Turkish FrameNet 2.0: Frame Prediction via Retrieval-Augmented LLMs and Automated Quality Estimation

Anonymous ACL submission

Abstract

Building FrameNet-style resources requires costly expert annotation, and predicate coverage remains limited for many languages. We present **Turkish FrameNet 2.0**, expanding the foundational Turkish FrameNet (Marşan et al., 2021) to **846 frames and 8,549 annotated verbal synsets**, and study a semi-automated workflow for 9,238 additional unlabeled synsets. The workflow retrieves 30 candidate frames, predicts frame labels with retrieval-augmented LLMs and LoRA fine-tuning, and uses LLM judges to prioritize predictions for human verification.

Our main findings are that retrieval is the central bottleneck (R@30=75.9%), Turkish definitions are more informative than English translations, and domain-adapted fine-tuning is strongest: LoRA-tuned Qwen2.5-7B reaches 0.393 accuracy, outperforming GPT-4o-mini and Gemini 2.5 Flash prompting. Judge-based triage is useful but model-dependent: API-scale judges are substantially more reliable than local 7–12B judges. On a separate 967-synset expert evaluation, GPT-4o-mini achieves 0.501 soft accuracy, while the candidate-constrained LoRA model reaches 0.480, supporting the use of the pipeline as an annotation aid rather than a replacement for human experts.[1]

1 Introduction

Frame semantics (Fillmore, 1976) describes predicates in terms of the situations they evoke and the participants in those situations. A verb such as *give*, for example, evokes a transfer event involving a giver, a recipient, and an item. Berkeley FrameNet (Baker et al., 1998) turns this idea into a structured inventory of frames and lexical units, and has supported work in semantic role labeling, question answering, and knowledge-rich NLP. Its influence has led to FrameNet-style resources for Spanish (Rüggeberg and Petruck, 2003), German (Burchardt et al., 2006), Japanese (Ohara et al., 2004), Chinese (You and Liu, 2005), and other languages.

The main obstacle to building such resources is annotation cost. For each predicate, annotators must determine which frame best captures its meaning and how that frame differs from nearby alternatives. This requires both linguistic expertise and familiarity with a large, fine-grained inventory. The problem is even harder for languages with less complete semantic resources, where many predicates have no prior frame annotation and closely related frames may appear plausible.

Turkish illustrates this bottleneck. The first Turkish FrameNet (Marşan et al., 2021) introduced 139 frames over approximately 2,700 annotated sentences, providing an important foundation but limited coverage of the verbal lexicon. We introduce **Turkish FrameNet 2.0**, which expands the resource to **846 frames and 8,549 annotated verbal synsets**. A synset is a set of near-synonymous lexical items associated with a shared sense definition (Miller, 1994; Oram, 2001); in our case, synsets come from Turkish WordNet/KeNet (Bakay et al., 2021). Assigning a frame to a synset therefore assigns it to the lexical meaning expressed by the definition, which may cover several Turkish verbs or verbal expressions.

The expanded resource still leaves 9,238 verbal synsets with definitions but no frame labels. Annotating them from scratch would require repeated search through a large frame inventory. The central question of this paper is whether models can reduce that search burden without replacing expert judgment. We frame the task as annotation support: a system retrieves plausible frames, predicts a label, and estimates whether the prediction is likely to be worth rapid human verification.

Our pipeline first retrieves 30 candidate frames

[1] Anonymous repository: https://anonymous.4open.science/r/trframenet-llm-013B/. A non-anonymous version will be released upon acceptance.

using lexical BM25 matching and embedding similarity. It then predicts a frame with either retrieval-augmented prompting or LoRA fine-tuning, and finally uses LLM judges to score the proposed assignment on a five-point appropriateness scale. We evaluate on a held-out portion of the annotated lexical list and on 967 previously unannotated synsets labeled independently by an expert.

Contributions and findings. We make four main contributions. First, we release Turkish FrameNet 2.0, a synset-level resource aligned with Berkeley FrameNet 1.7[2] and Turkish WordNet/KeNet. Second, we formulate the remaining expansion as a human-in-the-loop problem: models narrow the candidate space and prioritize likely labels, while annotators retain final authority. Third, we show that retrieval is the main bottleneck. The gold frame appears among the top 30 candidates for 75.9% of test instances, which sets a hard ceiling for candidate-constrained prediction. Fourth, we compare prompting, fine-tuning, input representations, and judge-based triage. LoRA-tuned Qwen2.5-7B reaches 0.393 accuracy, outperforming prompted GPT-4o-mini and Gemini 2.5 Flash; Turkish definitions are more useful than English translations; and judge-based filtering improves precision but remains reliable only with stronger API-scale judges. On the 967-synset expert evaluation, GPT-4o-mini reaches 0.501 soft accuracy, while the constrained LoRA model reaches 0.480, supporting the use of the pipeline as an annotation aid rather than an automatic replacement.

2 Related Work

2.1 Multilingual FrameNet and Annotation Scaling

FrameNet-style resources have been developed for several languages, including Spanish (Rüggeberg and Petruck, 2003), German (Burchardt et al., 2006), Japanese (Ohara et al., 2004), and Chinese (You and Liu, 2005). These projects show that the Berkeley FrameNet model can be transferred across languages, but also that coverage grows slowly because each new predicate must be mapped to a fine-grained semantic inventory.

For Turkish, predicate-semantic resources include Turkish PropBank and related verb-framing work (Kara et al., 2020), which define predicate senses and argument structures in the PropBank tradition. The most directly relevant FrameNet-style resource is Marşan et al. (2021), which introduced 139 frames over approximately 2,700 Turkish annotated sentences. We build on this sentence-level foundation but expand the resource at the synset level and study model-assisted annotation for thousands of remaining verbal meanings.

Cross-lingual projection can reduce annotation cost by transferring semantic labels through parallel data and word alignment (Daza and Frank, 2020). We do not use projection here because our target is a Turkish synset-level frame label, not token-level alignment in parallel sentences. Instead, we use existing Turkish annotations as supervision and predict new labels within Turkish.

2.2 LLM-Assisted Linguistic Annotation

LLMs have been used as annotation assistants for tasks such as named entity recognition, part-of-speech tagging, and semantic role labeling (Hartmann et al., 2017; Tosun et al., 2026). The relevant point for our setting is the human-in-the-loop role of the model: it can propose candidate labels or rationales, while expert annotators retain final control.

Recent frame-semantic work uses LLMs mainly for data generation, argument identification, or semantic role labeling. Cui and Swayamdipta (2024) generate English frame-semantic training sentences; Pancholy et al. (2021) augment data through sister lexical-unit substitution; Devasier et al. (2025) evaluate LLMs for frame-semantic argument identification; and Rai et al. (2025) use LoRA fine-tuning to improve zero-shot semantic role labeling. Our task is more lexically focused: given a Turkish synset, its definition, and a retrieved candidate list, the model must choose the most appropriate FrameNet label.

2.3 Generative Frame Semantics

Tayyar Madabushi et al. (2025) propose a Generative FrameNet setting in which LLMs generate frame-like structures directly, rather than selecting from the Berkeley inventory. Our setting differs because the output space is fixed and canonicalized to FrameNet 1.7. We therefore study frame selection, not frame generation. Definition-aware modeling is also relevant: Blevins and Zettlemoyer (2020) show that explicit sense or definition information can ground model predictions. Our primary setup follows this intuition by conditioning prediction on

[2] https://framenet.icsi.berkeley.edu/; throughout this paper, *FrameNet* refers to release 1.7.

Metric	Value
Total Sentences	20,372
Total Predicate Annots.	22,521
Unique Frames	716
Unique Lexical Units	7,347
Unique FE Roles	667
Avg. Sentence Length	9.33
Avg. FEs / Sentence	1.88

Table 1: Expanded Turkish FrameNet sentence corpus statistics.

Turkish synset definitions.

2.4 LLM-as-a-Judge

LLM-as-a-Judge methods use a model to rate another model's output. Zheng et al. (2023) show that such judgments can correlate with human preferences, while also identifying biases such as position and verbosity effects. For frame assignment, judging is especially challenging because many wrong labels are semantically close to the correct frame. We therefore evaluate judge reliability for Turkish frame assignment and use judge scores only as a triage signal for human annotation, not as a replacement for expert evaluation.

3 Dataset

We use three related resources. The first is a sentence corpus with frame and frame-element annotations inherited from earlier Turkish FrameNet releases. The second resource is the new human-annotated lexical list contributed in this paper: 8,549 Turkish synset–definition–frame triples. The third resource is the application target for our system, a set of 9,238 verbal synsets that have definitions but do not yet have frame labels. For this third resource, our system retrieves 30 candidate frames for each synset; a human annotator then reviews this candidate set and selects up to three plausible frames, ordered by appropriateness.

3.1 Sentence Corpus

The corpus integrates sentences from six available datasets: Kenet (6,844), Etstur (3,761), Penn-Treebank-20 (3,342) (Kuzgun et al., 2020), FrameNet (2,688), Atis (2,235) (Cesur et al., 2024), and PennTreebank (1,502). After deduplication and quality filtering, the corpus comprises **20,372 sentences**, **22,521 predicate annotations**, **716 unique frames**, and **667 FE roles** (Table 1; Figure 4).

3.2 Annotation Process

The expanded lexical annotation was carried out by five annotators using a shared spreadsheet. For each Turkish verbal synset, annotators examined the synset form and its definition, then selected the most appropriate FrameNet frame after consulting the Berkeley FrameNet database. A common strategy was to identify English lexical items semantically close to the Turkish definition and use their FrameNet entries as evidence for assigning the Turkish synset to a comparable frame.

This workflow also exposed the central difficulty addressed in this paper. Not all relevant English paraphrases are lexicalized in FrameNet, and Turkish definitions often encode fine-grained semantic distinctions that may plausibly correspond to multiple frames. Although annotators had broad familiarity with the frame inventory, exhaustively considering all possible frames for each synset is difficult in practice, especially given the size and semantic overlap of the inventory.

Berkeley FrameNet 1.7 contains approximately 1,200 semantic frames. The 846 frames instantiated in Turkish FrameNet 2.0 reflect frames with at least one Turkish lexicalization in the source synsets; frames evoked exclusively by culturally or linguistically English-specific concepts (e.g., certain sports-, law-, or institution-specific scenarios) were not populated. The expansion is therefore comprehensive with respect to the Turkish synset pool drawn from KeNet, not a complete transfer of the FrameNet 1.7 inventory.

3.3 Linguistic Diversity and Challenges

The resource has the long-tailed distribution typical of lexical semantic inventories. As shown in Figure 1, frame frequencies follow a Zipfian pattern, with the 20 most frequent frames accounting for 39.0% of occurrences. This matters for model evaluation because many frames have few training examples, so a system must generalize beyond frequent labels.

The data also reflects the morphological and lexical properties that make Turkish frame prediction difficult. Analysis with DİLBAZ morphological analyzer (Yıldız et al., 2019) shows that the average lexical unit contains 2.74 morphological segments. In addition, 583 of 6,938 unique expressions are associated with more than one frame. This 8.4% ambiguity rate motivates the use of definitions: the same or similar surface expression may evoke dif-

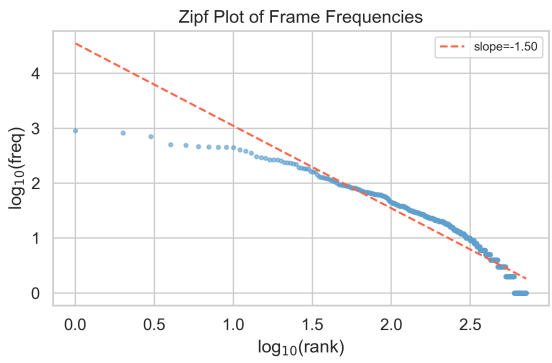

Figure 1: Zipf plot of frame frequencies (slope $= -1.50$, log-log scale), confirming a power-law distribution with a long tail of rare frames.

ferent frames depending on its intended sense.

3.4 Annotated Lexical List and Expansion Target

The **annotated lexical list** contains 8,549 Turkish verbal expressions (synsets) with gold frame labels across 846 canonical frames after normalization (98 case-inconsistency corrections). **76.3% are MWEs**, reflecting the idiomatic character of Turkish verbal predicates. Surface overlap with the sentence corpus is only 0.6% (exact string match), rising to 60.1% via morphological analyzer lemma-based matching, confirming that the low overlap stems from morphological inflection rather than semantic incompatibility.

The **unannotated expansion target** contains **9,238** verbal expressions with definitions but no frame labels, constituting the primary target for automated frame induction.

3.5 Data Partitioning

For Setups A and B (lexical prediction), we apply a stratified **80/20 split** to the annotated lexical list: **6,876 train / 1,673 test** instances, stratified by frame frequency group. For Setup C (sentence context), we use the separate sentence corpus split in Table 2. To bound inference cost across five models while preserving the frame distribution, we evaluate on a stratified random sample of 1,000 instances drawn from the 2,521-instance test partition rather than the full split.

4 Methodology

4.1 Candidate Retrieval

The full inventory contains over a thousand frame labels, which is too large for reliable direct prompting. We therefore retrieve a smaller candidate set before prediction. The retriever combines lexical matching with semantic similarity: BM25Okapi (Robertson and Zaragoza, 2009) indexes the synset and definition text, while a multilingual sentence-transformer (mMPNET; Reimers and Gurevych 2019) provides cosine similarity over pre-computed embeddings. Each component retrieves 15 unique frames, and their ordered union is truncated to 30 candidates.

This retrieval step defines the upper bound for any system that is forced to choose from the candidate list. On the 1,673-instance test set, the gold frame appears in the top candidate for 23.6% of instances, in the top five for 47.5%, in the top ten for 57.1%, and in the top thirty for 75.9%. We use 30 candidates because it substantially improves coverage while keeping prompts short enough for smaller local models. Increasing the list to 50 candidates increased latency by about 40% and gave only a modest recall gain.

Split	Sent.	Frames	Corpora
Train	14,739	716	6
Val	3,112	508	6
Test	2,521	390	6

Table 2: Sentence corpus splits (Setup C). The test partition contains 2,521 instances; Setup C results (Table **??**) are reported on a stratified 1,000-instance sample of this partition.

4.2 Experimental Setups

We evaluate three input configurations. **Setup A** (primary) provides the Turkish synset, its Turkish definition, and the 30 retrieved candidates. **Setup B** replaces the Turkish definition with its English translation obtained via multilingual embedding alignment. **Setup C** replaces synset+definition with a predicate verb in full sentence context, drawn from the sentence corpus and evaluated on a separate 1,000-instance test sample (Table 2).

4.3 Few-Shot Frame Prediction

Few-shot prompts ask the model to select exactly one frame from the retrieved candidate list. Each prompt includes a short task description, five in-context examples retrieved from the training set, and the target synset. The examples are chosen with BM25 but constrained to have distinct frame labels, so that the prompt shows several possible semantic distinctions rather than repeated instances

of the same frame. The model returns JSON containing the predicted frame, a confidence label, and a brief rationale. We also evaluate a zero-shot variant that removes the examples but keeps the same task instruction and candidate list. Decoding is deterministic wherever the model interface supports it.

4.4 Supervised Fine-Tuning

We fine-tune Qwen2.5-7B-Instruct (Qwen et al., 2025) using LoRA adapters (Hu et al., 2022), a parameter-efficient method that updates a small set of low-rank adapter weights instead of all model parameters. The model is trained on the human-annotated lexical list and then asked to generate the frame name directly. This unconstrained setting tests whether fine-tuning can internalize the frame inventory without relying on a retrieved candidate list at inference time.

4.5 LLM-as-a-Judge

Each judge receives (`SYNSET`, `DEFINITION`, `FRAME`) and rates semantic appropriateness 1–5 (JSON output with `score` and `reasoning`). Judges evaluate *gold-annotated* frames to isolate reliability from downstream prediction quality. Five judges: Qwen2.5-7B, LLaMA-3-8B, Gemma-3-12B, GPT-4o-mini, Gemini 2.5 Flash. Reliability: pairwise quadratic-weighted Cohen's κ (Cohen, 1960) on 1,000 stratified test instances.

Human validation protocol. An expert annotator labels 967 synsets from the unannotated expansion target using the same BM25+cosine +FrameNet-1.7 candidate list provided to the system, without access to model predictions. To accommodate cases where multiple frames could plausibly apply, the annotator may select up to three candidate frames per synset (a primary choice and up to two alternatives). Model predictions are then compared against this expert annotation using two metrics: *exact match* (model output equals the annotator's first choice) and *soft match* (model output matches any of the annotator's 1–3 selections).

4.6 Evaluation Metrics

We report four complementary metrics. Accuracy is the proportion of predictions that exactly match the gold frame. Cohen's κ is a chance-corrected agreement measure; it is useful here because frame labels are highly imbalanced, so raw accuracy alone can overstate agreement on frequent frames. Mean reciprocal rank (MRR) measures where the predicted frame appears in the retrieved candidate list, rewarding systems that rank a plausible frame highly even when it is not the exact gold label. Recall@k measures whether the gold frame appears among the top-k retrieved candidates and is therefore a property of the retrieval stage rather than of an individual model.

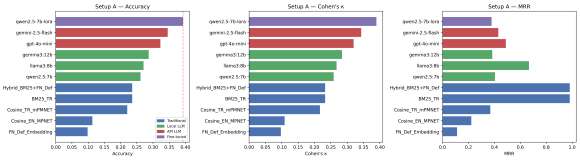

Figure 2: Setup A results: Accuracy (left), Cohen's κ (center), MRR (right). Fine-tuned Qwen2.5-7B leads on Acc/κ; LLaMA-3-8B leads on MRR.

5 Results

Figure 2 and Table 3 summarize the main frame prediction results. The task is difficult: the model must select one label from the entire frame inventory, and many frames are rare. Two simple baselines establish the lower bound. Always predicting the most frequent frame, STIMULATE_EMOTION, reaches only 0.020 accuracy, and frequency-weighted random prediction reaches 0.007. All retrieval and LLM systems therefore learn substantial task-specific signal beyond label frequency.

5.1 Frame Prediction

Finding 1: Domain-adapted fine-tuning is strongest. The best system is Qwen2.5-7B with LoRA fine-tuning, which reaches 0.393 accuracy. This is 4.7 percentage points higher than Gemini 2.5 Flash and 7.0 points higher than GPT-4o-mini in the same Setup A few-shot condition. The result suggests that, once several thousand in-domain annotations are available, adapting a smaller open model can be more effective than prompting a larger general model. The fine-tuned model also reaches 51.8% of the retrieval upper bound, since the gold frame is present in the top-30 candidate set for 75.9% of test instances. When the same fine-tuned model is forced to use the retrieved candidate list, accuracy drops to 0.353, suggesting that it has learned useful information about the frame label space during fine-tuning.

Finding 2: Turkish definitions are more informative than translations. Every model performs better with Turkish definitions than with English

Setup	Category	System	N	Acc	κ	MRR
A	Traditional	BM25_TR	1,673	0.236	0.233	0.236[†]
		Hybrid BM25+FN-Def	1,673	0.236	0.233	0.236[†]
		Cosine TR (mMPNET)	1,673	0.221	0.217	0.369
		Cosine EN (MPNET)	1,673	0.113	0.109	0.223
		FN Def Embedding	1,673	0.099	0.097	0.112
		Majority (STIM._EMOTION)	1,673	0.020	0.000	–
		Random (weighted)	1,673	0.007	0.002	–
	Local LLM	Gemma-3-12B (fs)	1,673	0.287	0.284	0.384
		LLaMA-3-8B (fs)	1,649	0.271	0.268	**0.668**
		Qwen2.5-7B (fs)	1,655	0.262	0.259	0.405
		Gemma-3-12B (zs)	1,673	0.234	0.231	0.251
		LLaMA-3-8B (zs)	1,673	0.213	0.211	0.308
		Qwen2.5-7B (zs)	1,648	0.188	0.185	0.231
	API LLM	Gemini 2.5 Flash (fs)	1,673	0.346	0.343	0.431
		GPT-4o-mini (fs)	1,673	0.323	0.320	0.489
		GPT-4o-mini (zs)	1,673	0.267	0.264	0.345
	Fine-tuned	Qwen2.5-7B-LoRA (unconstr.)	1,673	**0.393**	**0.390**	0.379
		Qwen2.5-7B-LoRA (constr.)	1,673	0.353	0.350	–
B	LLM (fs)	GPT-4o-mini	1,673	0.305	0.302	0.293
		Gemini 2.5 Flash	1,673	0.289	0.285	0.309
		Gemma-3-12B	1,673	0.284	0.281	0.271
		LLaMA-3-8B	1,670	0.265	0.261	0.268
		Qwen2.5-7B	1,673	0.249	0.246	0.245
C	LLM (fs)	Gemma-3-12B	1,000	0.376	0.370	0.507
		LLaMA-3-8B	998	0.351	0.344	0.667
		Qwen2.5-7B	997	0.291	0.284	0.445

Table 3: Frame prediction results across all systems and setups. (fs) = few-shot; (zs) = zero-shot. Recall@k is shared within a setup: A: R@1=0.236, R@5=0.475, R@10=0.571; B: R@1=0.113, R@5=0.234, R@10=0.308; C: R@1=0.388, R@5=0.525, R@10=0.574. [†]BM25 MRR equals R@1 by construction (1-NN retrieval). [‡]Setup C uses a different corpus and test distribution; results are diagnostic, not directly comparable to Setups A/B. **Bold**: best per metric.

translations, although the size of the accuracy gap varies by model. The larger difference appears in retrieval: the top-30 recall falls from 75.9% in Setup A to 45.1% in Setup B. Translation therefore loses information that is important for organizing Turkish predicates into FrameNet frames, especially for idiomatic multi-word expressions and case-marked constructions.

Finding 3: Sentence context helps ranking. Setup C is evaluated on a different sentence-level test set, so its accuracy is diagnostic rather than directly comparable to Setups A and B. Even so, the high MRR values show that sentence context helps models rank plausible frames. This is especially relevant for polysemous predicates, where the definition gives the lexical sense but the sentence clarifies the event being described.

Finding 4: Few-shot examples provide useful calibration. Few-shot prompting consistently improves over zero-shot prompting. The gains range from 5.3 accuracy points for Gemma-3-12B to 7.4 points for Qwen2.5-7B, with even larger relative improvements in MRR. The examples appear to help models interpret the candidate list and match the level of granularity expected by FrameNet labels.

Judge 1	Judge 2	κ
GPT-4o-mini	Gemini 2.5 Flash	**0.663**
Gemma-3-12B	GPT-4o-mini	0.470
Qwen2.5-7B	GPT-4o-mini	0.468
Qwen2.5-7B	Gemini 2.5 Flash	0.436
Gemma-3-12B	Gemini 2.5 Flash	0.384
Qwen2.5-7B	Gemma-3-12B	0.216
LLaMA-3-8B	Gemma-3-12B	0.136
LLaMA-3-8B	GPT-4o-mini	0.081
LLaMA-3-8B	Gemini 2.5 Flash	0.049
LLaMA-3-8B	Qwen2.5-7B	0.033

Table 4: Pairwise quadratic-weighted κ on 1,000 gold-annotated instances (sorted descending).

5.2 LLM-as-a-Judge

GPT-4o-mini and Gemini 2.5 Flash are the most consistent judges, with quadratic-weighted agree-

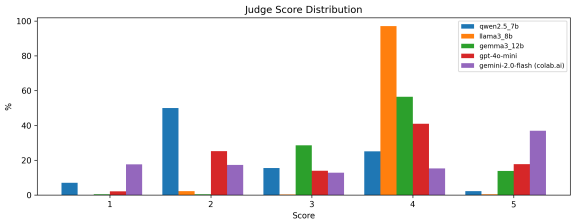

Figure 3: Judge score distributions (1–5 Likert scale). LLaMA-3-8B assigns score 4 to 97% of instances (leniency bias). GPT-4o-mini and Gemini exhibit the most balanced distributions.

Subset	N (%)	GPT-4o Acc	Gemma Acc
All instances	1,000 (100%)	0.291	0.270
Judge score ≥ 4	587 (58.7%)	**0.402**	**0.375**

Table 5: Judge-filtered precision on 1,000 stratified gold-annotated test instances from the annotated lexical list (distinct from the 967-instance expansion set in Table 6). Restricting to GPT-4o-mini score ≥ 4 raises accuracy by $+11.1$ points (GPT-4o-mini) and $+10.5$ points (Gemma-3-12B). The remaining 41.3% are candidates for human review.

Model	Exact	Soft	InCand
GPT-4o-mini	**0.322**	**0.501**	0.997
Gemma-3-12B	0.284	0.453	0.990
LLaMA-3-8B	0.260	0.418	0.995
Qwen2.5-7B	0.237	0.403	0.991
Qwen2.5-7B-LoRA (constrained)	0.308	0.480	0.949
Qwen2.5-7B-LoRA (unconstrained)	0.243	0.378	0.695

Table 6: Model–human alignment on 967 unannotated expansion instances (Study B). Exact: model matches annotator's first-choice frame. Soft: model matches any of annotator's 1–3 options. InCand: proportion within the top-30 candidate list. Single annotator; IAA not applicable.

ment of 0.663. Following the commonly used interpretation of Landis and Koch (1977), this falls in the "substantial" range. Their score distributions are also more balanced than those of the local judges: no single score accounts for more than 41% of their ratings. By contrast, LLaMA-3-8B assigns a score of 4 to 97% of instances, indicating a strong leniency bias. Qwen2.5-7B shows the opposite tendency, with half of its ratings at score 2. Gemma-3-12B is closer to the API models, but its agreement with GPT-4o-mini remains moderate rather than substantial.

These differences matter because judge scores are intended to guide annotation effort. Table 5 shows that a simple GPT-4o-mini threshold can identify a higher-precision subset: keeping only predictions scored 4 or 5 raises accuracy while still covering a majority of instances.

5.3 Human Annotation Comparison

Table 6 reports model–human alignment on the 967-instance unannotated expansion set. GPT-4o-mini achieves the highest exact match (0.322) and soft accuracy (0.501), indicating that its predictions fall within the annotator's acceptable option set in approximately half of cases. Local few-shot models follow the same relative ordering as in Setup A (Gemma-3-12B: soft=0.453, LLaMA-3-8B: soft=0.418, Qwen2.5-7B: soft=0.403), confirming that benchmark rankings transfer to the expansion setting.

The LoRA fine-tuned model exhibits an interesting interaction with the candidate constraint. Under unconstrained inference, it achieves soft accuracy of only 0.378 and generates predictions outside the candidate list in 30.5% of cases (InCand=0.695), as the model was trained without candidate list input and occasionally produces unlisted frame names. Under constrained inference—where the top-30 candidate list is appended to the prompt—accuracy rises substantially (exact=0.308, soft=0.480, InCand=0.949), surpassing all local few-shot models on soft accuracy. This pattern contrasts with Setup A results, where unconstrained LoRA outperformed the constrained variant (0.393 vs. 0.353): on the primary benchmark, the model has internalized sufficient frame knowledge to benefit from unconstrained generation, whereas on the unannotated expansion set—covering novel expressions outside the training distribution— candidate-constrained inference provides necessary grounding.

All few-shot models remain within the candidate list in over 99% of cases (InCand$\geq$0.990), confirming that retrieval-constrained prompting effectively eliminates hallucination regardless of model scale.

The expert annotator frequently used the option to record more than one plausible frame: 9% of synsets received a single frame, 31% received two alternatives, and 60% received three. This distribution confirms that many expansion instances are genuinely ambiguous and that soft accuracy is an informative complement to exact match.

6 Discussion

Cost-effectiveness of domain adaptation. The results support a practical conclusion for resource expansion: when several thousand in-domain examples are available, a 7B model adapted to the task can outperform stronger general-purpose prompted models. This is useful for lexicographic projects because API cost, data-sharing restrictions, and reproducibility concerns often make closed systems difficult to rely on as the main annotation engine.

Why Turkish definitions outperform translations. The comparison between Setups A and B shows that translation is not a neutral preprocessing step. Turkish definitions preserve cues that are relevant to frame choice, including light-verb constructions, case-marked nominals, and aspectual distinctions. English translations often normalize these cues, making different Turkish predicates appear more similar than they are for FrameNet annotation.

Error analysis. We analyze prediction errors for Gemma-3-12B in the Setup A few-shot condition, the strongest local model by accuracy. Performance is lower on rare frames: frames with fewer than 10 training examples obtain 0.251 accuracy, compared with 0.301 for more frequent frames. This five-point gap confirms that the long tail in Figure 1 remains a central challenge.

Multi-word expressions are only slightly harder than single-word predicates when Turkish definitions are available. Accuracy is 0.281 for multi-word expressions and 0.298 for single-token predicates, a 1.7-point difference. The most common errors instead involve semantically neighboring frames. For example, the model sometimes predicts POSTURE for CHANGE_POSTURE, SOUNDS for MAKE_NOISE, or MOTION for SELF_MOTION. Exact-match accuracy treats these errors the same as unrelated labels, but they are qualitatively different: the prediction is often a hypernym or closely related frame. This is why rank-based metrics and judge-based evaluation are useful complements to exact match.

Retrieval as the primary bottleneck. The top-30 retrieval ceiling of 75.9% means that one in four gold frames is unavailable to any candidate-constrained predictor. This limits the possible gains from better generation alone. Future work should therefore improve retrieval, for example with fine-tuned bi-encoders or retrieval models that use FrameNet definitions and frame relations more directly.

Judge reliability and pipeline precision. Only the API-scale judges reach substantial agreement on this task; the local 7–12B judges are not yet reliable substitutes for human evaluation. They may still be useful after task-specific calibration, but the present results caution against treating judge scores as gold labels. In a tiered annotation workflow, predictions that GPT-4o-mini scores as 4 or 5 would be routed to verification, while the remaining 41.3% of instances would receive full human review. This raises precision from 0.291 to 0.402 for GPT-4o-mini predictions and from 0.270 to 0.375 for Gemma-3-12B predictions. The actual time savings depend on how much faster verification is than annotation from scratch, which should be measured in a deployment study.

7 Conclusion

We presented **Turkish FrameNet 2.0**, an expanded resource with 846 frames and 8,549 annotated verbal synsets, and evaluated a semi-automated system for extending it further. The results show that retrieval-augmented prompting is useful, but that domain-adapted fine-tuning is stronger when enough annotated examples are available. They also show that Turkish definitions are better supervision than English translations, and that sentence context provides a complementary signal for ranking candidate frames.

For deployment, the most important result is that automation should be used as triage rather than replacement. The best models are not accurate enough to annotate the expansion target without expert review. However, candidate-constrained prediction and judge-based filtering can identify a subset of assignments that are more suitable for rapid verification. Future work will extend the framework to frame element prediction, fine-tune additional architectures, and measure the time savings of this workflow in a live annotation interface for lexicographers.

Limitations

Scope and coverage. All experiments are conducted on Turkish, so generalization to other languages, domains, and frame inventories remains to be tested. Turkish FrameNet 2.0 instantiates 846 of the approximately 1,200 Berkeley FrameNet 1.7 frames; frames without Turkish lexicalizations in

the source synset pool remain unpopulated, and frame-to-frame relations such as Inheritance, Subframe, and Causative_of are not yet modeled.

Retrieval and modeling limits. The gold frame appears in the top-30 candidate list for 75.9% of test instances, so one in four gold frames is unavailable to candidate-constrained predictors before generation begins. We also fine-tune only Qwen2.5-7B; other architectures, especially LLaMA-3-8B which achieved the highest few-shot MRR, may benefit from similar adaptation.

Annotation and evaluation limits. The expanded lexical list and the 967-instance expansion evaluation reflect expert judgments, but frame assignment can remain subjective for rare or polysemous predicates. The expansion evaluation set is currently annotated by one expert, so it supports external validation but not inter-annotator agreement; a second pass and adjudication are planned.

Judge reproducibility. The judge-filtered precision result depends on GPT-4o-mini and Gemini 2.5 Flash, which are commercial systems subject to unannounced updates. We preserve the judge outputs used in our experiments, but future replication will require access to comparable model snapshots or re-calibration against human ratings.

Ethical Considerations

Data sources. All data used in this work derives from publicly available Turkish corpora [3] and the Berkeley FrameNet (Baker et al., 1998). No personal or sensitive information is present in the data.

Human annotators. The expanded lexical list (Turkish FrameNet 2.0) and the 967-instance expansion evaluation were produced by expert annotators with prior frame-semantic experience. Annotators were either co-authors of this work or formal collaborators of the research group, compensated through their institutional appointments at standard rates for academic research assistants in their respective countries; they provided informed consent for the release of the resulting annotations.

Resource release. Turkish FrameNet 2.0 and the expansion evaluation set will be released under a Creative Commons Attribution license to support continued research on Turkish frame semantics and on methodologies for FrameNet expansion in other languages.

[3] Available at https://github.com/starlangsoftware

Use of generative AI. The authors used commercial large language models (Anthropic Claude, OpenAI ChatGPT) for limited language polishing of the manuscript and for code-generation assistance during pipeline implementation. All scientific contributions, experimental design, analyses, and final wording are the authors' own. The same models used by the system under study (GPT-4o-mini, Gemini 2.5 Flash, Qwen2.5-7B, LLaMA-3-8B, Gemma-3-12B) are documented in Sections 4.3–4.5 as objects of study, not as writing assistants.

References

Özge Bakay, Özlem Ergelen, Elif Sarmış, Selin Yıldırım, Bilge Nas Arıcan, Atilla Kocabalcıoğlu, Merve Özçelik, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Begüm Avar, and Olcay Taner Yıldız. 2021. Turkish WordNet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174, University of South Africa (UNISA). Global Wordnet Association.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Neslihan Cesur, Aslı Kuzgun, Mehmet Kose, and Olcay Taner Yıldız. 2024. Building annotated parallel corpora using the ATIS dataset: Two UD-style treebanks in English and Turkish. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 104–110, Torino, Italia. ELRA and ICCL.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Xinyue Cui and Swabha Swayamdipta. 2024. Annotating FrameNet via structure-conditioned language generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 681–692, Bangkok, Thailand. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Jacob Devasier, Rishabh Mediratta, and Chengkai Li. 2025. Can LLMs extract frame-semantic arguments? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30609–30622, Suzhou, China. Association for Computational Linguistics.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Neslihan Kara, Deniz Baran Aslan, Büşra Marşan, Özge Bakay, Koray Ak, and Olcay Taner Yıldız. 2020. TRopBank: Turkish PropBank v2.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2763–2772, Marseille, France. European Language Resources Association.

Aslı Kuzgun, Neslihan Cesur, Bilge Nas Arıcan, Merve Özçelik, Büşra Marşan, Neslihan Kara, Deniz Baran Aslan, and Olcay Taner Yıldız. 2020. On building the largest and cross-linguistic turkish dependency corpus. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Büşra Marşan, Neslihan Kara, Merve Özçelik, Bilge Nas Arıcan, Neslihan Cesur, Aslı Kuzgun, Ezgi Sanıyar, Oğuzhan Kuyrukçu, and Olcay Taner Yildiz. 2021. Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125, University of South Africa (UNISA). Global Wordnet Association.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Kyoko Hirose Ohara, Seiko Yamaguchi Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2004. The japanese framenet project: An introduction.

Peter Oram. 2001. Wordnet: An electronic lexical database. christiane fellbaum (ed.). cambridge,ma: Mit press, 1998. pp. 423. *Applied Psycholinguistics*, 22(1):131–134.

Ayush Pancholy, Miriam R L Petruck, and Swabha Swayamdipta. 2021. Sister help: Data augmentation for frame-semantic role labeling. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 78–84, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shahid Iqbal Rai, Danilo Croce, and Roberto Basili. 2025. Injecting frame semantics into large language models via prompt-based fine-tuning. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 31–47, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Carlos Subirats Rüggeberg and Miriam R. L. Petruck. 2003. Surprise: Spanish framenet! *International Congress of Linguists Workshop on Frame Semantics*, page 6.

Harish Tayyar Madabushi, Taylor Hudson, and Claire Bonial. 2025. Generative FrameNet: Scalable and adaptive frames for interpretable knowledge storage and retrieval for LLMs powered by LLMs. In *Proceedings of Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025*, pages 107–119, Abu Dhabi, UAE. ELRA and ICCL.

Ebubekir Tosun, Mehmet Emin Buldur, Özay Ezerceli, and Mahmoud ElHussieni. 2026. A hybrid protocol for large-scale semantic dataset generation in low-resource languages: The turkish semantic relations corpus. *Preprint*, arXiv:2601.13253.

Olcay Taner Yıldız, Begüm Avar, and Gökhan Ercan. 2019. An open, extendible, and fast Turkish morphological analyzer. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1364–1372, Varna, Bulgaria. INCOMA Ltd.

Liping You and Kaiying Liu. 2005. Building chinese framenet database. *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Dataset Visualizations

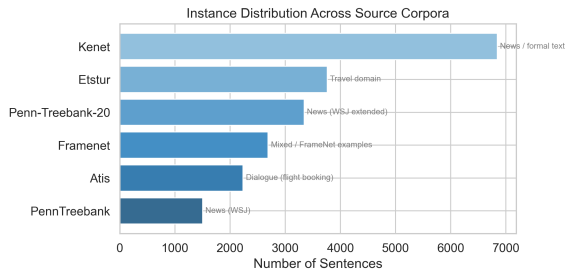

Figure 4: Instance distribution across the six source corpora.

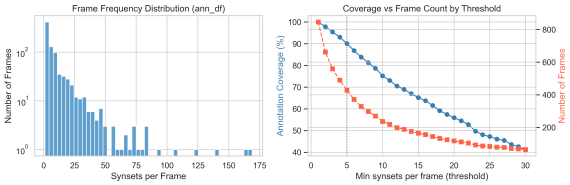

Figure 5: Frame frequency distribution in the annotated lexical list (left) and coverage vs. frame count by minimum-example threshold (right). Threshold$=5$ retains 428 frames, covering 90.1% of synsets.

B Prompt Templates

Setup A / B few-shot prompt (abbreviated).

```
You are a FrameNet expert for Turkish. Given a Turkish
verb expression (SYNSET) and its definition (DEFINITION),
select the single most appropriate FrameNet frame from
the candidate list.
Output ONLY valid JSON: {"predicted_frame": "...",
"confidence": "high|medium|low", "reasoning": "..."}
Candidate frames: <f_1>, ..., <f_30>
Examples:
SYNSET: <s_1>
DEFINITION: <d_1>
FRAME: <f_1>
[4 more examples]
SYNSET: <target>
DEFINITION: <target_def>
Answer:
```

Judge prompt (abbreviated).

```
You are a FrameNet expert evaluating frame assignment
quality for Turkish verb expressions.
Given: SYNSET, DEFINITION, ASSIGNED FRAME
Rate appropriateness 1-5:
1=Completely wrong  2=Mostly wrong
3=Partially correct 4=Mostly correct  5=Perfect
Output ONLY valid JSON:
{"score": <1-5>, "reasoning": "<one sentence>"}
```

C Retrieval Ablation

System	R@1	R@3	R@5	R@10	R@30
BM25 only	0.236	0.393	0.476	0.573	0.702
Cosine TR only	0.221	0.296	0.349	0.441	0.675
Hybrid (proposed)	0.236	0.392	0.475	0.571	**0.759**

Table 7: Retrieval recall ablation on the Setup A test set (N=1,673). Hybrid union consistently outperforms individual components.