | F$_1$)	Base (P | R | F$_1$)	+LoRA (P | R | F$_1$)
Llama-1B	0.19|0.23|0.21	0.66|0.63|**0.57**	0.15|0.14|0.13	0.24|0.35|**0.24**	0.16|0.33|0.21	0.33|0.34|**0.23**
Llama-3B	0.46|0.41|0.39	0.63|0.60|**0.56**	0.24|0.24|0.19	0.29|0.34|**0.31**	0.33|0.45|0.38	0.58|0.52|**0.50**
Mistral-7B	0.53|0.23|0.30	0.71|0.69|**0.70**	0.24|0.21|0.18	0.51|0.33|**0.38**	0.56|0.08|0.13	0.60|0.58|**0.57**
Llama-8B	0.51|0.48|0.49	0.71|0.70|**0.70**	0.20|0.22|0.13	0.36|0.29|**0.31**	0.82|0.18|0.22	0.64|0.55|**0.55**
Mistral-12B	0.52|0.51|0.49	0.70|0.68|**0.68**	0.26|0.22|0.17	0.38|0.29|**0.32**	0.58|0.54|0.53	0.63|0.57|**0.57**
Mistral-24B	0.58|0.47|0.48	0.73|0.71|**0.72**	0.28|0.36|0.27	0.52|0.33|**0.37**	0.63|0.52|0.53	0.65|0.60|**0.62**

Table 3: Macro-averaged Precision | Recall | F$_1$ for **Act**, **Emotion**, and **Sentiment**, computed directly from the source CSVs (Base indicated by `base` in the filename; +LoRA by `frac1`). In each model–category pair, the higher F$_1$ is in bold. Higher is better.

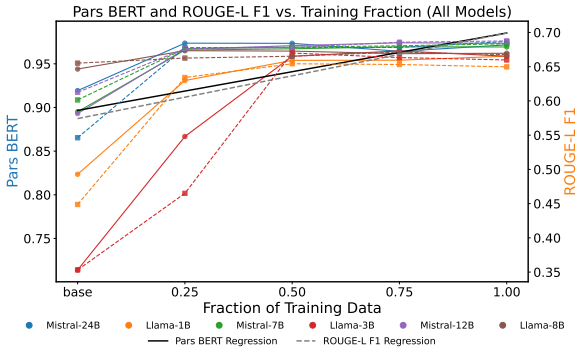

Figure 4: ParsBERT cosine (top) and ROUGE-L F1 (bottom) vs. fraction of the training data.

4.1 WIKI-FADIAL: Dialogue Generation

We evaluate each instruction-tuned LLM in its original **Base** form and after Low-Rank Adaptation (**+LoRA**). All models are fine-tuned on the WIKI-FADIAL [5] and decoding follows the six-turn schema described in Sec. 3.1.

Overall performance. Table 2 presents the comparison between the Base and +LoRA variants across four automatic metrics on the *test* set. Averaged over all checkpoints, +LoRA improves **ROUGE-L** by **+0.137** absolute, **BLEU-4** by **+0.170** absolute (≈ **+39%** relative over the Base average), **BERTScore$_F$** by **+0.210** absolute, and **BERTCos** by **+0.101** absolute.[6] Gains are especially pronounced on smaller models (e.g., LLAMA-3.2-1B and LLAMA-3.2-3B); larger models also benefit but with smaller deltas. Across models, the strongest +LoRA scores are: ROUGE-L and BERTCos for MISTRAL-NEMO, BLEU-4 and BERTScore$_F$ for MISTRAL-24B-SMALL.

Data efficiency. Figure 4 visualizes how performance scales with the fraction of fine-tuning data for all models, reporting ParsBERT cosine and ROUGE-L F1 at *Base*, 25%, 50%, 75%, and 100% of the training split. Across checkpoints, curves rise steeply from *Base* to *25%* and then flatten, indicating that roughly the first quarter to half of the corpus delivers most of the eventual gain, with diminishing returns thereafter.

4.2 DAILYDIALOG-FA: Turn-Level Act & Emotion—LLMs vs. Encoders

We compare six instruction-tuned decoder LMs (Base vs. +LoRA) and ten encoders (MLP vs. Full) for **Dialogue-Act** and **Emotion** classification (Tables 3–4).

Acts. Full-tuned encoders are strongest: FABERT attains **0.75** macro-F$_1$, with E5-BASE/LARGE, PARSBERT, XLM-R-BASE, and MBERT at **0.72–0.73**. LoRA-adapted LMs reach **0.70–0.72** (best: MISTRAL-24B **0.72**), far above Base (avg Act macro-F$_1$ improves (0.39 to 0.67), largely via higher recall).

Emotions. The task is harder for all models. LoRA LMs peak at **0.38** (MISTRAL-7B)

[5] In all experiments, we used a 90/5/5 split for training, validation, and testing.

[6] Averages computed across the six models in Table 2.

Model	Act		Emotion		Sentiment	
	MLP (P | R | F_1)	Full (P | R | F_1)	MLP (P | R | F_1)	Full (P | R | F_1)	MLP (P | R | F_1)	Full (P | R | F_1)
e5-small-100M	0.65|0.69|0.66	0.71|0.75|**0.72**	0.20|0.43|0.21	0.26|0.51|**0.30**	0.47|0.47|**0.47**	0.31|0.33|0.31
fabert-100M	0.67|0.72|0.68	0.73|0.77|**0.75**	0.25|0.55|0.28	0.29|0.52|**0.33**	0.49|0.50|**0.49**	0.16|0.33|0.21
roberta-base-100M	0.51|0.54|0.51	0.66|0.69|**0.67**	0.02|0.14|0.03	0.12|0.14|**0.13**	0.42|0.42|0.41	0.51|0.52|**0.51**
parsbert-110M	0.68|0.71|0.69	0.72|0.76|**0.73**	0.24|0.48|0.26	0.30|0.47|**0.33**	0.52|0.53|**0.49**	0.16|0.33|0.21
bert-base-multicased-200M	0.59|0.61|0.59	0.71|0.75|**0.72**	0.20|0.38|0.20	0.26|0.55|**0.29**	0.43|0.46|0.41	0.54|0.55|**0.54**
e5-base-300M	0.67|0.71|0.68	0.72|0.76|**0.73**	0.24|0.48|0.26	0.31|0.59|**0.36**	0.47|0.48|0.47	0.56|0.57|**0.56**
xlm-roberta-base-300M	0.54|0.56|0.54	0.72|0.76|**0.73**	0.12|0.14|0.13	0.30|0.60|**0.34**	0.50|0.50|**0.50**	0.31|0.38|0.32
roberta-large-400M	0.51|0.53|0.52	0.65|0.69|**0.66**	0.15|0.26|0.09	0.12|0.14|**0.13**	0.46|0.44|**0.44**	0.16|0.33|0.21
e5-large-600M	0.69|0.73|0.70	0.72|0.75|**0.73**	0.25|0.50|0.26	0.31|0.61|**0.36**	0.59|0.58|**0.58**	0.51|0.51|0.50
xlm-roberta-large-600M	0.53|0.56|0.54	0.72|0.76|**0.73**	0.12|0.14|**0.13**	0.02|0.14|0.03	0.57|0.52|**0.52**	0.40|0.50|0.45

Table 4: Corrected macro-averaged Precision | Recall | F_1 for **Act**, **Emotion**, and **Sentiment**, under two fine-tuning regimes (MLP vs. Full). In each triplet the higher F_1 is in bold.

and **0.37** (MISTRAL-24B); top encoders are E5-BASE/LARGE at **0.36**, followed by XLM-R-BASE **0.34** and Farsi-centric FABERT/PARSBERT **0.33**. Adaptation is the main driver: LoRA for decoder LMs and full fine-tuning for encoders. With LoRA, 7B LMs nearly match 12–24B. For acts, prefer Farsi-centric encoders (FaBERT/ParsBERT); for emotions, multilingual similarity encoders (Multilingual-E5). Beyond base size, scaling adds little; if one LM must cover both tasks, choose Mistral-7B with LoRA.

4.3 PLAYDIAL-FA: Sentiment Classification

We evaluate dialogue-level polarity (negative/neutral/positive) on expressive, theatrical Farsi exchanges. Decoder LMs are run in Base and +LoRA modes; encoders use MLP heads or full fine-tuning (Tables 3–4).

LLMs. LoRA markedly improves recall and overall F_1. Prompt-only models are unstable (often high precision, low recall), whereas +LoRA yields consistent gains: Mistral-24B reaches 0.62 macro-F_1, with Mistral-7B and Mistral-12B around 0.57, and Llama-8B at 0.55. Very small LMs benefit from LoRA but remain behind.

Encoders. Results are mixed: full fine-tuning helps some backbones (E5-Base: 0.56; mBERT: 0.54; RoBERTa-Base: 0.51), but shallow MLP heads outperform full fine-tuning on others (E5-Large + MLP: 0.58; XLM-R-Base + MLP: 0.50). Farsi-centric encoders (FaBERT, ParsBERT) degrade under full fine-tuning on this domain (≈ 0.21) yet are competitive with MLP heads (≈ 0.49). Primarily due to the small size of the dataset, full fine-tuning overfits and erodes generalizable features, whereas a shallow MLP head preserves them; in this setup, the MLP configuration performs better than full fine-tuning.

On PlayDial-FA, LoRA-adapted LMs slightly outperform the best encoder baselines (0.62 vs. 0.58) and offer robust recall on rhetorical language. For efficiency, E5-Base (full) or E5-Large (MLP) are strong choices. Common confusions include neutral vs. mild polarity and sarcasm; pooling across turns (hierarchical or mean pooling) helps.

5 Conclusion and Future Work

TALKFA provides a unified benchmark for Farsi dialogue across WIKI-FADIAL, DAILYDIALOG-FA, and PLAYDIAL-FA. A culturally aware pipeline combining LLM-assisted data creation, native-speaker verification, and LoRA/encoder fine-tuning yields a reproducible benchmark.

Findings: (RQ1) LoRA improves knowledge-grounded generation (+13 ROUGE-L F_1 points, $\sim$39% BLEU-4). (RQ2) FABERT gives the best act score (0.75 macro-F_1), LORA-MISTRAL-7B the best emotion score (0.38 macro-F_1), and XLM-R_{Base} remains a strong multilingual baseline (0.73/0.34 F_1). (RQ3) On sentiment, LoRA MISTRAL-24B reaches 0.62 macro-F_1 vs. 0.58 for the best encoder. (RQ4) LoRA consistently outperforms zero-shot, and small (1–3B) LLMs with LoRA rival larger base models. (RQ5) Using only 25–50% of WIKI-FADIAL recovers over 90% of the final gains.

Resources and Future Work: We release the datasets, guidelines, code, LoRA adapters/checkpoints, and baselines. Future work includes richer pragmatic annotations, targeted human evaluation, retrieval-grounded generation, speech/code-switching extensions, adaptation to Dari/Tajik, and robustness/safety suites with a public leaderboard.

6 Limitations

While our benchmark advances evaluation for Farsi dialogue, it also inherits constraints from its con-

struction and scope. We note four main limitations:

- **LLM-Assisted Data Creation Bias:** Relying on GPT-4o drafts followed by human verification and post-editing may introduce stylistic regularities that reduce linguistic variety compared with fully native-authored dialogues.
- **Structural Constraints in WIKI-FADIAL:** The fixed six-turn, 1–20-token format underrepresents longer, messier, or multi-party exchanges, potentially biasing models toward brief and orderly replies.
- **Cross-Lingual Mapping in DAILYDIALOG-FA:** Act and emotion inventories are inherited from English, and Farsi-specific pragmatics (e.g., ta'arof, indirectness, and honorifics) may not map cleanly.
- **Generalisation and Register Coverage:** The benchmark omits many real-world registers (e.g., social media slang, code-switching, and multi-party chats) and dialectal varieties such as Dari and Tajik, limiting out-of-domain robustness.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, and 1 others. 2018. Multiwoz: A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. *arXiv preprint arXiv:2405.17974*. Presented at LREC-COLING 2024.

Mahmoud El-Haj, Muhammad Abdul-Mageed, Abdel-Rahim Elmadany, and 1 others. 2023. Commonsense reasoning in arab culture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Google. 2024. google-bert/bert-base-multilingual-cased. HuggingFace model card. Multilingual BERT trained on 104 Wikipedia languages.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, and 1 others. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of Interspeech*, pages 1891–1895.

HooshvareLab. 2020a. Hooshvarelab/bert-base-parsbert-uncased. HuggingFace model card. Monolingual Persian BERT model (ParsBERT).

HooshvareLab. 2020b. Hooshvarelab/bert-fa-base-uncased. HuggingFace model card. Persian language model ParsBERT v2.0.

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. *arXiv preprint arXiv:2106.09685*.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, and 1 others. 2021. Parsinlu: A suite of language understanding challenges for persian. *arXiv preprint arXiv:2012.06154*.

Mike Lewis, Yinhan Liu, Naman Goyal, and 1 others. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, and 1 others. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*.

Zihan Li, Swaroop Sunkara, Xinyan Zhao, and 1 others. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.

Meta AI. 2019. Facebookai/xlm-roberta-base. HuggingFace model card. Multilingual RoBERTa pretrained on CommonCrawl (100 languages).

Meta AI. 2024a. Facebookai/roberta-base. HuggingFace model card. Pre-trained English RoBERTa-base model.

Meta AI. 2024b. Facebookai/roberta-large. HuggingFace model card. RoBERTa-large model introduced in the RoBERTa paper.

Meta AI. 2024c. Facebookai/xlm-roberta-large. HuggingFace model card. Large multilingual XLM-RoBERTa model.

Meta AI. 2024d. meta-llama/llama-3.1-8b-instruct. HuggingFace model card.

Meta AI. 2024e. meta-llama/llama-3.2-1b-instruct. HuggingFace model card.

Meta AI. 2024f. meta-llama/llama-3.2-3b-instruct. HuggingFace model card.

Microsoft. 2024a. intfloat/multilingual-e5-base. HuggingFace model card. Embedding model: 12 layers, embedding size 768.

Microsoft. 2024b. intfloat/multilingual-e5-large. HuggingFace model card. Embedding model: 24 layers, embedding size 1024; supports about 100 languages.

Microsoft. 2024c. intfloat/multilingual-e5-small. HuggingFace model card.

Microsoft. 2024d. microsoft/deberta-xlarge-mnli. HuggingFace model card. DeBERTa-XLarge model fine-tuned on MNLI.

Mistral AI. 2024a. mistralai/mistral-7b-instruct-v0.3. HuggingFace model card.

Mistral AI. 2024b. mistralai/mistral-nemo-instruct-2407. HuggingFace model card, developed with NVIDIA.

Mistral AI. 2025. mistralai/mistral-small-24b-instruct-2501. HuggingFace model card.

PDF2Go. 2025. Pdf2go: Convert pdf to text online. Online PDF conversion tool.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, and 1 others. 2025. Nusadialogue: Dialogue summarization and generation for underrepresented and extremely low-resource languages. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 82–100, Online. Association for Computational Linguistics.

Hannah Rashkin, Maarten Smith, Eric Sap, and 1 others. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Shahid Beheshti University NLP Group. 2024. sbunlp/-fabert. HuggingFace model card. Persian BERT-base model (FaBERT) trained on HmBlogs corpus.

Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and 1 others. 2024. Farseval-pkbets: A new diverse benchmark for evaluating persian large language models. *arXiv preprint arXiv:2504.14690*.

Sathya Krishnan Suresh, Mengjun Wu, Pranav Tushar, and Eng Siong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. *arXiv preprint arXiv:2409.19020*. Accepted at NAACL 2025.

TakBook. 2025. Takbook: Free online book download platform. Website.

Wikipedia contributors. 2009. Essential articles every wikipedia should have. https://fa.wikipedia.org.

Wikipedia contributors. 2014. Most viewed articles in persian wikipedia.

Wikipedia contributors. 2025a. 100 essential articles (persian wikipedia).

Wikipedia contributors. 2025b. Featured articles in persian wikipedia. https://fa.wikipedia.org.

Jingjing Xu, Wei Chen, Biao Tang, and Muhao Chen. 2023. Llm-based dialogue data generation for low-resource languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Linting Xue, Noah Constant, Adam Roberts, and 1 others. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kamayar Zeinalipour, Neda Jamshidi, Marco Maggini, and 1 others. 2024. Persianmcq-instruct: A comprehensive resource for generating multiple-choice questions in persian. *arXiv preprint arXiv:2403.09338*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Saizheng Zhang, Emily Dinan, Jack Urbanek, and 1 others. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Hao Zhou, Tom Young, Minlie Huang, and 1 others. 2018. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Full Dialogues of WIKI-FADIAL, DAILYDIALOG-FA, and PLAYDIAL-FA

Here, we present the complete dialogues from the three sources illustrated in Figure 1: WIKI-FADIAL (natural dialogues generated from Farsi Wikipedia paragraphs), DAILYDIALOG-FA (daily English conversations translated into culturally natural Farsi), and PLAYDIAL-FA (expressive multi-turn dialogues extracted from Farsi plays).

WIKI-FADIAL
Natural dialogue from a Persian Wikipedia paragraph.

تخم‌مرغ رنگی معمولاً یک تخم‌مرغ طبیعی (و گاه مصنوعی) است که آن را با روش‌های گوناگون تزئین می‌کنند و در جشن‌هایی مانند نوروز و عید پاک استفاده می‌شود. اغلب هنگام نوروز و در سفرهٔ هفت‌سین برای زیبایی به کار می‌رود. برای رنگ‌آمیزی آن از مواد طبیعی مانند لبو، پوست انار، زردچوبه و... یا رنگ‌های شیمیایی استفاده می‌کنند.تخم غاز، اردک و مرغ معمولاً خالی می‌شوند – در هر انتهای آن سوراخی ایجاد می‌شود و محتویات آن بیرون ریخته می‌شود. سپس تخم‌مرغ نقاشی‌شده، رنگ‌آمیزی‌شده و خشک می‌شود یا تزئین می‌گردد (با استفاده از تکنیک‌های مختلف). تزئین تخم‌مرغ به‌ویژه در کشورهای اروپای شرقی رواج دارد.

Colored eggs are usually natural (and sometimes artificial) eggs that are decorated using various methods and are used in celebrations such as Nowruz and Easter. They are often used during Nowruz on the Haft-Seen table for decoration. To dye them, natural materials such as beetroot, pomegranate peel, turmeric, and others, or chemical dyes are used.Goose, duck, and chicken eggs are usually emptied—holes are made at both ends of the egg, and the contents are removed. Then the egg is painted, dyed, and dried, or decorated using different techniques. Egg decoration is especially popular in Eastern European countries.

A: عید نزدیکه، باز وقت تخم مرغ رنگیه!
A: (*Nowruz is close—time for painted eggs again!)*

B: یاد بچگیم افتادم که با پوست انار رنگش میکردیم.
B: (It reminds me of childhood, when we dyed them with pomegranate skins.)

A: من همیشه تخم مرغ رو سوراخ میکردم که بتونم خالیش کنم.
A: (*I always poked holes in the eggs so I could empty them.)*

B: اروپای شرقیها استاد تزیینشن، واقعاً هنرمندن.
B: (*Eastern Europeans are masters at decorating them—they're true artists.)*

A: تو سفره هف تسین یه عالمه ازش میذارم، خیلی قشنگه.
A: (*I put lots of them on the Haft-Seen table—they look so pretty.)*

B: کاش امسال با رنگ طبیعی درستش کنیم، هم سالمتره هم اصیلتر.
B:(*I wish we'd make them with natural dyes this year—it's healthier and more authentic.)*

DAILYDIALOG-FA
Daily English conversations into culturally natural Persian dialogues.

سارا:سلام، مریم. امروز رنگت پریده، چی شده؟
(Sara: Hi, Mary. You look pale today. What's the matter with you?)

مریم:فکر کنم سرما خوردم. همه‌جای بدنم درد می‌کنه.
(Maryam: I think I've caught a cold. Now I'm aching all over.)

سارا:بهتره یه کم دارو بخوری و حسابی استراحت کنی. امیدوارم زودتر حالت خوب بشه.
(Sara: You'd better take some medicine and have a good rest. I hope you'll be yourself pretty soon.)

مریم:ممنونم. فکر کنم باید برم دکتر رو ببینم.
(Maryam: Thank you. I suppose I should go see a doctor.)

PLAYDIAL-FA

Extract expressive dialogues from Persian plays .

فیگارو: این همه احساس، فقط یه ابراز علاقهٔ سادەست؟
(Figaro: All this emotion — is it just a simple show of affection?)

سوزانا: می‌خوای محبت ببینی؟ خیلی خب، بیا جلو!
(Suzana: Want some love? Fine, come closer!)

فیگارو: آخ! چه مشت محکمی!
(Figaro: Ouch! What a strong punch!)

سوزانا: این یکی برای اون شب لعنتی... و این یکی برای دروغ‌هات! و این... اینم یکی دیگه!
(Suzana: That one's for that damn night... and this one for your lies! And this... one more!)

فیگارو: وایسا! حداقل فرصت نفس کشیدن بده!
(Figaro: Wait! At least let me catch my breath!)

سوزانا: بگیر! اینم! تا یاد بگیری چطور سر زن رو گول نزنی!
(Suzana: Take this too! So you learn not to fool a woman!)

B Prompt Used for WIKI-FADIAL Generation

System prompt: This GPT model takes Farsi texts and transforms each one into a casual, informal dialogue between two people. The generated dialogue should:

- be natural, relatable, and culturally appropriate in Farsi,
- include elements such as humor, honest reflection, light-hearted frustration, empathetic advice, and small talk,
- reflect personal or work-related topics and daily routines when appropriate,
- use natural idioms or slang and maintain an informal tone,
- contain exactly 6 turns, each between 1 and 20 words,
- include a `"reference"` field in each turn quoting or summarizing the relevant part of the source text for transparency.

Input format (JSON):

```
{
  "texts": [
    "First Farsi text...",
    "Second Farsi text...",
    "Third Farsi text...",
    "Fourth Farsi text...",
    "Fifth Farsi text..."
  ]
}
```

Output format: The model produces five independent JSON objects (one per input text). Each object contains only the generated dialogue and does not include the original input text:

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": "...",
      "reference": "..."
    },
    {
      "speaker": "B",
      "text": "...",
      "reference": "..."
    }
  ]
}
```

C Human Post-editing Protocol for WIKI-FADIAL

Below is the streamlined protocol followed by native-speaker annotators when refining WIKI-FADIAL dialogues. Two annotators work independently; a senior linguist resolves any conflicts.

> **Guideline:** When editing or regenerating these dialogues, make sure each has exactly 6 short turns (each 1–20 words) in casual, friendly, natural Farsi, sounding like two people in a real-life chat. The language must be informal, sprinkled with humor, light frustration, empathy, or small talk, using idioms and spoken expressions that feel truly Iranian.
>
> Each turn must include a reference that ties it clearly to the original text (a short quote or idea summary) to show transparency. Ensure the conversation is not a dry summary but a playful, human back-and-forth — like two friends reacting, joking, giving quick advice, or sharing small relatable complaints.
>
> After generating, carefully check the dialogue for all these points: if it fails (too stiff, formal, robotic, or missing references), regenerate it with a prompt asking for playful, informal style, or manually fix grammar, shorten long lines, replace unnatural words, or add emotional touches until it feels authentic.
>
> Throughout, keep the JSON clean (no trailing commas, correct brackets), so the final output is short, vivid, idiomatic, well-referenced, and technically valid.

Stage 1 – Coherence & Flow

- Ensure each question follows logically from the previous answer.
- Delete turns that introduce off-topic content or repeat earlier facts.

Stage 2 – Factuality & Reference

- Check every statement against the source paragraph; replace or remove hallucinations.
- Verify names, dates, and numbers; correct diacritics.

Stage 3 – Style & Register

- Keep each turn 1–20 tokens long; split or condense otherwise.
- Use informal Farsi: swap formal forms for idioms/slang; add discourse markers sparingly.
- Optional: insert light humour, empathy, or advice—without altering facts.

Stage 4 – Compliance & Safety

- check Hate checks; anonymise or delete flagged text.
- Remove any content violating Wikimedia policy (explicit, medical, self-harm).

Reproducibility Details

Hardware. Experiments are conducted on **3× NVIDIA RTX A6000** GPUs (48 GB each) with mixed-precision (bf16), FlashAttention, and gradient checkpointing.

Software environment. Python 3.11, PyTorch 2.3.0, HuggingFace Transformers $\geq$ 0.22, and BitsAndBytes 0.45 for 4-bit QLoRA. All package versions are pinned in `environment.yml`.

Random seeds. We fix `seed = 100` for model initialisation, dataloader shuffling, and NumPy/PyTorch RNGs. Results are averaged over three seeds for classification.

Hyper-parameters.

- *LLM fine-tuning (LoRA):* per-device batch size = 4, gradient accumulation = 2 (effective batch size = 24 across 3 GPUs), lr = 1×10^{-4} (AdamW, cosine decay), weight decay = 1×10^{-4}, warmup ratio = 0, max grad-norm = 1.0, epochs = 3, context = 512 tokens. LoRA configuration: rank = 128, α = 256, dropout = 0.1.
- *Encoder classification (Frozen encoder + MLP head):* batch = 64, lr = 1×10^{-5} (AdamW), epochs = 100, context = 256 tokens. 4-layer head: 512, 256, 128, 64 $\rightarrow$ `LogSoftmax`, NLLLoss.
- *Encoder classification (Full fine-tuning):* batch = 32, lr = 2×10^{-5} (AdamW), epochs = 3, context = 256 tokens.

Code & data. All corpora, scripts, and checkpoints will be released.

Compute footprint. LLM fine-tuning consumes 16 GPU-hours in total; encoder runs finish in $\sim$2 hours per task. This modest budget enables replication on most academic clusters.

D Illustrative Examples of Human Post-editing for WIKI-FADIAL

Below are five examples of human post-editing for WIKI-FADIAL. Each example includes the original Wikipedia passage, the dialogue generated by the model, and the final version after manual revision by human annotators.The following examples illustrate different types of interventions:

- **Example 1:** The opening sentence contained a grammatical error that obscured the intended meaning, while the remainder of the dialogue was fluent and well-structured. As a result, only the first sentence was manually revised.
- **Example 2:** The model overused repetitive openers such as “Have you heard...” or “Have you read...”. To increase conversational variety and avoid redundancy, annotators modified these introductions manually.
- **Example 3:** The generated sentences lacked syntactic coherence and logical progression, resulting in a disjointed and confusing dialogue. Consequently, annotators chose to regenerate the entire dialogue using the model.
- **Example 4:** The model produced an inaccurate and unrealistic analogy—e.g., comparing a storm to a funnel—that was not grounded in the original source content. This type of error, commonly seen in metaphorical or comparative contexts, was corrected by regenerating the dialogue, which yielded a coherent and accurate result on the second attempt.
- **Example 5:** No revisions were required. The model-generated dialogue was grammatically sound, culturally appropriate, and naturally reflected Farsi conversational style, even capturing a light sense of humor typical of everyday speech.

Example 1:

فَرگشت به‌معنی تغییر در خصوصیت‌های وراثتی جانداران در طول نسل‌های متوالی است. فرگشت بدان معناست که تمام جانداران با همهٔ تنوعی که دارند، از نیایی مشترک که حدود ۳/۵ تا ۳/۸ میلیارد سال پیش می‌زیسته، پدیدار گشته‌اند.
مطالعهٔ علمی فرگشت از نیمهٔ سدهٔ نوزدهم میلادی آغاز شد، زمانی که پژوهش‌ها بروی سنگواره‌ها و گوناگونی جانداران، بسیاری از دانشمندان را متقاعد کرده بود که همهٔ جانداران باید به‌نحوی فرگشت یافته باشند. سازوکارهای پیش‌برندهٔ فرگشت همچنان نامشخص باقی‌ماندند تا اینکه در اواسط قرن نوزدهم میلادی، نظریه علمی «فرگشت براساس انتخاب طبیعی» توسط چارلز داروین و آلفرد راسل والاس مطرح گردید که به توضیح چگونگی و علل فرگشت می‌پرداخت. این نظریهٔ علمی به‌طور خلاصه بیان می‌کرد که تنها نسلی از جانداران توانایی زنده‌ماندن و تولیدمثل را دارند که به‌دلیل جهش‌های ژنتیکی، هماهنگی بیشتری با تغییرات محیطزندگی خود کسب کرده باشند.
فرگشت علت پدیدهٔ گونه‌زایی است؛ که طی آن یک گونهٔ اجدادی منفرد به دو یا چند گونهٔ متفاوت تقسیم می‌شود. گونه‌زایی در شباهت‌های ساختاری، جنینی و ژنتیکی جانداران؛ پراکندگی جغرافیایی گونه‌های مرتبط با هم، و ثبت سنگواره‌ای تغییرات، قابل مشاهده‌است. نیای مشترک جانداران امروزی تا بیش از ۳/۵ میلیارد سال پیش؛ یعنی از زمان پیدایش حیات روی زمین، قدمت دارد. فرگشت چه به‌صورت درون‌جمعیتی و چه به‌صورت گونه‌زایی میان جمعیت‌ها، از طرق گوناگونی روی می‌دهد؛ آهسته و پیوسته به نام انتخاب انباشتی، یا به‌سرعت از یک موضع ایستا تا موضع بعدی؛ که تعادل نقطه‌ای خوانده می‌شود.
در اوایل قرن بیستم، «نظریهٔ فرگشت براساس انتخاب طبیعی»، با علم ژنتیک، دیرین‌شناسی و سامانه‌شناسی ادغام شدند که با پیوستن یافته‌های بعدی چون زیست‌شناسی مولکولی؛ تحت عنوان سنتز مدرن به اوج رسید. این تلفیق به یک بنیان اصلی در زیست‌شناسی بدل شد؛ چنان‌که تبیینی منسجم و یکپارچه، برای تنوع زیستی و تاریخ حیات روی زمین، فراهم ساخت. امروزه فرگشت در شاخه‌های مختلف علوم زیستی، همچون زیست‌شناسی بقا، جنین‌شناسی، بوم‌شناسی، فیزیولوژی، دیرین‌شناسی، کشاورزی و پزشکی مورد مطالعه قرار می‌گیرد. همچنین فرگشت بر دیگر حیطه‌های مطالعات بشری، همچون انسان‌شناسی، فلسفه و روان‌شناسی نیز اثرگذار بوده‌است. فرگشت؛ سرمشق و الگوی مسلط و چهارچوب اصلی زیست‌شناسی نوین است و پایه و اساس آن را تشکیل می‌دهد و با شواهد بسیاری پشتیبانی می‌گردد.

Evolution refers to the change in the hereditary characteristics of organisms over successive generations. It implies that all living beings, with all their diversity, have originated from a common ancestor that lived approximately 3.5 to 3.8 billion years ago.The scientific study of evolution began in the mid-19th century, when research on fossils and the diversity of organisms convinced many scientists that all living things must have evolved in some way. The mechanisms driving evolution remained unclear until Charles Darwin and Alfred Russel Wallace proposed the scientific theory of evolution by natural selection, which explained how and why evolutionary change occurs. In brief, this theory stated that only those individuals capable of surviving and reproducing—due to genetic mutations that gave them better adaptation to environmental changes—would pass on their traits to the next generation.Evolution is the cause of speciation—the process by which a single ancestral species splits into two or more distinct species. Speciation is evident in structural, embryological, and genetic similarities among organisms; in the geographical distribution of related species; and in the fossil record showing gradual changes. The most recent common ancestor of all living organisms is estimated to date back more than 3.5 billion years, to the origin of life on Earth.Evolution can occur both within populations and through speciation between populations, in different ways: gradually and continuously through cumulative selection**,** or in rapid shifts between relatively stable states, known as punctuated equilibrium**.**In the early 20th century, the theory of evolution by natural selection was integrated with genetics, paleontology, and systematics—an integration later expanded with findings from molecular biology—forming what is known as the modern synthesis**.** This synthesis became a foundational framework in biology, offering a unified explanation for the diversity of life and the history of organisms on Earth.Today, evolution is studied across many fields of biological science, including evolutionary biology, embryology, ecology, physiology, paleontology, agriculture, and medicine. It has also influenced disciplines such as anthropology, philosophy, and psychology. Evolution is the dominant paradigm and central framework of modern biology**,** supported by extensive empirical evidence.

Example 1:

Generated dialogue befor editing	**Generated dialogue after editing**
A: داشتی می‌دونی فرگشت یعنی چی؟ همون تغییر ژنی تو نسلاس. A: Were you knowing what evolution means? It's that gene change in the generashons. B: آره، یعنی همه‌مون از یه جد مشترک اومدیم، حتی گیاها! B: Yeah, like all of us came from one common ancestor, even the plants! A: نظریه‌اشم داروین و والاس مطرح کردن با انتخاب طبیعی. A: Darwin and Wallace brought up its theory too, with natural selection. B: یعنی اونایی زنده می‌مونن که با محیط بهتر جورن دیگه؟ B: So the ones stay alive that match better with environment, right? A: واسه همینم گونه‌های جدید کم‌کم به‌وجود میان. A: That's why new species gradually come into being. B: امروزه همه چی با فرگشت توضیح داده میشه، از ژنتیک تا پزشکی. B: Nowadays everything gets explained with evolution, from genetics to medicine.	A: فرگشت همون تغییر ژنی تو نسلاس. A: Evolution is like gene change in generashons. B: یعنی همه‌مون از یه جد مشترک اومدیم، حتی گیاها! B: So we all came from a common ancestor — even plants! A: نظریه‌اشم داروین و والاس مطرح کردن بر اساس انتخاب طبیعی. A: Darwin and Wallace also proposed the theory, based on natural selection. B: یعنی اونایی زنده می‌مونن که با محیط بهتر جورن دیگه؟ B: You mean the ones that match better with the environment survive, right? A: واسه همینم گونه‌های جدید کم‌کم به‌وجود میان. A: That's why new species appear gradually. B: امروزه همه چی با فرگشت توضیح داده میشه، از ژنتیک تا پزشکی. B: These days everything is explained by evolution — from genetics to medicine.

Example 2:

چهارمضراب یک فرم از موسیقی سازی (بدون آواز) و متریک است که در موسیقی سنتی ایرانی رواج دارد. چهارمضراب معمولاً جزو تندترین بخش‌های موسیقی سنتی ایران دانسته می‌شود و عموماً میزان‌بندی 68 یا 616 دارد.
اجرای چهارمضراب در ابتدا روی سازهای تار و سه‌تار مرسوم بود و احتمال دارد که نام آن به نواختن هر چهار سیم این سازها توسط نوازنده اشاره کند یا به این اشاره کند که در اجرای این گونه موسیقی، زخمه‌هایی که با مضراب به سیم‌ها زده می‌شد، الگویی چهارتایی داشته‌اند. با این حال، چند وجه تسمیه احتمالی دیگر نیز برای چهارمضراب مطرح شده‌است. در دوره‌های معاصرتر، چهارمضراب برای سازهای دیگر از جمله ویولن و کمانچه نیز نوشته و اجرا شده‌است و اجرای آن با همراهی یک ساز ضربی نظیر تنبک یا با کمک دو ساز مختلف نیز متداول گشته‌است. در توصیف نحوهٔ اجرای چهارمضراب از اصطلاح‌های «پایه» و «واخوان» استفاده می‌شود که پایه به الگوی ساده‌ای که در زمینهٔ چهارمضراب مرتب تکرار می‌شود، اشاره می‌کند و واخوان به نت‌های تک تکی که بین الگوی پایه اجرا می‌شوند و ملودی چهارمضراب را شکل می‌دهند.
موسیقی‌شناسان در مورد کاربرد چهارمضراب اتفاق نظر ندارند؛ برخی آن را وسیلهٔ تمرین نوازنده می‌دانسته و از این جهت آن را با اتود در موسیقی کلاسیک غربی مقایسه کرده‌اند، دیگران آن را حاصل خلاقیت آهنگسازان برای کاهش یکنواختی اجرای موسیقی ردیفی دانسته‌اند و از دید عده‌ای هدف از اجرای چهارمضراب نمایش دادن مهارت نوازندهٔ ساز و استراحت دادن به خواننده بوده‌است.
در بین فرم‌های موسیقی سنتی ایرانی، چهارمضراب قدمت کمتری دارد. شکل‌گیری چهارمضراب‌های جدید را به درویش‌خان نسبت داده‌اند و از حبیب سماعی، ابوالحسن صبا، مرتضی محجوبی، جواد معروفی، علی تجویدی و فرامرز پایور به عنوان سازندگان چهارمضراب‌های معروف نام برده می‌شود. از پایور آلبومی به نام چهارمضراب نیز برجای مانده‌است.

Chaharmezrab is a form of instrumental (non-vocal) and metric music that is common in traditional Iranian music. It is usually considered one of the fastest sections of Iranian traditional music and is generally written in 6/8 or 6/16 time signature.Chaharmezrab was originally performed on instruments like the *tar* and *setar*, and its name may refer either to the act of plucking all four strings of these instruments by the musician or to a four-part picking pattern used in this style of music. However, several other possible etymologies have also been suggested.In more recent times, *chaharmezrab* compositions have also been written and performed on other instruments such as violin and *kamancheh*, and performances accompanied by a percussion instrument like *tonbak* or even with two different instruments have become common.In describing the performance of *chaharmezrab*, the terms *paye* ("base") and *vākhan* ("response") are used. *Paye* refers to a simple recurring pattern that repeats in the background, while *vākhan* denotes the individual notes played between the base patterns, which form the melody of the *chaharmezrab*.Musicologists do not fully agree on the purpose of *chaharmezrab*. Some consider it a technical exercise for the performer, comparing it to an *étude* in Western classical music. Others view it as a creative element added by composers to reduce the monotony of radif (canonical repertoire) performance. Still others believe its aim is to showcase the instrumentalist's skill and give the vocalist a chance to rest.Among the forms of traditional Iranian music, *chaharmezrab* has a relatively shorter history. The development of modern *chaharmezrabs* is attributed to *Darvish Khan*, and well-known composers of the form include *Habib Samaei*, *Abolhasan Saba*, *Morteza Mahjoubi*, *Javad Maroufi*, *Ali Tajvidi*, and *Faramarz Payvar*. An entire album titled *Chaharmezrab* has also been recorded by *Payvar*.

Example 2:

Generated dialogue befor editing	**Generated dialogue after editing**
A: شنیدی می‌گن چهارمضراب تندترین بخش موسیقی سنتیه؟ A: Have you heard they say Chaharmezrab is the fastest part of traditional music? B: آره، مخصوصاً وقتی با تنبک اجرا میشه، جون‌دار میشه حسابی! B: Yeah, especially when it’s played with tonbak, it really becomes full of life! A: اسمش از اون چهارتا زخمه با مضرابه نه؟ A: Its name comes from those four strokes with the pick, right? B: بعضیا هم می‌گن تمرینیه، مثل اتود تو موسیقی کلاسیک. B: Some people also say it's like a practice piece, like an étude in classical music. A: ولی پایور و صبا باهاش کلی کار هنری کردن. A: But Payvar and Saba did a lot of artistic stuff with it. B: آلبوم «چهارمضراب» پایور هم که دیگه شاهکاره! B: Payvar’s "Chaharmezrab" album is just a masterpiece!	A: می‌گن چهارمضراب تندترین بخش موسیقی سنتیه. A: They say Chaharmezrab is the fastest part of traditional music. B: آره، مخصوصاً وقتی با تنبک همراهی می‌شه، واقعاً پرانرژی و زنده‌ست! B: Yes, especially when it’s accompanied by the tonbak, it becomes truly energetic and lively! A: اسمش هم از اون چهار زخمه با مضراب گرفته شده، درسته؟ A: Its name comes from those four strokes played with the pick, right? B: بعضی‌ها هم می‌گن نوعی تمرین برای نوازنده‌هاست، مثل اتود در موسیقی کلاسیک. B: Some also say it’s a kind of technical exercise for performers, like an étude in classical music. A: اما استادانی مثل پایور و صبا ازش آثار هنری جدی ساختن. A: But masters like Payvar and Saba created serious artistic works with it. B: آلبوم «چهارمضراب» اثر پایور هم که واقعاً شاهکاره! B: And Payvar’s album *Chaharmezrab* is truly a masterpiece!

Example 3:

مکبث (انگلیسیMacbeth, King of Scotland :؛ c. 1005 – ۱۵ اوت ۱۰۵۷) پادشاه اسکاتلند بود. او ابتدا سردار نبرد با نروژی‌ها و حاکم خطه گلامیس بود که با پیروزی در نبرد با نروژ از طرف شاه دانکن، شاه اسکاتلند، حاکم خطه کادور و سومین ولیعهد بعد از پسران شاه، ملکوم و دونالبین شد. مکبث به تحریک همسرش شاه دنکان را در خواب کشته ، خود بر تخت سلطنت اسکاتلند نشسته و اتگشت اتهام قتل را به سوی پسران شاه قرار داد. دونالبین به ایرلند رفته و ملکوم به لندن، پیش ادوارد خستو پادشاه انگلستان رفته و با کمک شاه انگلیس و سرداران شورشی مکبث از جمله مکدف، مکبث و سپاهش را در نبرد دنسی نین شکست داد و کشت و بر تخت سلطنت نشست.
Macbeth (English: *Macbeth, King of Scotland*; c. 1005 – 15 August 1057) was the king of Scotland. He was initially a general fighting against the Norwegians and the ruler of the region of Glamis. After a victory in the war against Norway, he was appointed by King Duncan of Scotland as the ruler of the region of Cawdor and became the third heir to the throne after the king's sons, Malcolm and Donalbain.Encouraged by his wife, Macbeth killed King Duncan in his sleep, seized the throne of Scotland, and directed the accusation of murder toward the king’s sons. Donalbain fled to Ireland and Malcolm went to London, seeking refuge with Edward the Confessor, the king of England. With the help of the English king and Macbeth's rebellious generals, including Macduff, Malcolm defeated Macbeth and his army in the Battle of Dunsinane, killed him, and ascended the throne.

Example 3:

Generated dialogue befor editing	**Generated dialogue after editing**
A: مکبث اولش کلی وفادار بود، سردار جنگ با نروژ! A: Macbeth was super loyal at first — a warlord fighting Norway! B: آره ولی زنش افتاد به جونش، شاه رو تو خواب زد! B: Yeah, but his wife pushed him — he killed the king in his sleep! A: پسرای شاه هم فرار کردن، بندازن تقصیر اونا! A: The king's sons ran away — so they could blame it on them! B: ولی ملکوم برگشت با لشکر انگلیس، زد لت‌وپارش کرد! B: But Malcolm came back with the English army and totally smashed him! A: اون مکدفم بود که انتقام گرفت، دل همه خنک شد! A: And it was Macduff who got revenge — everyone felt satisfied! B: نتیجه جاه‌طلبی بی‌حساب همینه دیگه، می‌افتی ته دره. B: That's what blind ambition gets you — you end up at the bottom.	A: بابا مکبثم آخرش رفت رو دور قتل و خیانت، نه؟! A: Man, even Macbeth ended up on a killing and betrayal spree, right?! B: آره، اون اول سردار بود، ولی با تحریک زنش پرید وسط کار خلاف! B: Yeah, he was a commander at first, but his wife pushed him into shady stuff! A: بعدشم نقشو انداخت گردن پسرای شاه که خودش پادشاه شه. A: Then he blamed it on the king's sons so he could take the throne himself. B: اما خب، ملکوم برگشت با لشکر انگلیسی‌ها حالشو گرفت. B: But yeah, Malcolm came back with the English army and wrecked him. A: مکدفم که انتقام خون شاه رو گرفت، مکبث دیگه راهی نداشت. A: And Macduff avenged the king's blood — Macbeth had no way out after that. B: دیدی آخر عاقبت جاه‌طلبی بی‌حساب چی میشه؟ فقط تهش حسرت و مرگ! B: See what blind ambition gets you in the end? Just regret and death.

Example 4:

پیچند، توفان پیچنده یا تورنادو(Tornado) ، نوعی از گردباد است که به توده هوای به‌شدت متلاطم گویند که با سطح زمین و یک ابر کومولونیمبوس تماس برقرار کرده، و در برخی موارد اساساً یک ابر کومولونیمبوس است. از آنان با عنوان چرخندها یا طوفان‌ها نام‌می‌برند، اگرچه کلمه چرخند در هواشناسی در یک مفهوم گسترده‌تر، به هر گردش بسته کم‌فشار گفته می‌شود. پیچندها در شکل‌ها و اندازه‌های مختلفی ایجاد می‌شوند ولی در کل شکل ابر قیفی را دارند، که طرف باریک آن با سطح زمین در تماس است و اغلب اوقات با ابری از آوار و گرد و غبار همراه هستند. بیشتر پیچندها سرعتی کمتر از ۱۱۰ مایل بر ساعت (۱۷۷ کیلومتر/ساعت) و ارتفاعی به اندازهٔ ۲۵۰ فوت (۷۶ متر) دارند و پیش از پراکنده شدن، مسافتی در حدود چندین کیلومتر را می‌پیمایند. قدرتمندترین پیچندها، پیچند سال ۱۹۹۹ پل کریک-مور، می‌توانند سرعتی بیش از ۳۰۰ مایل بر ساعت (۴۸۳ کیلومتر بر ساعت) و پهنایی بیش از دو مایل (۳/۲ کیلومتر) داشته باشند و می‌توانند ده‌ها مایل (بیش از ۱۰۰ کیلومتر) بر روی زمین باقی بمانند.

گونه‌های مختلف پیچندها شامل، فواره زمین، پیچند چندگرداب و پیچند دریایی می‌شود. تنوره‌های دریایی، به‌صورت بادهای مارپیچی قیفشکل شناخته، که به ابرهای عظیم کومه‌ای یا کومه‌ای بارا متصل می‌شوند. این‌ها به‌عنوان پیچندهای غیرمخرب دسته‌بندی می‌شوند که بر سطح آب ایجاد و در این‌که آیا این پیچندها از پیچندهای واقعی محسوب می‌گردند اختلاف نظر وجود دارد. این بادهای مارپیچی قیفی‌شکل، بیشتر در مناطق گرمسیری نزدیک به استوا رخ می‌دهند و در سرزمین‌های قطبی کمتر شایع هستند. دیگر پدیده‌های شبیه به پیچند در طبیعت، تندباد کوتاه‌مدت، تنوره دیو، گردباد آتشین و تنوره بخار هستند؛ بیشتر بادهای رعدوبرق‌زا با پیچندها اشتباه گرفته می‌شوند درحالی‌که عملکردشان کاملاً متفاوت است.

پیچندها در تمامی قاره‌ها به‌جز جنوبگان دیده شده‌اند. درحالی‌که اکثریت قریب به‌اتفاق پیچندها درمنطقه پیچند، منطقه‌ای در ایالات متحده آمریکا، رخ می‌دهند، ولی امکان رخداد آن در همه جای آمریکای شمالی وجود دارد. همچنین در مناطقی از جنوب آسیا و آسیای شرقی، شمال و شرق و مرکز آمریکای جنوبی، جنوب آفریقا، شمال‌غربی و جنوب شرقی اروپا، غرب و جنوب شرقی استرالیا و نیوزیلند دیده می‌شوند. پیچندها را می‌توان از طریق رادار دوپلر پالس با استفاده از الگوهای سرعت و بازتاب پیش‌بینی کرد، همچنین امکان استفاده از اکو هوک یا حوزه‌های تخریب یا استفاده از لکه‌های طوفان وجود دارد.

چندین معیار برای اندازه‌گیری قدرت و میزان تخریب پیچندها وجود دارد. مقیاس فوجیتا پیچندها را بر اساس قدرت تخریب و خسارتی که برجای گذاشته‌اند، می‌سنجد و امروزه در بسیاری از کشورها نسخهٔ به‌روز شدهٔ آن مقیاس فوجیتای افزوده جایگزین شده‌است. یک پیچند F0 یاEF0 ، ضعیف‌ترین نوع، می‌تواند به درختان خسارت زده اما به تأسیسات اساسی صدمه‌ای نمی‌زند. یک پیچند F5 یاEF5 ، قوی‌ترین نوع، به پایه‌های ساختمان‌ها آسیب می‌زند و می‌تواند آسمان‌خراش‌ها را دچار حادثه کند. مقیاس TORPO نیز از T0 برای پیچندهای بسیار ضعیف تا T11 برای قدرتمندترین پیچندها دسته‌بندی می‌شود. اطلاعات رادار هواشناسی داپلر، تصویرسنجی، و الگوهای چرخش زمین (علائم چرخزاد) نیز برای تعیین شدت و دسته‌بندی پیچندها به کار می‌روند.

A *tornado* (also known as a whirlwind or twister) is a type of violently rotating air mass that makes contact with the ground and a cumulonimbus cloud — and in some cases, it essentially *is* a cumulonimbus cloud. These phenomena are referred to as cyclones or storms, although in meteorology the word *cyclone* has a broader meaning, referring to any low-pressure closed circulation.Tornadoes come in various shapes and sizes but generally take the form of a funnel cloud, with the narrow end in contact with the ground, often accompanied by a cloud of debris and dust. Most tornadoes have wind speeds below 110 miles per hour (177 km/h), diameters of about 250 feet (76 meters), and travel a few kilometers before dissipating. The most powerful tornadoes — such as the 1999 Bridge Creek–Moore tornado — can reach speeds over 300 miles per hour (483 km/h), have widths over two miles (3.2 km), and remain on the ground for more than 100 kilometers.Different types of tornadoes include *landspouts*, *multiple vortex tornadoes*, and *waterspouts*. Waterspouts are spiraling funnel-shaped winds that connect to large cumulus or cumulonimbus clouds. They are generally classified as non-destructive tornadoes that form over water, and there is debate over whether they should be considered “true” tornadoes. These spiraling funnel-shaped winds occur more frequently in tropical areas near the equator and are rare in polar regions.Other natural phenomena similar to tornadoes include *gustnadoes*, *dust devils*, *fire whirls*, and *steam devils*. Many thunderstorm-related winds are often mistaken for tornadoes, though their mechanisms are entirely different.Tornadoes have been observed on every continent except Antarctica. While the vast majority occur in “Tornado Alley” — a region in the United States — they can happen anywhere in North America. Tornadoes are also found in parts of South and East Asia, northern, eastern, and central South America, southern Africa, northwestern and southeastern Europe, western and southeastern Australia, and New Zealand.Tornadoes can be predicted using Doppler radar by analyzing velocity and reflectivity patterns. Other indicators include hook echoes, damage paths, or storm spotter reports.Several systems exist to measure the strength and destructiveness of tornadoes. The Fujita scale rates tornadoes based on the damage they cause; today, many countries use an updated version known as the Enhanced Fujita (EF) scale. An F0 or EF0 tornado — the weakest — may damage trees but not significantly affect major infrastructure. An F5 or EF5 tornado — the strongest — can destroy building foundations and damage skyscrapers. The TORRO scale also classifies tornadoes from T0 (very weak) to T11 (extremely powerful). Doppler weather radar data, imaging tools, and ground rotation patterns (tornadic signatures) are also used to assess and categorize tornado intensity.

Example 4:

Generated dialogue befor editing	Generated dialogue after editing
A: داداش اون روز مستند پیچند دیدم، یه قیف افتاده بود وسط زمین! A: Dude, I watched a tornado documentary the other day — a funnel had dropped right onto the ground! B: قیف؟ یعنی هوا داشت آش درست می‌کرد؟ B: A funnel? So the air was cooking soup or what? A: نه بابا! سرعتش بعضی وقتا تا ۴۸۰ کیلومتر می‌رسه! A: Nah, man! Sometimes it goes up to 480 kilometers per hour! B: یعنی اگه بزنه به خونه ما، کل محله می‌ره سفر! B: If that hits our house, the whole neighborhood's going on a trip! A: یه مدلشم هست بهش می‌گن تنوره دریایی، رو دریاست! A: There's also a kind called a waterspout — it happens over the sea! B: دیگه کوسه هم بیاد وسطش بشه فیلم تخیلی کامل میشه! B: Just add a shark in the middle of it and boom — you've got a full sci-fi movie!	A: یه طوفان قیفی رو تو اخبار نشون می‌داد که نصف شهر رفت تو هوا! A: They were showing a funnel storm on the news — half the city flew into the air! B: یه پیچند اومد. با خودش گرد و غبار آورد! B: A tornado came. It brought dust and debris with it! A: می‌گن بعضیاش سرعتش از ۴۰۰ کیلومترم بیشتره! A: They say some of them go over 400 kilometers an hour! B: با اون سرعت می‌تونه از خونه من بپره تو خونه خاله‌م! B: At that speed it could jump from my house into my aunt's! A: یه مدل دیگه‌شم هست بهش می‌گن تنوره دریایی، فقط رو آبه! A: There's another type too — they call it a waterspout, it only happens over water! B: خب پس با قایق نریم شمال، یه وقت گیر نیفتیم توش! B: Then let's not go boating up north — what if we get stuck in one?!

Example 5:

جستجوگر گوگل (به انگلیسی Google Search):یا گوگل یک موتور جستجوی وب است که توسط گوگل توسعه داده شده‌است. گوگل از راه این وبگاه روزانه صدها میلیون دلار درآمد به‌دست می‌آورد .Marketplace. American Public Media. Retrieved December 9, 2017")(به انگلیسی). این وبگاه در سال ۱۹۹۷ بنیاد نهاده شد. دامنهٔ اصلی این سایت در ماه مهٔ ۲۰۰۸ ۱۳۵ میلیون بار بازدید شده‌است. این موتور جستجو بیشترین بازدیدکننده در بین کاربران را دارد. موتور گوگل روزانه چندصدمیلیون بار به طرق مختلف استفاده می‌شود. مهم‌ترین هدف گوگل، یافتن متن موردنظر در میان صفحات وب است. اما انواع دیگر اطلاعات به وسیلهٔ قسمت‌های دیگر آن مثل جستجوی تصاویر گوگل، نیز مورد جستجو قرار می‌گیرند. جستجوگر گوگل توسط لَری پیج و سرگئی برین در سال ۱۹۹۷ ساخته شد. این جستجوگر به‌جز جستجوی واژگان، ۲۲ حالت جستجوی دیگر نیز دارد؛ مانند جستجوی مترادف‌ها، پیش‌بینی هوا، محدوده‌های زمانی (وقت محلی)، قیمت سهام، اطلاعات زلزله، زمان نمایش فیلم‌ها، اطلاعات فرودگاه و... همچنین مختص اعداد، امکانات ویژه‌ای وجود دارد؛ مانند بازه (۷۰...۸۰)، دماها، واحدهای پول و تبدیل اینها به هم، عبارات محاسباتی و... ترتیب قرارگرفتن نتایج جستجوی گوگل بستگی به عاملی به نام رتبهٔ (به انگلیسی rank): صفحه دارد. جستجوی گوگل با به‌کاربردن عملگرهای جبر بولی مانند شمول و عدم شمول، گزینه‌های زیادی را برای کاربران قابل تنظیم کرده‌است. به انگلیسی(Advanced search)
Google Search (in English: *Google Search*), or simply *Google*, is a web search engine developed by Google. This website generates hundreds of millions of dollars in daily revenue for Google. ("Marketplace. American Public Media. Retrieved December 9, 2017"). The website was founded in 1997. Its main domain received 135 million visits in May 2008. Google's search engine has the highest number of users among all search engines. It is used hundreds of millions of times daily in various ways.The main goal of Google is to find the desired text among web pages. However, other types of information can also be searched using its additional services, such as *Google Image Search*.Google Search was created in 1997 by *Larry Page* and *Sergey Brin*.In addition to keyword search, this search engine supports 22 other types of searches, including searching for synonyms, weather forecasts, time zones (local time), stock prices, earthquake information, movie showtimes, airport data, and more. It also provides special features for numbers, such as ranges (e.g., 70...80), temperatures, currencies and their conversions, mathematical expressions, and so on.The order of Google search results depends on a factor called *page rank*. Google Search, using Boolean logic operators such as inclusion and exclusion, offers many customizable options for users (in English: *Advanced search*).

Example 5:

Generated dialogue

A:
والا من هر چی می‌خوام بدونم، اول از گوگل می‌پرسم!
A : Honestly, whatever I want to know, I ask Google first!

B:
منم حتی اگه یه دونه جوش بزنم، اول می‌زنم: جوش عجیب روی صورت 😅
B : Same here! Even if I get a single pimple, I type: "weird pimple on face" 😅

A:
یادته یه زمانی یاهو سرچ داشتیم؟ الان گوگل خدای جستجوعه!
A : Remember when we used Yahoo Search? Now Google is the god of searching!

B:
راستی اون قسمت حساب کردن و هواشناسی‌شم خیلی باحاله.
B : By the way, the calculator and weather forecast parts are really cool too.

A:
من از جستجوی عکساش بیشتر استفاده می‌کنم، بخصوص برای استیکر!
A : I mostly use its image search, especially for stickers!

B:
فقط یه چیزی، گوگل از ما بیشتر از مامانمون می‌دونه!
B : One thing though — Google knows more about us than our moms do!

E Model Checkpoints

Large-language models.

1. `meta-llama/Llama-3.2-1B-Instruct`
2. `meta-llama/Llama-3.2-3B-Instruct`
3. `meta-llama/Llama-3.1-8B-Instruct`
4. `mistralai/Mistral-7B-Instruct-v0.3`
5. `mistralai/Mistral-Nemo-Instruct-2407`
6. `mistralai/Mistral-Small-24B-Instruct-2501`

Transformer encoders.

1. `intfloat/multilingual-e5-large`
2. `intfloat/multilingual-e5-base`
3. `intfloat/multilingual-e5-small`
4. `HooshvareLab/bert-base-parsbert-uncased`
5. `sbunlp/fabert`
6. `google-bert/bert-base-multilingual-cased`
7. `FacebookAI/roberta-base`
8. `FacebookAI/roberta-large`
9. `FacebookAI/xlm-roberta-large`
10. `FacebookAI/xlm-roberta-base`

Translation Prompt

The following prompt was used to generate culturally authentic Farsi dialogues from English sources:

Your task is to translate a given dialogue from English to Farsi, ensuring the following guidelines are followed:

1. Farsi Names: Use Farsi names in the translated dialogue while maintaining the gender distinctions present in the English version.

2. Natural Language: Craft the conversation in everyday spoken Farsi, utilizing common expressions and adhering to cultural norms of politeness.

3. Engaging Flow: Ensure the dialogue consists of at least six exchanges between two characters. Maintain a natural, logical, and engaging flow. Where relevant, include elements such as questions, disagreements, or problem-solving.

4. Cultural Authenticity: Reflect authentic Farsi cultural nuances in the conversation.

5. Dialogue Length: Keep the length of the dialogue the same as in the English version. Do not add any new conversations or exchanges.

6. Output Format: Provide the complete conversation in a valid JSON format. The dialogue should be entirely in Farsi, with no additional English commentary or context. Maintain logical coherence and consistency throughout.

Expected JSON Output Format:

```
{
  "conversation": [
    {
      "speaker": "first person",
      "text": "{{response_1}}"
    },
    ...
  ]
}
```

English dialogue: dialog

F Human Post-editing Protocol for DAILYDIALOG-FA

Below is the streamlined protocol followed by native-speaker annotators when refining *DailyDialog-FA* dialogues. Two annotators work independently to post-edit the model-generated Farsi dialogues to ensure naturalness, cultural relevance, and structural fidelity to the original English source.

Editing Procedure

Once the model generates the Farsi dialogue from English input, human annotators follow this step-by-step process:

1. **Fluency and Naturalness**
   Ensure that each utterance reads like natural, everyday spoken Farsi. Sentences should flow conversationally, avoiding overly literal translations or stiff, unnatural phrasing. Aim for speech that reflects how native speakers actually talk.

2. **Cultural Localization**
   Adapt all cultural elements to fit an Iranian context. This includes:
   - Replacing foreign currencies (e.g., USD, Euro) with the appropriate Iranian currency.
   - Substituting references to globally known locations (e.g., Niagara Falls) with locally recognizable or culturally familiar places.
   - Adjusting units of measurement or public references (e.g., holidays, social norms) to culturally meaningful equivalents.

3. **Preserving Dialogue Structure**
   Keep the number of dialogue turns exactly the same as the original English version. Do not split, merge, or remove turns unless absolutely necessary for clarity or fluency. Preserve the logical progression and intention behind each exchange.

4. **Speaker Identity and Naming**
   Assign culturally appropriate Farsi names that match the gender and style of the original English characters. Avoid names that may seem foreign or uncommon in the Iranian context. Gender agreement between speakers' names and pronouns must be consistent throughout.

5. **Grammatical and Semantic Corrections**
   Correct any errors in grammar, word choice, or sentence structure. Fix mistranslations that may alter or obscure the intended meaning. Pay special attention to:
   - Verb tense consistency
   - Proper use of prepositions and conjunctions
   - Subject-verb agreement
   - Pronoun resolution

6. **Removal of Artifacts**
   Eliminate any residual English words, incomplete phrases, formatting markers, or annotation artifacts (e.g., brackets, placeholder tokens). Ensure that the dialogue reads cleanly, without machine-generated traces.

7. **Tone and Politeness**
   Maintain an appropriate register for informal dialogue. The tone should feel friendly, polite, and contextually appropriate for casual interaction. Avoid archaic, poetic, or overly formal language unless explicitly present in the original.

8. **Formatting and Output**
   Ensure the final dialogue is submitted in valid JSON format, with clearly structured key-value pairs for each turn. No syntax errors should be present, and spacing or punctuation should follow standard conventions.

G Illustrative Examples of Human Post-editing for DAILYDIALOG-FA

Below are five representative examples from DAILYDIALOG-FA. Each example includes the original English dialogue, the raw Farsi translation generated by the model, and the final version after human post-editing. These examples illustrate the types of edits performed and the reasoning behind them.

- **Example 1:** References to "dollars" were converted to "tomans" to reflect the local currency used in Iran.
- **Example 2:** A grading scale out of 100 was adapted to a 0–20 scale, which is standard in the Iranian education system.
- **Example 3:** Mention of China's Huangguoshu Waterfall was replaced with Iran's Laton Waterfall to create a culturally familiar context.
- **Example 4:** An American store name was substituted with a well-known Iranian brand.
- **Example 5:** The model autonomously replaced an American personal name with a natural-sounding Farsi name, requiring no human edits.

Each case illustrates the specific types of transformations necessary to produce high-quality, culturally grounded dialogues that are suitable for native Farsi speakers.

Example1:

Original Text:

['I need to copy a document immediately . '
' We have a copy machine in our computer lab , located on the first floor . '
' Great ! How much is it for a copy ? '
' The price per copy is ten cents . '
" That sounds reasonable . I'll be down there immediately . "
" There shouldn't be any long lines , sir . "]

Translation before editing:	**Translation after editing:**
رضا: من باید همین الان یک مدرک رو کپی بگیرم. Reza: I need to make a copy of a document right now.	رضا: من باید همین الان یک مدرک رو کپی بگیرم. Reza: I need to make a copy of a document right now.
سارا: ما توی آزمایشگاه کامپیوترمون یه دستگاه کپی داریم، توی طبقه اول هست. Sara: We have a copy machine in our computer lab—it's on the first floor.	سارا: ما توی آزمایشگاه کامپیوترمون یه دستگاه کپی داریم، توی طبقه اول هست. Sara: We have a copy machine in our computer lab—it's on the first floor.
رضا: عالیه! هزینه‌ی هر کپی چقدره؟ Reza: Great! How much does each copy cost?	رضا: عالیه! هزینه‌ی هر کپی چقدره؟ Reza: Great! How much does each copy cost?
سارا: قیمت هر کپی ده سنت هست. Sara: Each copy costs ten cents.	سارا: قیمت هر کپی ده تومان هست. Sara: Each copy costs ten tomans.
رضا: خیلی معقوله، همین الان می‌رم پایین. Reza: That’s very reasonable—I'll head down right now.	رضا: خیلی معقوله، همین الان می‌رم پایین. Reza: That’s very reasonable—I'll head down right now.
سارا: نباید صف طولانی باشه، آقا. Sara: There shouldn’t be a long line, sir.	سارا: نباید صف طولانی باشه، آقا. Sara: There shouldn’t be a long line, sir.

Example 2:

Original Text:

['What do you want to know about me ? '
' How about your academic records at college ? '
' The average grade of all my courses is above 85 . '
' In which subject did you get the highest marks ? '
' In mathematics I got a 98 . ' ' Have you received any scholarships ? '
' Yes , I have , and three times in total . '
' Have you been a class leader ? '
' I have been a class commissary in charge of studies for two years . '
' Did you join in any club activities ? '
' I was an aerobics team member in college . '
' What sport are you good at ? '
' I am good at sprint and table tennis . ' ' You are excellent . ']

Translation before editing:	**Translation after editing:**
مریم: دوست داری چی در مورد من بدونی؟ Maryam: What would you like to know about me?	مریم: دوست داری چی در مورد من بدونی؟ Maryam: What would you like to know about me?
علی: مثلاً درباره‌ی سوابق تحصیلی‌ات در دانشگاه؟ Ali: For example, your academic background in university?	علی: مثلاً درباره‌ی سوابق تحصیلی‌ات در دانشگاه؟ Ali: For example, about your academic background at university?
مریم: میانگین نمرات همه‌ی درسام بالای ۸۵ بوده. Maryam: My average grade in all subjects has been above 85.	مریم: میانگین نمرات همه‌ی درسام بالای ۱۹ بوده. Maryam: My average grade in all subjects has been above 19.
علی: تو کدوم درس بالاترین نمره رو گرفتی؟ Ali: In which subject did you get the highest score?	علی: تو کدوم درس بالاترین نمره رو گرفتی؟ Ali: In which subject did you get the highest grade?
مریم: در ریاضی نمره‌ی ۹۸ گرفتم. Maryam: I got a 98 in mathematics.	مریم: در ریاضی نمره‌ی ۱۹.۵ گرفتم. Maryam: I got a 19.5 in mathematics.
علی: تا حالا بورس تحصیلی گرفتی؟ Ali: Have you ever received a scholarship?	علی: تا حالا بورس تحصیلی گرفتی؟ Ali: Have you ever received a scholarship?
مریم: بله، سه بار بورس تحصیلی گرفتم. Maryam: Yes, I've received a scholarship three times.	مریم: بله، سه بار بورس تحصیلی گرفتم. Maryam: Yes, I've received a scholarship three times.

Example 3:

Original Text:
['Well , do you have any experience in guiding ? ' ' Yes , but I only have few experiences . I guided a few groups of foreign tourists from America last winter vacation . I showed them around some places of interest . ' ' What places of interest did you show them ? ' ' Too many . I will just give you one example , Huangguoshu Falls , the largest waler-fall in China . ']

Translation before editing:	**Translation after editing:**
سارا: خب، آیا تا حالا تجربه‌ای در راهنمایی گردشگران داشتی؟ Sara: So, have you ever had experience guiding tourists? مهدی: بله، اما خیلی زیاد نه. زمستان سال گذشته چند گروه از گردشگران خارجی از آمریکا را راهنمایی کردم و آن‌ها را به چند جای دیدنی بردم. Mehdi: Yes, but not a lot. Last winter, I guided several groups of foreign tourists from the U.S. and took them to some tourist attractions. سارا: چه جاهای دیدنی‌ای را به آن‌ها نشان دادی؟ Sara: What tourist sites did you show them? مهدی: خیلی جاها! ولی فقط یک نمونه می‌گویم، آبشار هووان گوشوکه بزرگ‌ترین آبشار در چین است. Mehdi: Many places! But I'll just mention one: Huangguoshu Waterfall, which is the largest waterfall in China.	سارا: خب، آیا تا حالا تجربه‌ای در راهنمایی گردشگران داشتی؟ Sara: So, have you ever had experience guiding tourists? مهدی: بله، اما خیلی زیاد نه. زمستان سال گذشته چند گروه از گردشگران خارجی از آمریکا را راهنمایی کردم و آن‌ها را به چند جای دیدنی بردم. Mehdi: Yes, but not a lot. Last winter, I guided a few groups of foreign tourists from the U.S. and took them to some attractions. سارا: چه جاهای دیدنی‌ای را به آن‌ها نشان دادی؟ Sara: What tourist spots did you show them? مهدی: خیلی جاها! ولی فقط یک نمونه می‌گویم، لاتون که بزرگ‌ترین آبشار در ایران است. Mehdi: Many places! But I'll just mention one: Laton, which is the largest waterfall in Iran.

Example 4:

Original Text:

["Excuse me . I'm looking for Bluemingdails . Could you tell me how to get there ? "
" Sure . It's very close actually.You go straight down this road . Then you turn left , at the next junction . "
' Left the next junction . Ok . '
" Bluemingdails is on the corner of that block . You see it as you turn left.Seriously , You can't miss it.It ' s enormous . "
' Thanks very much . ']

Translation before editing:

شخص اول: ببخشید، من دارم دنبال فروشگاه بلومینگدیلز می‌گردم. می‌تونید بهم بگید چطور می‌تونم برم اونجا؟
Person 1: Excuse me, I'm looking for the Blumingdale's store. Can you tell me how I can get there?

شخص دوم: البته! خیلی نزدیکه. این خیابون رو مستقیم برو، بعد سر چهارراه بعدی به چپ بپیچ.
Person 2: Of course! It's very close. Go straight on this street, then turn left at the next intersection.

شخص اول: چهارراه بعدی به چپ، درسته؟
Person 1: Left at the next intersection, right?

شخص دوم: بله، فروشگاه بلومینگدیلز سر نبش اون بلوک قرار داره. به محض این‌که بپیچی، می‌بینیش. واقعاً محاله از دستش بدی، خیلی بزرگه!
Person 2: Yes, Blumingdale's is right on the corner of that block. As soon as you turn, you'll see it. It's really impossible to miss—it's huge!

شخص اول: خیلی ممنون!
Person 1: Thank you so much!

Translation after editing:

شخص اول: ببخشید، من دارم دنبال فروشگاه پرسپولیس می‌گردم. می‌تونید بهم بگید چطور می‌تونم برم اونجا؟
Person 1: Excuse me, I'm looking for the Persepolis store. Can you tell me how I can get there?

شخص دوم: البته! خیلی نزدیکه. این خیابون رو مستقیم برو، بعد سر چهارراه بعدی به چپ بپیچ.
Person 2: Of course! It's very close. Go straight on this street, then turn left at the next intersection.

شخص اول: چهارراه بعدی به چپ، درسته؟
Person 1: Left at the next intersection, right?

شخص دوم: بله، فروشگاه پرسپولیس سر نبش اون بلوک قرار داره. به محض این‌که بپیچی، می‌بینیش. واقعاً محاله گمش کنی، خیلی بزرگه!
Person 2: Yes, the Persepolis store is right on the corner of that block. As soon as you turn, you'll see it. It's really impossible to miss—it's huge!

شخص اول: خیلی ممنون!
Person 1: Thank you so much!

Example 5:

Original Text:

['Marketing manage office , can I help you ? ' ' Mr Green , please . '
" Sorry , he's out . " ' What time you suspect back them ? '
' Sorry , I am not sure , can I take a message ? '
" No , thanks . I'll call back later . "
" That's fine.please call again later . " ' Ok , thank you.Goodbye . ']

Translation:

منشی **:**مدیریت بازاریابی، بفرمایید؟ چطور می‌توانم کمک کنم؟
Receptionist: Marketing department, how may I help you?

احمد **:**آقای کریمی لطفاً.
Ahmad: Mr. Karimi, please.

منشی **:**متأسفم، ایشان بیرون هستند.
Receptionist**:** I'm sorry, he's out.

احمد **:**حدس می‌زنید چه ساعتی برمی‌گردند؟
Ahmad: Do you have any idea what time he'll be back?

منشی **:**متأسفم، مطمئن نیستم. مایلید پیغامی بگذارید؟
Receptionist**:** I'm sorry, I'm not sure. Would you like to leave a message?

احمد **:**نه، ممنون. بعداً تماس می‌گیرم.
Ahmad**:** No, thanks. I'll call back later.

منشی **:**مشکلی نیست، لطفاً بعداً دوباره تماس بگیرید.
Receptionist: No problem, please try again later.

احمد **:**باشه، ممنون. خداحافظ.
Ahmad**:** Okay, thanks. Goodbye.

H Segmentation prompt

Task Description

Your duty is to receive the text and segment it strictly based on the rules below. The output must be in **Farsi** and presented in a **clear, organized table format**.

Segmentation Rules

1. **Scene Continuity:**
   - Strictly follow scene continuity.
   - Start a new segment only when there is an **explicit scene change**, such as:
     - Change in **location**
     - Change in **time**
     - A significant shift in the **conversation context** (not a minor one)
2. **Dialogue Preservation:**
   - Preserve the original dialogues **exactly as they appear in the text**.
   - Do **not** omit, alter, summarize, or paraphrase any part.
   - Keep interruptions, incomplete sentences, repetitions, and pauses.
   - Clearly indicate the **speaker** for each line.
   - Double-check for completeness to ensure no dialogue lines are missing after segmentation.
3. **Output Format:**
   - Present the output in a table with the following columns:
     (a) **Title:** A short, descriptive title summarizing the essence of the scene.
     (b) **Characters:** List all characters who have dialogues or are mentioned in the segment.
     (c) **Dialogues:** Present all dialogues exactly as in the original script, clearly marking each speaker.
     (d) **Overall Sentiment:** Choose from *Positive*, *Negative*, or *Neutral*.
     (e) **Reference:** Always write the name of the theatre play.

Additional Instructions

- Ensure each scene is **complete** before moving to the next one.
- Do **not** omit any part of the text.
- Double-check for any missing dialogues after segmentation.
- Maintain clarity and organization for easy analysis.

I Annotation and Editing Protocol for Play-Based Dialogues

To ensure consistency and quality during corpus construction, annotators followed a structured protocol while cleaning, extracting, reformulating, and validating dialogue segments from the Farsi play scripts.

- **Remove non-dialogue content.** Delete stage directions, narrator descriptions, scene headers, page numbers, and formatting artifacts introduced by OCR. Only spoken dialogue lines should remain.
- **Preserve original dialogue content.** The objective of the extraction stage is to retain the spoken dialogue exactly as it appears in the original plays. Annotators must not paraphrase, rewrite, or introduce new content at this stage.
- **Correct OCR artifacts.** Fix obvious OCR errors such as broken characters, incorrect punctuation, or merged tokens while preserving the intended wording of the original script.

- **Maintain speaker attribution.** Ensure that each dialogue line remains associated with the correct character in the original script. If OCR errors obscure the speaker name, the annotator should verify it by consulting the original PDF.
- **Preserve conversational order.** Dialogue lines must remain in the exact sequence in which they appear in the original play. Reordering dialogue turns is not permitted.
- **Remove stage interruptions within dialogue.** If stage directions appear inside dialogue lines (e.g., gestures, pauses, or descriptions of actions), remove them while keeping the surrounding spoken text intact.
- **Verify extraction against the original script.** After cleaning the OCR output, annotators compare the extracted dialogue with the original play to ensure that no dialogue lines were accidentally omitted or altered.
- **Check scene boundaries for segmentation.** When scripts are segmented into dialogue scenes, ensure that a new dialogue begins only when the scene clearly changes in the original play. Segmentation should preserve the natural conversational flow within each scene.
- **Verify reformulated dialogues for copyright compliance.** After segmentation, dialogue segments are reformulated using GPT-4o to avoid verbatim reproduction of copyrighted play scripts. Annotators compare each reformulated dialogue with the original scene to ensure that the conversational intent, speaker interactions, and thematic content are preserved while avoiding direct copying of the source text.
- **Preserve conversational structure.** During reformulation verification, annotators ensure that the number of turns and the overall conversational flow remain consistent with the original scene.
- **Prevent verbatim overlap.** Annotators ensure that reformulated dialogues do not contain long verbatim spans from the original plays, thereby preventing redistribution of copyrighted text.
- **Flag ambiguous or corrupted cases.** If OCR errors, missing dialogue, or unclear speaker attribution are encountered, annotators flag the issue for discussion and resolve it during the adjudication stage.
- **Validate sentiment labels.** For sentiment annotation, annotators review the dialogue within the context of the original scene to determine whether the overall sentiment is *positive*, *neutral*, or *negative*. When disagreements occur, annotators discuss the case and finalize the label through consensus.

J Example of Segmentation, Reformulation, and Human Verification

This appendix provides an illustrative example of the pipeline used to construct dialogues in PLAYDIALFA. For each scene, the original script excerpt is first segmented into a dialogue, then reformulated by GPT-4o to avoid verbatim copying, and finally verified and corrected by human annotators.

Example 1

Original Dialogue	After Reformulation
.جسیکا: این بهترین فیلمیه که تو عمرم دیدم .یونا: من خوشم نیومد .جسیکا: باید برین دنبالش یونا. باید پیداش کنین یونا: می‌دونی رفتن به نیویورک چقدر خرج داره؟ جسیکا: هیشکی نمی‌دونه. نرخ‌ها هر روز عوض می‌شن. تو چقدر پول داری؟ .یونا: هشتاد دلار جسیکا: من هم چهل و دو دلار دارم. فکر کنم پول تاکسی‌هات رو بتونی بدی. یونا: اما چه جوری تا اونجا برم؟ مادر جسیکا: عزیزم من چند دقیقه می‌رم بیرون... می‌شه چند دقیقه مراقب دفتر باشی تا من برگردم؟ .جسیکا: باشه جسیکا: می‌خوای صندلیت وسط باشه یا کنار پنجره؟ .یونا: کنار پنجره	!جسیکا: این فیلم شاهکاره .یونا: من خیلی باهاش حال نکردم .جسیکا: باید بری پیداش کنی یونا: می‌دونی بلیت نیویورک چقدر گرونه؟ جسیکا: کسی قیمت‌ها رو دقیق نمی‌دونه. هر روز فرق می‌کنه. پول چقدر داری؟ .یونا: هشتاد دلار جسیکا: منم چهل و دو تا دارم. برای کرایه تاکسی شاید بس باشه. یونا: ولی چطوری برم اونجا؟ مادر جسیکا: من چند دقیقه می‌رم بیرون، مواظب دفتر باشی؟ .جسیکا: حله جسیکا: می‌خوای کنار پنجره بشینی یا وسط؟ .یونا: پنجره رو ترجیح می‌دم
Jessica: This is the best movie I've ever seen in my life. Yuna: I didn't like it. Jessica: You have to go after him, Yuna. You have to find him. Yuna: Do you know how expensive it is to go to New York? Jessica: Nobody knows. Prices change every day. How much money do you have? Yuna: Eighty dollars. Jessica: I have forty-two dollars too. I think you can at least pay for your taxis. Yuna: But how do I get there? Jessica's mother: Honey, I'm going out for a few minutes... Can you watch the office for a few minutes until I come back? Jessica: Okay. Jessica: Do you want a middle seat or a window seat? Yuna: A window seat.	Jessica: This movie is a masterpiece! Yuna: I didn't really enjoy it that much. Jessica: You have to go find him. Yuna: Do you know how expensive a ticket to New York is? Jessica: No one knows the exact prices. They change every day. How much money do you have? Yuna: Eighty dollars. Jessica: I have forty-two as well. Maybe it's enough for taxi fares. Yuna: But how am I supposed to get there? Jessica's mother: I'm stepping out for a few minutes—can you watch the office? Jessica: Sure. Jessica: Do you want to sit by the window or in the middle? Yuna: I prefer the window.

Example 2

Original Dialogue	After Reformulation
بهروز: قربون اشکات برم، سرتو بگیر بالا. پرتو؟ عزیزم؟ .پرتو: گریه نمیکنم، حالم خوبه !بهروز: همیشه واسم مهم بودی بهروز: قهوتو بخور... اصلا بذار خودم واست شکر بریزم. تو حواست نیست، زیادی شیرینش می‌کنی. یه قاشق خوبه؟ دو .قاشق؟ بسه؟ پرتو فقط بهم بگو چند قاشق؟ جون بهروز پرتو: بسه بهروز. چرا دیوونه بازی در می‌آری؟ بهروز: واسه خنده تو. زود بخور عزیزم. اینم سرد میشه آ... .نوش جون ..پرتو: ممنون بابت قهوه	...بهروز: عزیز دلم، این اشکاتو پاک کن، ببین منم اینجام .پرتو: من گریه نمی‌کنم، فقط خسته‌م .بهروز: همیشه برام مهم بودی. هنوزم هستی قهوتو بخور... نذار سرد بشه. حتی اگه بخوای، خودم شیرینش .می‌کنم. یک قاشق؟ دو تا؟ بگو چندتا؟ فقط بگو پرتو: بس کن بهروز، باز داری شوخی درمیاری؟ .بهروز: فقط برای اینکه بخندی... نمی‌خوام قهوت تلخ بمونه .پرتو: مرسی، واسه قهوه
Behrouz: Oh, let me kiss those tears of yours—lift your head up. Parto? Dear? Parto: I’m not crying. I’m fine. Behrouz: You’ve always been important to me! Behrouz: Drink your coffee… Actually, let me add the sugar for you. You’re not paying attention—you make it too sweet. Is one spoon good? Two spoons? Enough? Parto, just tell me how many spoons? For Behrouz’s sake. Parto: That’s enough, Behrouz. Why are you acting crazy? Behrouz: For your smile. Drink it quickly, dear. Otherwise it’ll get cold… Enjoy. Parto: Thanks for the coffee.	Behrouz: My dear, wipe away those tears—look, I’m here too. Parto: I’m not crying, I’m just tired. Behrouz: You’ve always been important to me. You still are. Behrouz: Drink your coffee… don’t let it get cold. If you want, I’ll even sweeten it for you myself. One spoon? Two? Tell me how many—just say it. Parto: Stop it, Behrouz. Are you joking again? Behrouz: Just to make you laugh… I don’t want your coffee to stay bitter. Parto: Thanks for the coffee.