Automated Detection of Cognitive Distortions in Azerbaijani for Mental Health Applications

Laman Guluzada [1,2,3,*] and Samir Rustamov [1,3]

[1] School of Information Technology, ADA University, Baku, Azerbaijan
[2] School of Engineering and Applied Science, The George Washington University, Washington, DC, USA
[3] MegaSec AI LLC, Baku, Azerbaijan
Email: lguluzada4602@ada.edu.az (L.G.); samirrustamov@ada.edu.az (S.R.)
*Corresponding author

***Abstract*—This paper examines cognitive distortion classification in Azerbaijani under low-resource conditions. We adapt the Chinese C2D2 dataset and the English Therapist Question-and-Answer dataset through structured translation, and we construct a consensus-filtered synthetic Azerbaijani dataset. The translation workflow combines fixed label mappings with automated field-level quality checks. Across lexical, multilingual transformer, Qwen-based, and hybrid classifiers, dataset structure was as influential as model choice. The hybrid Term Frequency-Inverse Document Frequency (TF-IDF) + multilingual Bidirectional Encoder Representations from Transformers (mBERT) model achieved 0.6762 accuracy and 0.6276 macro-averaged F1-Score on the controlled C2D2-AZE benchmark, whereas performance was markedly lower on Therapist Question-and-Answer (Q&A) dataset. Qwen2.5-3B-Instruct supervised fine-tuning (SFT) achieved 0.8014 accuracy and 0.7793 macro-averaged F1-Score on Synthetic-AZE, showing that the task is learnable under curated conditions. Distorted-span extraction and retrieval-augmented reframing are included as exploratory proof-of-utility extensions. Overall, the study shows that Azerbaijani cognitive distortion detection is feasible, but that meaningful progress depends on careful data construction, label stability, and dataset-aware evaluation.**

***Keywords*—cognitive distortion detection, Azerbaijani Natural Language Processing (NLP), low-resource NLP, mental health NLP, text classification, Qwen**

I. Introduction

Cognitive distortions are systematic biases in information processing that can lead people to interpret life events, interpersonal interactions, and self-evaluations in excessively negative, inflexible, or maladaptive ways [1]. In cognitive theory, they are recurring patterns of reasoning rather than isolated mistakes, and they shape emotional responses and behavior [2, 3]. Because distorted thinking is associated with depression, anxiety, loneliness-related distress, and therapeutic change, its detection in text is increasingly relevant to mental health NLP [4–7].

Research on automatic cognitive distortion detection has progressed from early feasibility studies [8] to supervised benchmark systems [9, 10], multilingual resources [11], and reasoning-oriented Large Language-Model (LLM) approaches [12]. Nevertheless, results remain difficult to compare because datasets differ in taxonomy, text length, domain, and annotation practice. Fine-grained multi-class classification is particularly difficult for longer patient-style text [10, 13]. Azerbaijani text classification has previously been studied in general-domain settings. Earlier work examined sentiment and polarity classification in social news articles [14] and, more recently, in news and social media data [15]. Other studies introduced a large annotated Azerbaijani news classification dataset [16] and investigated data augmentation for low-resource text classification [17]. These studies demonstrate progress in Azerbaijani classification, but they focus mainly on polarity, subjectivity, and topic classification. Comparable resources and evaluation benchmarks are not yet available for mental-health-oriented classification in Azerbaijani, and cognitive distortion detection remains unexplored.

To our knowledge, this is the first systematic Azerbaijani benchmark-adaptation and evaluation pipeline for cognitive distortion detection. The contribution is data-centered rather than architectural: the study adapts established resources, constructs a consensus-filtered synthetic resource, and evaluates representative approaches under transparent low-resource conditions. It addresses the following research questions:

- **RQ1:** How does dataset structure, particularly the contrast between controlled thought-level examples and realistic support-seeking texts, affect cognitive distortion classification performance?
- **RQ2:** What level of performance can be obtained from translation-based and synthetic Azerbaijani data under low-resource conditions?
- **RQ3:** How do lexical, multilingual transformer, Qwen-based, and prompted LLM approaches compare across these settings?

Manuscript received May 28, 2026; revised July 24, 2026; accepted September 18, 2026; published November 12, 2026.

The main contributions of the paper are as follows:

- We present the first systematic Azerbaijani benchmark adaptation and evaluation pipeline for cognitive distortion detection, built from source-dataset benchmarking, translation-based adaptation, synthetic data construction, and consensus filtering.
- We compare lightweight lexical baselines, multilingual transformer models, and Qwen-based classifiers across existing and Azerbaijani-adapted datasets.
- We show that dataset structure shaped performance as much as model architecture, with controlled thought-level benchmarks proving much easier than realistic support-seeking texts.
- We include distorted-span extraction and retrieval-augmented cognitive reframing as exploratory proof-of-utility extensions rather than as primary benchmark tasks.

II. Related Work

Previous work shows that cognitive distortion detection can be formulated as an NLP task, but also that the task definition matters. Simms *et al.* [8] framed it as machine-learning text analytics, whereas Zhao *et al.* [18] applied neural text classification to fine-grained categories. Later studies reported stronger performance on controlled or crowdsourced data than on realistic therapeutic text. Shickel *et al.* [9] found that both machine-learning and neural methods performed better on a balanced crowdsourced dataset than on smaller, noisier clinical-style data. Mostafa *et al.* [19] examined journaling text, while Shreevastava and Foltz [10] reported fine-grained F1 below 0.30 for therapist-patient interactions despite stronger binary detection. The difference between simplified and authentic data remains a central issue in this field.

Recent research has expanded the task in three directions. First, multilingual and cross-lingual work has grown. Alhaj *et al.* [20] extended the task into Arabic social media, Wang *et al.* [11] introduced C2D2 as a large Chinese benchmark, Sastra *et al.* [21] contributed an Indonesian dataset, and Qi *et al.* [22] investigated more generalizable cross-lingual distortion detection with standardized annotations. Second, reasoning-oriented approaches have explored therapist-like prompting, span extraction, and multi-step pipelines. Chen *et al.* [12] proposed diagnosis-of-thought prompting with LLMs, Lim *et al.* [23] combined reasoning refinement with distortion-relevant spans, and Phan *et al.* [24] argued for a span-first pipeline before final classification. Third, surveys and generalization studies have made explicit that inconsistent taxonomies, task definitions, and evaluation setups hinder cumulative progress. Sage *et al.* [25] synthesized this problem at the field level, while Kuber *et al.* [26] showed that transfer across language and register remains fragile.

Taken together, this literature shows that reported performance depends strongly on the dataset rather than on model size alone. Existing resources differ in language, text unit, label granularity, and whether they contain controlled thoughts or naturally occurring support-seeking language. These differences make raw scores difficult to compare across studies and leave little evidence about how established approaches transfer to Azerbaijani. The present study therefore focuses on transparent adaptation and within-dataset comparison, rather than treating one score as a universal measure of cognitive distortion detection.

The present work fits into this literature by focusing on a low-resource language that has not yet been systematically studied for cognitive distortion detection. It also adopts the broader lesson emerging from recent work: cognitive distortion classification is not only a modeling problem, but also a data quality, label stability, and dataset comparability problem.

III. Methodology

A. Research Design

The study uses a staged experimental design for cognitive distortion classification and Azerbaijani adaptation. Because no established Azerbaijani pipeline was available, existing approaches were first tested on source datasets. The study then moved to Azerbaijani resource construction, translated-data experiments, span extraction, and reframing. This progression reflects the fact that performance in cognitive distortion detection depends heavily on dataset structure, text length, label definitions, and whether the input consists of controlled examples or realistic support-seeking text. Fig. 1 summarizes the workflow.

B. Datasets and Task Formulation

The main task was formulated as single-label multiclass classification. Although several distortions may co-occur in natural discourse, the single-label formulation was used because the source datasets were annotated in this way and because it allowed consistent comparison across experiments.

Two external resources were used as source datasets: C2D2 and Therapist Q&A. C2D2 is a controlled Chinese dataset of 7500 thought examples centered on everyday scenes and target distortions [11], whereas Therapist Q&A contains longer, more realistic support-seeking texts and was used here in a working version of 2530 annotated rows [10]. These resources represent complementary classification conditions: C2D2 is short and label-focused, while Therapist Q&A is longer, more narrative, and more ambiguous.

Since no large-scale Azerbaijani cognitive distortion benchmark was available, Azerbaijani experiments were built through two routes. First, both source datasets were translated into Azerbaijani, producing controlled benchmark adaptations of C2D2 and Therapist Q&A. Second, a separate synthetic Azerbaijani dataset of help-seeking texts was generated and then filtered for label stability through repeated annotation and cross-configuration agreement. Table I demonstrates the dataset summary for each stated dataset.

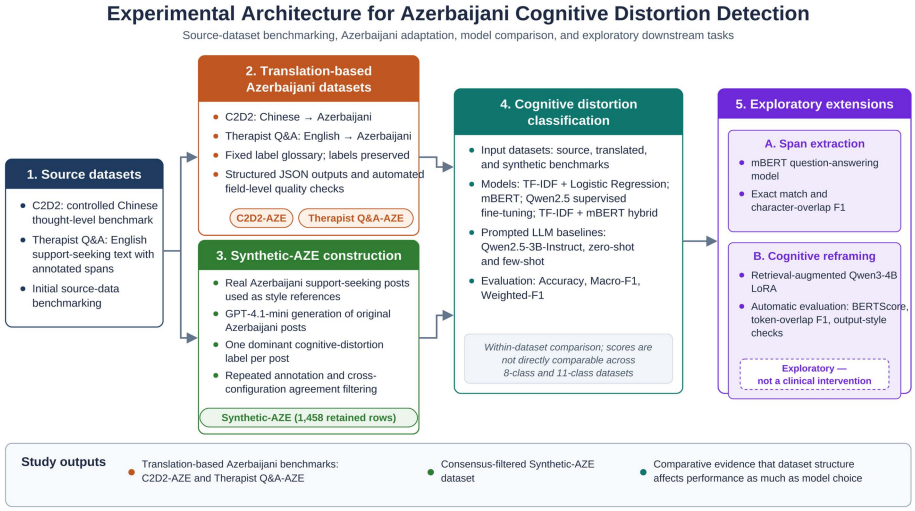


Fig. 1. Overall experimental pipeline. The reframing extension was evaluated using automatic semantic-overlap and output-style checks, as described in Section IV.

TABLE I. DATASET SUMMARY

Dataset	Language/Origin	Rows	Classes	Text Unit	Adaptation Route
C2D2	Chinese source	7500	8	Scenario + thought	Original benchmark; source for Azerbaijani translation
Therapist Q&A	English source	2530	11	Patient question + span	Original benchmark; source for Azerbaijani translation
C2D2-AZE	Azerbaijani adaptation of C2D2	7500	8	Scenario + thought	Structured translation with glossary-mapped labels
Therapist Q&A-AZE	Azerbaijani adaptation of Therapist Q&A	2530	11	Patient question + span	Structured translation with glossary-mapped labels
Synthetic-AZE	Synthetic Azerbaijani data	1458	11	Help-seeking post	LLM generation followed by cross-configuration consensus filtering

C. Data Preprocessing

Preprocessing standardized the text fields and protected the evaluation against leakage. Empty examples were removed, and label names were normalized where they referred to the same underlying distortion. Each dataset nevertheless retained its native class inventory during modeling. Depending on the task, translated data were represented as *thought only*, *scenario + thought*, *span only*, or *full question*. Train, validation, and test splits were created at the original-item level before these derived views were constructed.

D. Taxonomy Harmonization and Experiment-Specific Label Sets

Taxonomy harmonization was limited to label names used in preprocessing and reporting; each dataset retained its native class inventory. C2D2 and C2D2-AZE therefore used the original eight-class taxonomy, whereas Therapist Q&A, Therapist Q&A-AZE, and Synthetic-AZE used eleven classes. In Synthetic-AZE, *catastrophizing* is reported as *Magnification* to align with the CBT terminology used elsewhere in the paper. No classes were collapsed, expanded, or imputed.

Table II gives the label mappings used for cross-experiment reporting.

E. Translation-Based Azerbaijani Adaptation

C2D2-AZE was produced through a controlled Azerbaijani translation procedure. The Chinese distortion labels were mapped through a fixed Azerbaijani glossary rather than translated freely; only the *scenario* and *thought* fields were translated. This design reduced variation in how categories were expressed and preserved label consistency. The mapping was *非扭曲* → *Təhrif yoxdur*, *读心术* → *Fikir oxuma*, *过度泛化* → *Həddindən artıq ümumiləşdirmə*, *情绪化推理* → *Emosional əsaslandırma*, *乱贴标签* → *Etiketləmə*, *个人化归责* → *Şəxsiləşdirmə və özünü günahlandırma*, *非黑即白* → *Ağ-qara düşüncə*, and *算命* → *Gələcəyi mənfi proqnozlaşdırma*. Prompts instructed the model to retain psychological meaning, emotional intensity, first-person inner-speech style, and cues associated with the target distortion, while avoiding rationalization, correction, and explanatory reformulation. Outputs were returned as structured JSON with the original row identifier and translated *scenario_az* and *thought_az* fields. Translation used GPT-5.4 through the OpenAI Responses API with default decoding parameters.

Therapist Q&A-AZE was adapted using the same controlled-label approach, with adjustments for the source data structure. Dominant and secondary labels were mapped through a fixed Azerbaijani glossary, and the full patient question and annotated distorted span were translated together for each row while preserved as separate output fields. Accordingly, *patient_question_az* contains the full translated support-seeking text, whereas *distorted_part_az* contains the translated distorted segment only. If no distorted span was annotated in the source row, the corresponding Azerbaijani field was left empty. For the span-extraction experiments, the translated distorted span was then matched back to the translated full question, and rows with unmatched spans were excluded.

TABLE II. TAXONOMY MAPPING USED FOR CROSS-EXPERIMENT REPORTING

Canonical Reported Label	C2D2 / C2D2-AZE	Therapist Q&A/Therapist Q&A-AZE	Synthetic-AZE Raw Label
All-or-Nothing Thinking	非黑即白 (All-or-nothing thinking)	All-or-nothing thinking	all_or_nothing
Emotional Reasoning	情绪化推理 (Emotional reasoning)	Emotional Reasoning	emotional_reasoning
Fortune-telling	算命(Fortune-telling)	Fortune-telling	fortune_telling
Labeling	乱贴标签 (Putting labels on people)	Labeling	labeling
Magnification	Not in C2D2	Magnification	catastrophizing
Mental filter	Not in C2D2	Mental filter	mental_filter
Mind Reading	读心术 (Mind reading)	Mind Reading	mind_reading
No Distortion	非扭曲 (No distortion)	No Distortion	no_distortion
Overgeneralization	过度泛化 (Overgeneralization)	Overgeneralization	overgeneralization
Personalization	个人化归责 (Personalization)	Personalization	personalization
Should statements	Not in C2D2	Should statements	should_statements

The annotated 11-class label mapping used for Therapist Q&A-AZE was: *No Distortion → Təhrif yoxdur*, *Mind Reading → Fikir oxuma*, *Overgeneralization → Həddindən artıq ümumiləşdirmə*, *Magnification → Şişirtmə*, *Labeling → Etiketləmə*, *Personalization → Şəxsiləşdirmə və özünü günahlandırma*, *Fortune-telling → Gələcəyi mənfi proqnozlaşdırma*, *Emotional Reasoning → Emosional əsaslandırma*, *Mental filter → Mental filtr*, *Should statements → Gərək/mütləq tipli fikirlər*, and *All-or-nothing thinking → Ağ-qara düşüncə*. As in C2D2-AZE, the original class structure of the dataset was retained, and the label mapping was used only to ensure consistency in preprocessing and presentation.

F. Translation Quality Control and Scope

GPT-5.4 was selected for the translation stage because it supports multilingual text processing, instruction following, and structured outputs through the OpenAI Responses API [27]. This made it suitable for a controlled workflow in which source labels were preserved through a fixed glossary and text fields were returned in a row-linked JSON schema. Prior work has found GPT-family models to be competitive for machine translation across a range of language directions, while also noting limitations for low-resource and domain-specific settings [28]. This evidence motivated the choice of translation model and procedure, but was not treated as direct validation that every Azerbaijani output was error-free.

Translation quality control was implemented as a controlled data-adaptation procedure rather than as a one-pass general translation task. The prompts required preservation of source meaning, emotional intensity, first-person self-talk, and cues associated with the assigned cognitive distortion, while prohibiting correction, explanation, or softening of the thought. For C2D2-AZE, automated checks examined row identity and ordering, missing Azerbaijani outputs, leftover Chinese text, unusual length changes relative to the source, and highly repeated scenario–thought pairs. For Therapist Q&A-AZE, the checks examined missing *patient_question_az* values, consistency between source and translated distorted-span fields, label-field consistency after glossary mapping, and unusually short, long, or potentially untranslated outputs. These checks targeted missing translations, residual source-language text, length anomalies, row-level mismatches, and annotation-structure inconsistencies; flagged records were written to separate review files for inspection.

An Azerbaijani-native author reviewed the translated dataset for fluency, readability, field completeness, and consistency between each text and its assigned cognitive-distortion label. This review complemented the automated quality checks. To supplement it, a stratified sample of C2D2-AZE records was examined using an external machine-translation system: the original Chinese fields and their Azerbaijani translations were separately rendered into English and compared for preservation of the central meaning, emotional tone, and cues associated with the assigned label. These checks were used to identify clear translation discrepancies, but they do not replace an independent Chinese–Azerbaijani bilingual human evaluation of source-to-target semantic fidelity.

G. Synthetic Azerbaijani Data Construction

Synthetic-AZE was constructed to provide a low-resource classification setting beyond direct translation. Publicly accessible Azerbaijani Facebook and Reddit posts were used to inform the informal style of the generated texts. They were identified using support-seeking themes such as *problem* (problem), *psixoloji sağlamlıq* (mental health), *depressiya* (depression), *ailə münasibətləri* (family relationships), *stress* (stress), and *narahatlıq* (anxiety). The source material was not included verbatim in the final dataset. GPT-4.1-mini then generated new Azerbaijani social-media-style posts. The prompt required an original post with exactly one dominant target label and explicitly prohibited copying or close paraphrase of the source material. Generation used temperature 0.8, batches of five, eight examples per prompt, and a fixed seed of 42 for source-text selection.

The initial pool contained 2279 examples. Because linguistically plausible generated text may still be ambiguous with respect to its target label, the pool was not used directly for training. It was instead subjected to repeated annotation under two configurations: *cfg_a* used GPT-4.1-mini at temperature 0.7, and *cfg_b* used GPT-4.1 at temperature 0.5. Each configuration annotated the full pool five times. A configuration reached consensus only when the same dominant label appeared in at least four of the five runs. Table III. lists the generation and filtering protocol used for synthetic-Aze data construction.

The final acceptance rule was stricter than single-configuration self-consistency. A synthetic example was retained only if (i) *cfg_a* reached within-configuration consensus, (ii) *cfg_b* reached within-configuration consensus, and (iii) both configurations agreed on the same final dominant label. If either configuration failed to reach the 4/5 threshold, or if the two consensus labels differed, the example was marked cross-configuration ambiguous and removed. Under this protocol, 1458 of the 2279 generated examples were retained and 821 were discarded as unstable or cross-configuration ambiguous.

The two-stage criterion was designed to retain examples whose labels remained stable across repeated runs and across two model configurations. GPT-4.1-mini was used for generation and for *cfg_a*, while GPT-4.1 provided the second annotation configuration. This provides a reproducible LLM-based reliability safeguard for low-resource synthetic-data construction; expert human annotation remains a separate form of validation.

Configuration A reached consensus for 2036 of 2279 rows, with Fleiss' κ = 0.8757 across its five annotation rounds; configuration B reached consensus for 2133 rows, with Fleiss' κ = 0.9254. The final dataset comprises the 1458 examples that met all three acceptance conditions. Table IV reports the generated and retained class distributions. Because the final consensus label could differ from the label assigned during generation, the two columns describe distributions at different stages rather than per-label retention rates.

TABLE III. SYNTHETIC-AZE GENERATION AND FILTERING PROTOCOL

Stage	Model	Setting	Outcome
Generation	GPT-4.1-mini	Temp 0.8; batch 5; 8 style anchors; seed 42	2279 rows generated
Annotation cfg_a	GPT-4.1-mini	Temp 0.7; 5 runs; consensus = 4 of 5	Internal consensus labels
Annotation cfg_b	GPT-4.1	Temp 0.5; 5 runs; consensus = 4 of 5	Internal consensus labels
Final retention	Cross-config rule	Both configs had to reach consensus and agree	1458 retained; 821 discarded

TABLE IV. SYNTHETIC-AZE CLASS DISTRIBUTIONS BEFORE AND AFTER CROSS-CONFIGURATION FILTERING

Reported Label	Generated Pool (2279)	Retained Set (1458)	Distributional Note
All-or-Nothing Thinking	205	168	-
Emotional Reasoning	283	22	-
Fortune-telling	160	150	-
Labeling	215	151	-
Magnification	213	171	Generated label recorded as catastrophizing
Mental filter	247	101	-
Mind Reading	100	168	-
No Distortion	227	145	-
Overgeneralization	221	108	-
Personalization	190	145	-
Should statements	218	129	-

H. Reframing Data Construction

The reframing extension used two separate Azerbaijani resources with different roles. The first was the cognitive-reframing dataset released by Sharma *et al.* [29], which contains 600 practitioner-authored situations, automatic thoughts, and reference reframes. We translated this resource into Azerbaijani. Before splitting, 16 records matching the predefined crisis-content screen were excluded because this exploratory component was not designed for crisis response; 584 records remained. To prevent near-duplicate situation-thought pairs from crossing partitions, the data were grouped by the translated pair and divided with seed 42 into 468 training, 58 validation, and 58 test records. The final Qwen model was not trained on this translated resource: its 58-record test partition was reserved as a separate held-out Azerbaijani evaluation set.

The selected Qwen3-4B reframing model was trained instead on a synthetic C2D2-AZE reframe set. For each Azerbaijani C2D2 scenario, thought, and distortion label, GPT-5.4 generated a short, second-person balanced reframe with temperature 0.2. Generated outputs were automatically screened for generation failures, blank text, abnormal length, first-person wording, missing second-person wording, prohibited clinical phrasing, and verbatim copying of the input thought. This process yielded 987 usable synthetic rows, which were split with seed 42 into 789 training, 99 validation, and 99 synthetic test rows. The selected model was trained only on the 789 synthetic training rows. The 58-row translated reframing test split was kept outside this training pipeline and was used as a separate held-out evaluation set.

Retrieval was built from the 789 synthetic training rows only, after deduplication, using multilingual E5 representations and top-3 retrieval. For validation and test examples, retrieved neighbors were drawn only from this training memory. Thus, the reported synthetic-test results measure performance on held-out synthetic examples, whereas the 58-row translated test results provide a separate check on transfer to the original translated reframing distribution.

I. Ethics and Release Considerations

This study was conducted as a low-resource NLP investigation rather than as the development of a clinical decision system. The experiments were based on adapted versions of previously existing datasets rather than on newly collected clinical records. The C2D2 source data were obtained directly from the dataset authors, whereas the Therapist Q&A source data were accessed through a publicly available Kaggle release. The reframing source

dataset of Sharma *et al.* [29] is distributed under the CC BY-NC-ND 4.0 license; its Azerbaijani translation will therefore not be publicly released without explicit permission from the rights holders. Any release of the other Azerbaijani translated or derived resources will likewise remain subject to the licensing and redistribution conditions of their source datasets. Where redistribution is not permitted, the reproducible components of the pipeline, including code, prompts, label mappings, and processing procedures, can still be shared.

The mental-health setting also requires caution in how the outputs are interpreted and used. Cognitive distortion classification and automated reframing may be useful for research, reflective support, or educational purposes, but they should not be treated as substitutes for professional psychological assessment or therapy. In particular, automatically generated reframes may be incomplete, inappropriate, or unsafe if deployed without qualified human oversight, especially in cases involving acute distress, self-harm, suicidality, or complex psychiatric conditions. For this reason, the reframing component in this study should be understood as an exploratory downstream task rather than as a standalone intervention system.

A further limitation concerns benchmark status. C2D2-AZE and Therapist Q&A-AZE were created through structured translation, and Synthetic-AZE was created through controlled generation and consensus filtering. These resources are useful for comparative benchmarking under low-resource conditions, but they do not replace naturally authored and expert-validated Azerbaijani mental-health datasets. Accordingly, if C2D2-AZE is used as a benchmark resource, it should be interpreted as a translation-based research benchmark whose release remains conditional on source-data permissions rather than as an unrestricted standalone dataset.

J. Models and Evaluation

The experiments compared TF-IDF baselines, multilingual E5 embeddings with Logistic Regression, mBERT, Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct sequence classifiers, Qwen2.5-3B-Instruct supervised fine-tuning for generative classification, and a hybrid TF-IDF + mBERT ensemble. Classification models were evaluated with accuracy, macro-F1, and weighted-F1, with macro-F1 treated as the most important metric because of class imbalance and minority labels. The Qwen variants served different experimental roles rather than a single model-size comparison: Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct were examined as sequence classifiers to test whether increased capacity improved fine-grained classification, while Qwen3-4B was used only for the separate Low-Rank Adaptation (LoRA)-based reframing generation extension.

Across the classification workflows, train/validation/test splits were created with stratified label-aware partitioning using a fixed random seed of 42, and the reported results correspond to single fixed-split runs unless otherwise stated.

For classification, the split sizes in Table V are reported after expansion into task-specific input views, although the splits themselves were created at the original-item level to reduce leakage.

TABLE V. TRAIN/VALIDATION/TEST SPLIT SIZES USED IN THE MAIN CLASSIFICATION EXPERIMENTS

Experiment Dataset	Train	Val.	Test	Note
C2D2-AZE classification	12004	1498	1498	Built from 7500 original items using *thought only* and *scenario + thought* input views
Therapist Q&A-AZE classification	3305	411	411	Built from 2530 original items using *span only*, *full question*, and *no distortion full question* views
Synthetic-AZE classification	1166	147	145	Consensus-filtered dataset with 1458 retained rows

Table VI reports the key settings of the main model families discussed in the paper. Most experiments were run once on the fixed splits described above. In addition, for C2D2-AZE we conducted a five-seed follow-up analysis for TF-IDF Logistic Regression, mBERT, and the hybrid TF-IDF + mBERT ensemble using seeds 42–46. For each seed, mBERT was independently fine-tuned from the pretrained multilingual BERT checkpoint with the same hyperparameter setting as the main run, and the hybrid model was rebuilt with the corresponding seed-specific mBERT output.

The intended classification split ratio was 80%/10%/10% for training, validation, and testing, respectively, with stratification by label and seed 42. The row counts in Table V reflect the expanded input views; grouping and splitting occurred before expansion. The mBERT run used a fixed four-epoch schedule and no patience-based early stopping or checkpoint selection, because epoch checkpoints were disabled and the final epoch model was evaluated. The Qwen sequence-classification runs used fixed three-epoch schedules, with validation and checkpointing at each epoch; the best saved checkpoint was selected by validation macro-F1, but no patience-based early-stopping callback was used. The mBERT and Qwen classifier scripts used AdamW optimization with the learning-rate, warmup, and weight-decay settings reported in Table VI. In contrast, the reframing SFT run used 8-bit AdamW with a cosine learning-rate schedule, validation and checkpointing every 25 steps, and best-checkpoint selection by validation loss.

The mBERT, Qwen, and span experiments were run in Kaggle GPU notebook environments documented in the thesis as providing two Tesla T4 GPUs. The final Qwen3-4B reframing SFT experiment was run on the university SLURM GPU cluster using one NVIDIA Quadro RTX 6000 GPU with approximately 24 GB memory, CUDA 12.4, and PyTorch 2.10.0+cu128. The reproducibility code uses Python with PyTorch, Transformers (version 4.53.0 or later in the classification environment), Datasets, Accelerate, PEFT, BitsAndBytes, and scikit-learn.

TABLE VI. KEY TRAINING SETTINGS FOR THE MAIN MODEL FAMILIES

Model Family	Representative Model	Main Optimization Settings	Additional Settings
TF-IDF baselines	TF-IDF + Logistic Regression / Linear SVM	Word n-grams 1–2; character n-grams 3–5; max_features = 50,000	Logistic Regression used class-balanced loss, *saga* solver, and random seed 42; Linear SVM used class balancing and random seed 42
Embedding baseline	multilingual-e5-base + Logistic Regression	Embedding batch size = 64; Logistic Regression max_iter = 2000	Class-balanced loss, *lbfgs* solver, and random seed 42
mBERT classifier	bert-base-multilingual-cased	Max length = 512; epochs = 4; learning rate = 2e−5; train/eval batch size = 4/8	Gradient accumulation = 4; warmup ratio = 0.06; weight decay = 0.01; max grad norm = 1.0; class-weighted loss; seed 42
Existing benchmark Qwen classifier	Qwen2.5-1.5B-Instruct with LoRA	Max length = 512; epochs = 3; learning rate = 1e−4; train/eval batch size = 4/8	Gradient accumulation = 4; warmup ratio = 0.05; weight decay = 0.01; LoRA (r, α, dropout) = (16, 32, 0.05); 4-bit loading; bf16; seed 42
Qwen sequence classifier	Qwen2.5-3B-Instruct with LoRA	Max length = 384; epochs = 3; learning rate = 1e−5; train/eval batch size = 1/1	Gradient accumulation = 16; LoRA (r, α, dropout) = (16, 32, 0.05); max grad norm = 0.5; 4-bit loading; seed 42
Synthetic Qwen SFT classifier	Qwen2.5-3B-Instruct SFT with LoRA	Max sequence length = 1024; epochs = 3; learning rate = 5e−5; train/eval batch size = 1/1	Gradient accumulation = 8; warmup ratio = 0.05; weight decay = 0.01; LoRA (r, α, dropout) = (16, 32, 0.0); 4-bit loading; seed 42
Span QA	bert-base-multilingual-cased	Max length = 512; doc stride = 128; epochs = 4; learning rate = 3e−5; train/eval batch size = 8/8	Gradient accumulation = 2; warmup ratio = 0.06; weight decay = 0.01; max grad norm = 1.0; seed 42
Reframing SFT	Qwen3-4B LoRA	Max sequence length = 1536; epochs = 3; learning rate = 3e−5; train/eval batch size = 1/1	Gradient accumulation = 4; LoRA (r, α, dropout) = (16, 32, 0.0); 4-bit loading; seed 42

The hybrid classifier combined the probability outputs of TF-IDF Logistic Regression and mBERT using validation-tuned interpolation:

$$P_{final} = \alpha P_{tfidf} + (1 - \alpha) P_{mbert}, \quad (1)$$

where $\alpha = 0.8$ in the best C2D2-AZE setting. Candidate interpolation weights were swept on the validation split in steps of 0.05, and the setting that gave the best validation macro-F1 was then applied unchanged to the test split.

In addition to classification, we explored two downstream extensions. Distorted-span extraction was modeled as extractive question answering on Therapist Q&A-AZE. Cognitive reframing was modeled through a retrieval-augmented supervised fine-tuning setup using Qwen3-4B LoRA and multilingual E5 retrieval. The construction and separation of the translated and synthetic reframing resources are described in the preceding subsection. These components are included as focused extensions of the classification pipeline rather than as equally developed primary tasks.

IV. RESULTS AND DISCUSSION

Interpretive remarks from Table VII are discussed in the corresponding subsections below rather than being embedded in the summary table itself.

TABLE VII. MAIN EXPERIMENTAL RESULTS

Component	Dataset	Best/Selected Method	Main Result
Existing benchmark classification	C2D2	Qwen2.5-1.5B sequence classifier	Macro-F1 = 0.6267
Existing benchmark classification	Therapist Q&A	Qwen2.5-1.5B sequence classifier	Macro-F1 = 0.3458
Synthetic Azerbaijani classification	Synthetic-AZE	Qwen2.5-3B-Instruct SFT	Macro-F1 = 0.7793
C2D2-AZE classification	C2D2-AZE	Hybrid TF-IDF + mBERT	Macro-F1 = 0.6276
Therapist Q&A-AZE classification	Therapist Q&A-AZE	E5-base + Logistic Regression	Macro-F1 = 0.3125
Span extraction	Therapist Q&A-AZE	mBERT extractive QA	Character-overlap F1 = 0.7042
Reframing extension	RAG-SFT data	Qwen3-4B LoRA with retrieval	Exploratory downstream generation experiment

A. Initial Qwen Benchmarking

Before the Azerbaijani experiments, Qwen-based sequence classification was evaluated on the source datasets to establish a reference point. Qwen2.5-1.5B achieved 0.6644 accuracy and 0.6267 macro-F1 on C2D2, but only 0.3711 accuracy and 0.3458 macro-F1 on Therapist Q&A. This contrast already indicated that controlled thought-level examples are substantially easier than realistic patient-style text

B. Synthetic Azerbaijani Results

The synthetic Azerbaijani experiments show that the consensus-filtered dataset is learnable under curated low-resource conditions. Qwen2.5-1.5B-Instruct sequence classification achieved 0.4685 macro-F1, Qwen2.5-3B-Instruct achieved 0.7605, and Qwen2.5-3B-Instruct SFT achieved the best result: 0.8014 accuracy and 0.7793 macro-F1. The synthetic data therefore aligned particularly well with instruction-style training.

At the same time, this result should not be overinterpreted. The synthetic test set was cleaner and more label-focused than naturally occurring support-seeking language. Strong performance here should therefore be read as evidence of learnability under curated conditions rather than as proof of full real-world generalization.

C. C2D2-AZE Classification

C2D2-AZE was the main Azerbaijani classification benchmark. In this setting, lightweight lexical methods

remained highly competitive: TF-IDF with Logistic Regression achieved 0.6058 macro-F1, exceeding both mBERT and Qwen2.5-3B when used alone. The TF-IDF + mBERT ensemble performed best, with 0.6762 accuracy, 0.6276 macro-F1, and 0.6731 weighted F1. Classification results on C2D2-AZE dataset can be found at Table VIII.

TABLE VIII. C2D2-AZE CLASSIFICATION RESULTS

Model	Acc.	Macro-F1	Weighted-F1
TF-IDF + Logistic Regression	0.6562	0.6058	0.6527
TF-IDF + Linear SVM	0.6402	0.5835	0.6333
E5-base + Logistic Regression	0.5801	0.5436	0.5808
mBERT	0.6262	0.5749	0.6255
Qwen2.5-3B sequence classifier	0.6368	0.5824	0.6312
Hybrid TF-IDF + mBERT	0.6762	0.6276	0.6731

This result suggests that lexical and contextual signals are complementary, although lexical information remains more influential for this dataset. The best interpolation weight was 0.8 for TF-IDF and 0.2 for mBERT, consistent with strong lexical and phrasal cues in C2D2-AZE.

A five-seed follow-up analysis on C2D2-AZE further showed that the ranking between the strongest systems remained stable under modest random variation. Across seeds 42–46, TF-IDF + Logistic Regression reached 0.6058 ± 0.0000 macro-F1 and 0.6562 ± 0.0000 accuracy on the test split. mBERT reached 0.5796 ± 0.0093 macro-F1 and 0.6219 ± 0.0076 accuracy. The hybrid TF-IDF + mBERT ensemble remained strongest, with 0.6287 ± 0.0034 macro-F1 and 0.6773 ± 0.0032 accuracy. The validation-selected interpolation weight remained stable across runs, with mean $\alpha_{tfidf} = 0.8000 \pm 0.0500$ and $\alpha_{mbert} = 0.2000 \pm 0.0500$.

Using the five runs and a two-sided t interval, the 95% confidence interval for macro-F1 was [0.5681, 0.5911] for mBERT and [0.6244, 0.6329] for the hybrid ensemble. The corresponding accuracy intervals were [0.6125, 0.6313] and [0.6734, 0.6812], respectively. These intervals provide a compact measure of run-to-run uncertainty for the main C2D2-AZE comparison.

We also tested the main fixed-split C2D2-AZE predictions using 10,000 paired bootstrap resamples and a two-sided paired randomization test over the same 1,498 test examples. The hybrid ensemble exceeded mBERT by 0.0527 macro-F1 (95% paired bootstrap CI [0.0331, 0.0724], $p < 0.0001$) and TF-IDF Logistic Regression by 0.0219 macro-F1 (95% paired bootstrap CI [0.0059, 0.0379], $p = 0.0057$). These tests support the main fixed-split comparison; the five-seed analysis above separately characterizes run-to-run variation.

Input format also mattered for Qwen. When C2D2-AZE examples were represented as *scenario + thought* rather than *thought only*, Qwen2.5-3B improved from 0.5676 to 0.5969 macro-F1. This indicates that additional context helps when the distorted thought alone is short or underspecified.

D. Therapist Q&A-AZE Classification

The same model families performed substantially worse on Therapist Q&A-AZE. Multilingual E5 with Logistic Regression performed best, reaching 0.3820 accuracy and 0.3125 macro-F1. TF-IDF with Logistic Regression reached 0.3034 macro-F1, and TF-IDF with Linear SVM reached 0.2608. The lower scores are consistent with the nature of the dataset: texts are longer, include more background narration, contain overlapping cues, and do not always map cleanly from full-document wording to one dominant label.

To provide a prompted LLM reference point, we evaluated Qwen2.5-3B-Instruct in zero-shot and few-shot settings with one, three, and five demonstrations per class. On C2D2-AZE, macro-F1 increased steadily from 0.1162 in the zero-shot setting to 0.1335, 0.2256, and 0.2851 with one, three, and five demonstrations per class, respectively. On Therapist Q&A-AZE, the strongest prompted macro-F1 was 0.1148 with three demonstrations per class, compared with 0.0671 in the zero-shot setting. The pattern was not monotonic on this longer and more heterogeneous dataset: five demonstrations per class gave 0.1102 macro-F1. Table IX illustrates these results for each experiment setting discussed. Even at their strongest settings, the prompted baselines remained below the supervised classifiers, indicating that direct instruction prompting alone was not sufficient for reliable fine-grained distortion classification in Azerbaijani under the present setup.

TABLE IX. PROMPTED QWEN2.5-3B-INSTRUCT BASELINES ON AZERBAIJANI CLASSIFICATION DATASETS

Dataset	Setting	Acc.	Macro-F1	Invalid
C2D2-AZE	Zero-shot	0.1689	0.1162	2
C2D2-AZE	Few-shot (1 per class)	0.1903	0.1335	1
C2D2-AZE	Few-shot (3 per class)	0.2490	0.2256	1
C2D2-AZE	Few-shot (5 per class)	0.3057	0.2851	0
Therapist Q&A-AZE	Zero-shot	0.1363	0.0671	1
Therapist Q&A-AZE	Few-shot (1 per class)	0.1825	0.0638	1
Therapist Q&A-AZE	Few-shot (3 per class)	0.1290	0.1148	16
Therapist Q&A-AZE	Few-shot (5 per class)	0.1241	0.1102	14

Invalid outputs denote generations that did not resolve to one of the permitted labels and were scored using the evaluation script's deterministic fallback label. Their higher frequency in the three- and five-shot Therapist Q&A-AZE runs is reported for transparency and further supports interpreting prompted results on this dataset cautiously.

Because the classification experiments span both 8-class and 11-class settings, the macro-F1 values in Fig. 2 should be read as within-dataset comparisons rather than as directly scale-equated performance numbers across all tasks. To make relative task difficulty more interpretable, Table X reports simple test-set accuracy reference baselines for the datasets summarized in the figure.

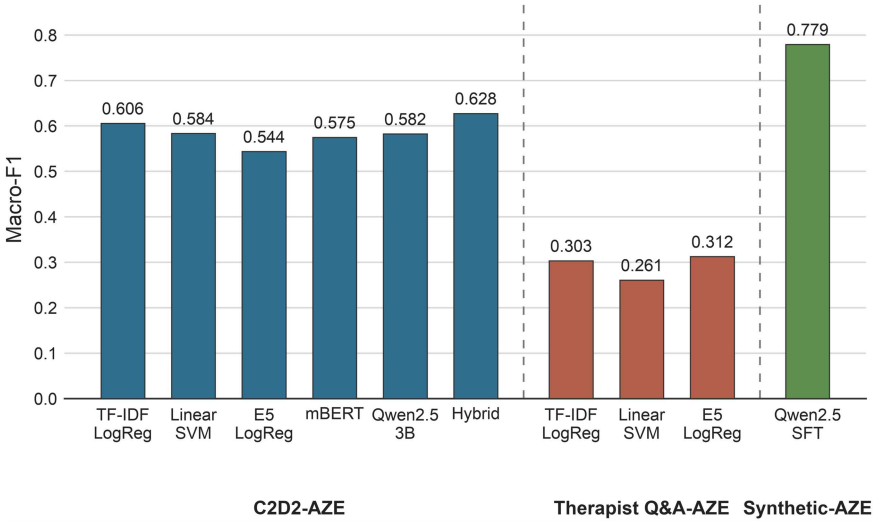


Fig. 2. Macro-F1 comparison across the main classification experiments. Because the figure includes both 8-class and 11-class tasks, it is intended as a compact overview of within-dataset model behavior rather than as a strict cross-dataset scale comparison.

The original Therapist Q&A benchmark row in Table X corresponds to the 10-class benchmark subset used in the initial Qwen comparison stage, whereas Therapist Q&A-AZE was evaluated under the later 11-class adapted setting.

TABLE X. ACCURACY REFERENCE BASELINES FOR THE CLASSIFICATION DATASETS SHOWN IN FIG. 2

Dataset	Classes	Majority	Random
C2D2	8	0.2718	0.1250
Therapist Q&A benchmark subset	10	0.1447	0.1000
C2D2-AZE	8	0.2737	0.1250
Therapist Q&A-AZE	11	0.2263	0.0909
Synthetic-AZE	11	0.1172	0.0909

E. Class-Level and Downstream Findings

Table XI and Fig. 3 provide class-level evidence for the best C2D2-AZE model, namely the hybrid TF-IDF + mBERT ensemble. The strongest classes were *No Distortion* (F1 = 0.8603), *Mind Reading* (F1 = 0.7635), *Personalization* (F1 = 0.7123), and *Labeling* (F1 = 0.6823). The weakest class was *Fortune-telling* (F1 = 0.4146), followed by *All-or-Nothing Thinking* (F1 = 0.4924) and *Overgeneralization* (F1 = 0.5281).

TABLE XI. PER-CLASS TEST PERFORMANCE FOR THE BEST C2D2-AZE MODEL

Class	Prec.	Rec.	F1	Supp.
All-or-Nothing Thinking	0.5159	0.4710	0.4924	138
Emotional Reasoning	0.5621	0.5733	0.5677	150
Fortune-telling	0.4636	0.3750	0.4146	136
Labeling	0.6581	0.7083	0.6823	144
Mind Reading	0.7181	0.8150	0.7635	200
No Distortion	0.8722	0.8488	0.8603	410
Overgeneralization	0.5281	0.5281	0.5281	178
Personalization	0.6933	0.7324	0.7123	142

The confusion matrix shows that the strongest classes were also the most internally stable. *No Distortion* and *Mind Reading* had the clearest diagonals, whereas *Fortune-telling* was often confused with *Mind Reading* and *Overgeneralization*. *All-or-Nothing Thinking* and *Overgeneralization* also showed more diffuse error patterns across neighboring classes. Taken together, these patterns are consistent with the view that some classes rely less on a single lexical cue and more on broader reasoning structure, which may make them harder to separate reliably.

The most common errors also illustrate why these neighboring categories are difficult to separate. A gold *Fortune-telling* example, "Müqaviləni imzalayanda niyə içindəkiləri başa düşmədim, niyə hüquqi şüurum olmadı? Məncə, valideynlərim mütləq məndən məyus olacaqlar." ("Why did I not understand the contract when I signed it? I think my parents will certainly be disappointed in me."), was predicted as *Mind Reading*; the anticipated judgment of others overlaps with the future-oriented conclusion. Another gold *Fortune-telling* example, "Bu dəfə məni rədd etdi, bilirəm, mən həmişə tək qalacağam." ("This time I was rejected; I know that I will always be alone."), was predicted as *Overgeneralization*, because it combines a future claim with an absolute generalization from one event. These examples support the confusion-matrix pattern without treating lexical overlap as a complete explanation of model behavior.

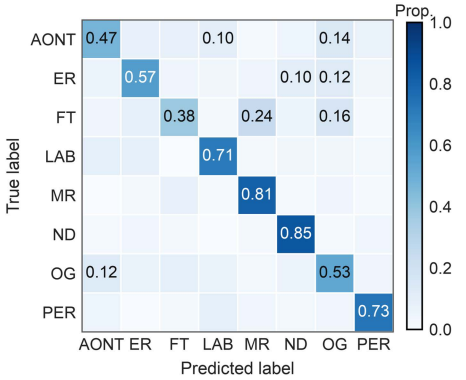


Fig. 3. Row-normalized confusion matrix for the best C2D2-AZE model. Abbreviations: AONT = All-or-Nothing Thinking, ER = Emotional Reasoning, FT = Fortune-Telling, LAB = Labeling, MR = Mind Reading, ND = No Distortion, OG = Overgeneralization, PER = Personalization.

Two downstream experiments were included as supplementary analyses. The span-extraction model achieved 0.2042 exact match and a mean character-overlap F1 of 0.7042. Table XII therefore presents a partial-overlap example rather than a perfect match, as this better illustrates the boundary errors observed in practice.

TABLE XII. ILLUSTRATIVE DOWNSTREAM EXAMPLES WITH ENGLISH RENDERINGS

Task	Example
Span extraction (partial match)	**Text excerpt:** "Mən və ərim artıq bir ildən çoxdur evliyik. Görüşdüyümüz vaxtlarda mən düşünürdüm ki, onda şərəf, sədaqət, bütövlük, dürüstlük və xeyirxahlıq kimi keyfiyyətlər var, həm də zəhmətkeş biri və ümumilikdə hərtərəfli formalaşmış bir insan xüsusiyyətləri göstərirdi. İki aydan sonra onu artıq düz beş ildir tanıyır olacağam və bu müddətin hamısında cütlük olmuşuq." **Gold span:** "mən düşünürdüm ki, onda şərəf, sədaqət, bütövlük, dürüstlük və xeyirxahlıq kimi keyfiyyətlər var, həm də zəhmətkeş biri və ümumilikdə hərtərəfli formalaşmış bir insan xüsusiyyətləri göstərirdi." **Predicted span:** "Görüşdüyümüz vaxtlarda mən düşünürdüm ki, onda şərəf, sədaqət, bütövlük, dürüstlük və xeyirxahlıq kimi keyfiyyətlər var, həm də zəhmətkeş biri və ümumilikdə hərtərəfli formalaşmış bir insan xüsusiyyətləri göstərirdi." **Character-overlap F1:** 0.9438
English span rendering	Gold: "I thought that he had qualities such as honor, loyalty, integrity, honesty, and kindness; he also appeared hardworking and, overall, well-rounded." Prediction: "When we met, I thought that he had qualities such as honor, loyalty, integrity, honesty, and kindness; he also appeared hardworking and, overall, well-rounded."
Span extraction (failure case)	**Text excerpt:** "Hər baxımdan, mən çox uğurlu olmalıyam. Bunu bilirəm, çünki məni o qədər də yaxşı tanımayan adamlar həmişə məndən təsirlənirlər. Mən kifayət qədər yaraşıqlıyam, yüksək IQ-m var, hazırcavabam, cazibədaram, istənilən adamla istənilən mövzuda söhbət başlada bilirəm və tez həll yolları tapa bilirəm." **Gold span:** "Hər baxımdan, mən çox uğurlu olmalıyam." **Predicted span:** "Mən kifayət qədər yaraşıqlıyam, yüksək IQ-m var, hazırcavabam, cazibədaram, istənilən adamla istənilən mövzuda söhbət başlada bilirəm və tez həll yolları tapa bilirəm." **Character-overlap F1:** 0.3010
English span rendering	Gold: "In every respect, I should be very successful." Prediction: "I am attractive enough, have a high IQ, am quick-witted and charming, can start a conversation with anyone about any subject, and can find solutions quickly."
Reframing (useful case)	**Distorted thought:** "Mən bu işdən rədd edildim, deməli heç vaxt uğurlu olmayacağam." **Predicted label:** Fortune-telling **Retrieved/generative reframe:** "Bir işdən rədd olunmağın gələcəkdə heç vaxt uğurlu olmayacağın demək deyil. Bu, sadəcə bir nəticədir və növbəti imkanlarda daha uyğun şəraitlə uğur qazanmağın mümkündür."
English span rendering	Thought: "I was rejected for this job, so I will never be successful." Reframe: "Being rejected for one job does not mean that you will never be successful in the future. It is only one outcome, and success may still be possible in a more suitable future opportunity."
Reframing (failure case)	**Context/thought summary:** grief after the death of a pet **Reference reframe:** "Bu itki çox ağırdır və darıxmağın kədərini hiss etməyin başadüşüləndir, amma bu ağrı bundan sonrakı bütün həyatının yalnız belə qalacağını göstərmir." **Generated reframe:** "Bu itin gəlməsi sənin üçün çox ağır və qorxucu bir an ola bilər, amma bu hadisəni yalnız 'darıxmağın kədəri' kimi görmək məcburi deyil. İndi həmin itin vəziyyətinin hər şeyi daha da gücləndirdiyini hiss edirsən; bu, sadəcə bu anda çox ağır bir dəyişiklik olduğunu göstərir. Sən indi çox pərdəsi var, amma bu hissən hər şeyin tamamilə uğursuz olduğu anlamına gəlmir." **Interpretation:** semantic drift and awkward phrasing remain possible even when the overall output stays in a supportive style.
English span rendering	Reference: "This loss is very painful, and it is understandable that you feel the sadness of missing your pet, but this pain does not mean that the rest of your life will remain this way." Generated: "The arrival of this dog may be a very difficult and frightening moment for you, but you do not have to see this event only as the sadness of missing it. You feel that the dog's situation is intensifying everything; this only shows a very difficult change at this moment. You now have many curtains, but this part of you does not mean that everything has completely failed."

We also evaluated an exploratory retrieval-augmented Qwen3-4B LoRA model for Azerbaijani reframing. The model was trained on 789 synthetic examples and evaluated on a 99-row synthetic test split and a separate 58-row translated held-out test split. As more than one reframe can be acceptable, BERTScore and token-overlap F1 were used instead of exact match. On the synthetic test set, the model obtained a BERTScore F1 of 0.7142 and a mean token-overlap F1 of 0.2142; the corresponding scores on the translated held-out test set were 0.6489 and 0.0791. Output checks found second-person wording in 95.96% and 94.83% of outputs, respectively, with no first-person leakage and very little truncation. The lower scores on the translated test set indicate limited transfer beyond the synthetic training distribution. These downstream results are reported as supplementary experiments, not as evidence of a clinically validated reframing system.

V. LIMITATIONS

The clearest finding is that dataset structure shaped performance as much as model choice. C2D2-AZE behaved as a controlled benchmark, with short label-focused examples and strong lexical cues. Therapist Q&A-AZE was closer to a realistic language-understanding task, where longer support-seeking texts contain several emotional and cognitive signals. The performance difference is therefore not merely numerical; it reflects a change in the nature of the task and in its unit of analysis.

This also explains why lightweight methods remained competitive. TF-IDF-based models were strong on C2D2-AZE because many distortion categories are partly expressed through repeated lexical or phrasal patterns. The best overall result did not come from the largest model, but from combining lexical and contextual signals. This is a useful low-resource lesson: stronger performance did not require a single dominant architecture, but rather an appropriate match between data structure and modeling assumptions.

The synthetic Azerbaijani pipeline further showed that data generation alone is not enough. Consensus filtering improved label reliability by removing examples whose distortion label was unstable across repeated or cross-configuration judgments. This does not replace expert human annotation, but it offers a practical intermediate route for low-resource experimentation when naturally annotated data are unavailable. The downstream results are best interpreted in the same spirit: they show that the classifier can support interpretable and generative

extensions, but they are not yet presented as fully mature standalone tasks.

Several further limitations should be kept in view when interpreting the reported scores. First, although a five-seed follow-up analysis and paired significance tests were added for the main C2D2-AZE comparison, most other model results in the paper still come from single fixed-split runs and were not subjected to formal statistical testing. Accordingly, relatively small score differences outside the main comparison should still be interpreted cautiously. Second, the prompted LLM comparison included zero-shot and one-, three-, and five-shot-per-class settings, but these were intentionally kept as reference baselines rather than exhaustively optimized through broader prompt engineering, self-consistency, or larger in-context example sets. Third, although the reframing extension was supplemented with automatic semantic-overlap evaluation, it was not designed or validated as a clinically reliable intervention system. Manual inspection still revealed occasional semantic drift and awkward phrasing, so this component should be interpreted as an exploratory downstream demonstration rather than as a mature standalone application.

VI. Conclusion

This paper presented an Azerbaijani cognitive distortion analysis pipeline that combines source-dataset benchmarking, translation-based adaptation, synthetic data construction, consensus filtering, and focused downstream extensions. The results show that Azerbaijani cognitive distortion classification is feasible, but that reliable progress depends on data quality, label stability, and dataset-aware evaluation. The strongest Azerbaijani classification result came from the hybrid TF-IDF + mBERT model on C2D2-AZE, while the lower Therapist Q&A-AZE scores show that realistic support-seeking text remains much harder than controlled thought-level examples.

The findings should be read with several limitations in mind. C2D2-AZE and Therapist Q&A-AZE are translation-based adaptations, so translationese or subtle shifts in pragmatic meaning may affect how their results generalize to naturally authored Azerbaijani text. Automated quality checks reduced obvious translation and structural errors, but they do not replace independent bilingual or expert review; full manual validation of the translated benchmarks remains outside the scope of this study. In addition, the single-label formulation follows the source annotations but cannot represent co-occurring distortions in natural discourse. Finally, the synthetic test set was constructed through generation and consensus filtering and is consequently cleaner and more label-stable than naturally occurring support-seeking language. Its results should therefore be interpreted as evidence of learnability under curated conditions, not as an estimate of real-world clinical performance.

More broadly, the study supports a methodological as well as an empirical conclusion. Low-resource cognitive distortion detection should not be approached only through larger models or additional training data. It also requires careful data construction, transparent quality control, comparison of simple and complex baselines, and modular extensions that support interpretation and downstream use. Future work should develop naturally annotated Azerbaijani resources, explore multi-label formulations, and assess reframing outputs through human judgments of relevance, specificity, and safety. Subject to the licensing and redistribution conditions of the source datasets, the translated and derived resources should be released for research use together with the reproducible pipeline components.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Laman Guluzada and Samir Rustamov: conceptualization, methodology, validation, investigation, writing—original draft preparation, review and editing; Laman Guluzada: software, formal analysis, data curation, visualization; Samir Rustamov: supervision; all authors had approved the final version.

Acknowledgment

This work was conducted at the MegaSec AI Laboratory under the Center for Data Analytics Research, ADA University.

References

[1] F. Zaiden, M. Mahfar, A. A. Senin, and F. M. Fakhruddin, "Global research pattern of cognitive distortion: A bibliometric analysis," *Sage Open*, vol. 13, no. 4, 2023.

[2] A. T. Beck, "Thinking and depression: I. Idiosyncratic content and cognitive distortions," *Archives of General Psychiatry*, vol. 9, no. 4, pp. 324–333, 1963.

[3] C. L. Yurica and R. A. DiTomasso, "Cognitive distortions," in *Encyclopedia of Cognitive Behavior Therapy*, Springer, 2005, pp. 117–122.

[4] K. Rnic, D. J. A. Dozois, and R. A. Martin, "Cognitive distortions, humor styles, and depression," *Europe's Journal of Psychology*, vol. 12, no. 3, pp. 348–362, 2016.

[5] S. Jager-Hyman, A. Cunningham, A. Wenzel, S. Mattei, G. K. Brown, and A. T. Beck, "Cognitive distortions and suicide attempts," *Cognitive Therapy and Research*, vol. 38, no. 4, pp. 369–374, 2014.

[6] K. Floyd, C. D. Ray, and J. K. Boumis, "Cognitive distortions associated with loneliness: An exploratory study," *Behavioral Sciences*, vol. 15, no. 8, 1061, 2025.

[7] J. B. Persons, C. D. Marker, and E. N. Bailey, "Changes in affective and cognitive distortion symptoms of depression are reciprocally related during cognitive behavior therapy," *Behaviour Research and Therapy*, vol. 166, 104338, 2023.

[8] T. Simms, C. Ramstedt, M. Rich, M. Richards, T. R. Martinez, and C. G. Giraud-Carrier, "Detecting cognitive distortions through machine learning text analytics," in *Proc. IEEE Int. Conf. Healthcare Informatics*, 2017, pp. 508–512.

[9] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, "Automatic detection and classification of cognitive distortions in mental health text," in *Proc. 2020 IEEE 20th Int. Conf. Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 275–280.

[10] S. Shreevastava and P. W. Foltz, "Detecting cognitive distortions from patient-therapist interactions," in *Proc. Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pp. 151–158.

[11] B. Wang, P. Deng, Y. Zhao, and B. Qin, "C2D2 Dataset: A Resource for Cognitive Distortion Analysis and Its Impact on

Mental Health," in *Proc. Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 10149–10160.
[12] Z. Chen, Y. Lu, and W. Y. Wang, "Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting," in *Proc. Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 4295–4304.
[13] J. S. Tauscher, K. Lybarger, X. Ding, A. Chander, W. J. Hudenko, T. Cohen, and D. Ben-Zeev, "Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness," *Psychiatric Services*, vol. 74, no. 4, pp. 407–410, 2023.
[14] S. Mammadli, S. Huseynov, H. Alkaramov, U. Jafarli, U. Suleymanov, and S. Rustamov, "Sentiment polarity detection in Azerbaijani social news articles," in *Proc. Int. Conf. Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 703–710.
[15] N. Guliyev, Z. Rustamov, and S. Rustamov, "Analysis of public sentiment in Azerbaijani news and social media," in *Proc. Int. Conf. Computing, Machine Learning and Data Science (CMLDS)*, 2024, 22, pp. 1–6.
[16] S. Rustamov, F. Hajiyev, A. A. Ziyaden, A. Yelenov, and A. A. Pak. (2023). AzerNewsV1: Azerbaijani news classification dataset. *IEEE DataPort*. [Online]. Available: https://ieee-dataport.org/documents/azernewsv1-azerbaijani-news-classification-dataset
[17] A. A. Ziyaden, A. Yelenov, F. Hajiyev, S. Rustamov, and A. A. Pak, "Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages," *PeerJ Computer Science*, vol. 10, e1974, 2024.
[18] X. Zhao, C. Miao, and Z. Xing, "Identifying cognitive distortion by convolutional neural network based text classification," *International Journal of Information Technology*, vol. 23, no. 1, pp. 1–12, 2017.
[19] M. Mostafa, A. El Bolock, and S. Abdennadher, "Automatic detection and classification of cognitive distortions in journaling text," in *Proc. 17th Int. Conf. Web Information Systems and Technologies (WEBIST)*, 2021, pp. 444–452.
[20] F. Alhaj, A. Al-Haj, A. Sharieh, and R. Jabri, "Improving Arabic cognitive distortion classification in Twitter using BERTopic," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 854–860, 2022.
[21] N. P. Sastra *et al.,* "Dataset on cognitive distortions for text classification in Indonesian language," *Data in Brief*, vol. 61, 111836, 2025.
[22] H. Qi *et al.,* "Generalizable cross-lingual cognitive distortion detection with standardized annotations and multi-task learning," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 16041–16051.
[23] S. Lim, Y. Kim, C.-H. Choi *et al.,* "ERD: A framework for improving LLM reasoning for cognitive distortion classification," in *Proc. 6th Clinical Natural Language Processing Workshop*, 2024, pp. 292–300.
[24] M.-C. Phan *et al.,* "From span extraction to classification: A multi-step framework for cognitive distortion analysis," in *Proc. 39th Pacific Asia Conf. Language, Information and Computation*, 2025, pp. 384–393.
[25] A. Sage, J. Keppens, and H. Yannakoudakis, "A survey of cognitive distortion detection and classification in NLP," in *Proc. Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 14884–14899.
[26] A. Kuber, E. Liscio, R. Zhang, C. Figueroa, and P. K. Murukannaiah, "Signs of struggle: Spotting cognitive distortions across language and register," in *Proc. Findings of the 14th Int. Joint Conf. Natural Language Processing and the 4th Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025, pp. 1041–1054.
[27] OpenAI. (2026). GPT-5.4 model. *OpenAI API Documentation*. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-5.4
[28] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, T. Matsushita, Y. J. Kim, M. Afify, and H. Hassan Awadalla, "How good are GPT models at machine translation? A comprehensive evaluation," arXiv preprint, arXiv:2302.09210, 2023.
[29] A. Sharma, K. Rushton, I. Lin, D. Wadden, K. Lucas, A. Miner, T. Nguyen, and T. Althoff, "Cognitive reframing of negative thoughts through human-language model interaction," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 9977–10000.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).