Family Matters: Cross-Provider LLM-Judge Committees and Iterative Prompts for Turkish PropBank Argument-Frame Prediction

Anonymous ACL submission

Abstract

Large language models are increasingly used not only to label data but also to judge whether another model's labels are correct. Using Turkish PropBank argument-frame prediction as a testbed, we study two questions. First, does agreement between LLM judges reflect shared capability or shared provider family? Across four judges from two provider families (two Gemini, two OpenAI), within-family agreement consistently exceeds cross-family agreement. However, a two-judge committee combining one model from each family improves exact-match precision by 11.9 points over the unfiltered baseline, more than doubling the gain of the best single-judge filter, while adding a third same-family judge yields little benefit. Second, how do targeted prompt revisions affect other error types? A prompt edit for Turkish intransitive activity verbs improves the intended ARG0-only class by 21.7 points but also increases ARG2 omissions by 30.9 points; a follow-up clarification partially recovers overall performance. These results suggest that LLM-judge committees benefit more from family diversity than larger committee size, and that prompt revisions for closed semantic taxonomies should be evaluated by error class rather than aggregate metrics alone.

1 Introduction

PropBank-style annotation links each predicate sense to the semantic participants licensed by that sense (Palmer et al., 2005). These participants are represented with core roles such as ARG0, ARG1, and ARG2.[1] The framework is closely related to Frame Semantics (Fillmore, 1976; Baker et al., 1998) and has been widely studied in semantic role labeling (Gildea and Jurafsky, 2000; Carreras and Màrquez, 2005; He et al., 2017; Shi and Lin, 2019).

Turkish makes sense-level argument-frame prediction particularly challenging because evidence for semantic roles is distributed across lexical meaning, derivational morphology, and case marking. As an agglutinative language, Turkish encodes voice, valency, and case-related information through suffixation (Oflazer, 1993; Göksel and Kerslake, 2005; Marşan et al., 2021). As a result, surface syntax does not always transparently reveal the underlying argument frame. A verb may appear intransitive while still licensing both ARG0 and ARG1, and dative-marked complements may function as core arguments rather than adjuncts. These distinctions make Turkish PropBank annotation highly dependent on expert linguistic analysis.

Large language models (LLMs) offer a potential way to reduce this annotation burden, but they also introduce new methodological questions. Recent work shows that LLMs can predict frame-like semantic structures from definitions and examples (Tayyar Madabushi et al., 2025), especially when combined with retrieval-augmented prompting (Lewis et al., 2020; Brown et al., 2020). However, prompt revisions are often evaluated only through aggregate metrics, without examining which error classes improve or deteriorate. In parallel, LLMs are increasingly used as evaluators (Zheng et al., 2023; Wang et al., 2024), yet it remains unclear whether agreement among LLM judges reflects semantic reliability or shared provider-family biases.

This paper studies Turkish PropBank argument-frame prediction as a closed multi-label task. Given a predicate sense, definition, and optional example sentence, models predict which of ARG0, ARG1, and ARG2 are licensed. We evaluate five prediction models under zero-shot, static few-shot, and retrieval-based prompting on 500 held-out senses from a 17,692-sense gold standard, and add a targeted GPT-5-mini ablation to test whether a reasoning-enabled OpenAI model changes the

[1] We adopt the PropBank core-role inventory (ARG0–ARG4), but predict only ARG0, ARG1, and ARG2 because they cover more than 99.5% of valid frames in the gold standard.

family-level picture.

Our findings are threefold. First, LLM-judge agreement shows a clear but asymmetric provider-family effect: within-family agreement exceeds cross-family agreement, but Gemini judges are substantially more internally consistent than OpenAI judges. Second, two-judge committees that combine one OpenAI and one Gemini model outperform larger single-family committees, improving exact-match precision by 11.9 points over the unfiltered baseline. Third, targeted prompt edits redistribute errors rather than uniformly improving performance: guidance that fixes ARG0 errors for Turkish intransitive activity verbs also substantially increases ARG2 omissions. Together, these results suggest that LLM-judge committees benefit more from family diversity than committee size, and that prompt revisions for closed semantic taxonomies should be evaluated by error class rather than aggregate metrics alone.

2 Related Work

Turkish PropBank and Turkish NLP infrastructure. Turkish PropBank extends PropBank-style semantic role annotation to Turkish verb senses by recording which core roles each sense licenses. The resource builds on earlier Turkish NLP infrastructure, including morphological analysis (Oflazer, 1993; Yıldız et al., 2019) and Turkish Universal Dependencies treebanks (Şahin and Adalı, 2018; Tsarfaty et al., 2013). The sense inventory of Marşan et al. (2021) provides ARG0–ARG4 role descriptions in a synset-style organisation; our experiments use the resulting inventory of 17,692 senses with at least one filled core-role field.

The distinctions relevant to this task are standard in semantic role theory. Unergative intransitives typically license an agent-like ARG0, unaccusatives license an undergoer-like ARG1, and transitive or triadic predicates may additionally license ARG2 (Levin, 1993; Göksel and Kerslake, 2005). In Turkish, these distinctions are often expressed through morphology and case marking, making sense-level frame prediction linguistically challenging.

Semantic role labelling and generative frame prediction. Traditional semantic role labelling identifies argument spans in a sentence and assigns semantic roles using statistical or neural models (Gildea and Jurafsky, 2000; Carreras and Màrquez, 2005; He et al., 2017; Shi and Lin, 2019). Our setting differs in that we predict the set of roles licensed by a predicate sense rather than labeling spans in context. Tayyar Madabushi et al. (2025) show that LLMs can generate FrameNet-style frame assignments, motivating our use of generative models for closed multi-label PropBank prediction.

In-context learning and retrieval-augmented prompting. In-context learning enables LLMs to adapt from instructions and examples without parameter updates (Brown et al., 2020), while retrieval methods improve prompting by selecting relevant demonstrations at test time (Lewis et al., 2020). Prior work shows that demonstration selection strongly affects performance (Liu et al., 2022; Rubin et al., 2022; Min et al., 2022). We therefore compare static and retrieved few-shot prompting using BM25 (Robertson and Zaragoza, 2009), dense multilingual sentence representations (Reimers and Gurevych, 2019; Karpukhin et al., 2020; Devlin et al., 2019; Conneau et al., 2020), and Maximal Marginal Relevance for diversity (Carbonell and Goldstein, 1998).

LLM-as-a-Judge and evaluator bias. LLM-based evaluation has become increasingly common (Zheng et al., 2023; Wang et al., 2024), but prior work shows that LLM judges are sensitive to factors such as output length, verbosity, and presentation order (Liu et al., 2023; Chiang and Lee, 2023; Dubois et al., 2023). These concerns are especially relevant for semantic role prediction, where outputs may be nearly correct despite differing by a single role. We therefore examine not only agreement with gold labels, but also whether judges from the same provider family behave more similarly than judges from different families.

Evaluation methodology. Because each predicate sense may license multiple roles, the task is a multi-label classification problem (Tsoumakas and Katakis, 2007). We report exact match together with macro- and micro-F1. For judge reliability, we use quadratic-weighted Cohen's κ (Cohen, 1960; Landis and Koch, 1977). Prompt comparisons are evaluated with paired bootstrap confidence intervals (Efron and Tibshirani, 1993; Dror et al., 2018), which estimate the stability of differences under resampling of matched evaluation items.

3 Methodology

3.1 Prompting Setups

We compare three prompting conditions that differ in how much task guidance is provided to the model. The zero-shot setup includes only the role definitions and decision rules, testing whether the model can apply the taxonomy without examples. The static few-shot setup adds six manually selected examples covering the main linguistic classes. One example is intentionally contrastive: in an eating event, a fork is treated as an adjunct rather than ARG2 because it is an instrument, not a participant licensed by the verb sense itself. The retrieved few-shot setup instead selects five examples separately for each test sense using a hybrid of lexical retrieval and embedding similarity, with Maximal Marginal Relevance used to reduce redundancy among the retrieved examples.

- **Zero-shot (ZS).** The model receives only the role taxonomy and prediction instructions.
- **Static few-shot (STATIC_FS).** The prompt includes six fixed examples representing major argument-frame patterns and one contrastive adjunct case.
- **Retrieved few-shot (RETR_MMR).** The prompt includes five examples retrieved dynamically for each target sense using a BM25+cosine hybrid with MMR diversification (k=5).

Output format. All prompts require structured JSON output containing the predicted argument labels, role descriptions, confidence score, and a short explanation. A tolerant parser removes Markdown code fences when models include them.

3.2 Multi-Tour Prompt Iteration

Prompts are revised over three tours, with each revision targeting errors observed in the previous round.

Tour 1 (t1). The baseline prompt defines the five argument-frame classes and provides a lexical test for ARG2, asking whether a third participant is required by the verb sense and answers a role-specific question such as "to whom?" or "into what state?".

Tour 2 (t2). Pilot analysis showed that Gemini models frequently misclassified ARG0-only intransitive activity verbs as ARG1-only. We therefore added explicit guidance that animate, agent-like participants in sound, behaviour, and expression verbs should generally be treated as ARG0. This tour also introduced a compact explanation of Turkish derivational morphology, including causative, passive, reciprocal, and reflexive markers that often affect argument structure.

Tour 3 (t3). Although t2 improved ARG0-only prediction, it also increased ARG2 omissions. Inspection suggested that models were interpreting the morphology cues too narrowly, as if ARG2 required explicit derivational marking. The third tour therefore clarified that morphology can provide evidence for ARG2 but is not required: some verbs, including "explain", "give", and certain light-verb constructions, license ARG2 without overt derivational morphology. Full prompt additions are given in Appendix B.

3.3 LLM-as-a-Judge Protocol

Each judge receives the predicate sense, definition, optional example sentence, and predicted frame, but not the gold frame. Judges rate the prediction on a five-point ordered scale, where 5 indicates a correct frame and 1 indicates no meaningful match. Outputs are returned as JSON containing a numeric `score` and brief `reasoning`.

We evaluate four judges from two provider families:

- `gpt-4o-mini` (OpenAI)
- `gpt-5-mini` (OpenAI, reasoning-enabled)
- `gemini-2.5-pro` (Google)
- `gemini-3-flash-preview` (Google)

`gpt-5-mini` is used both as a prediction model and as a judge, allowing comparison between within-family and cross-family agreement.

Judge reliability is measured with quadratic-weighted Cohen's κ (Cohen, 1960; Landis and Koch, 1977) on the 496 items for which all judges produced valid scores. We also compare single-judge filtering with committee rules: union accepts an item if any judge assigns the target score, majority requires at least two judges, and unanimous requires agreement from all judges in the committee. Finally, we examine score spread between judges as a possible signal for human review.

4 Experiments

We evaluate six LLMs from three provider families (Table 1). The Google Gemini models (Team et al., 2025) represent a recent proprietary API family. The OpenAI models include GPT-4o-mini and GPT-5-mini (reasoning-enabled), following instruction-tuned model lines (Ouyang et al., 2022). We also include Qwen2.5-7B-Instruct as an open-weights baseline (Qwen et al., 2025), similar in spirit to other open releases such as LLaMA (Touvron et al., 2023).

Provider	Model	Access
Google	`gemini-2.5-flash`	API
Google	`gemini-2.5-pro`	API
Google	`gemini-3-pro-preview`	API
OpenAI	`gpt-4o-mini`	API
OpenAI	`gpt-5-mini` (reasoning)	API
Alibaba	`qwen2.5:7b-instruct`	local

Table 1: Models evaluated.

Decoding. We use deterministic decoding when supported (temperature 0.0, seed 42). For models with controllable reasoning, we use the lowest available reasoning budget unless stated otherwise. Outputs that fail JSON parsing are treated as empty predictions, which penalizes recall rather than being discarded. We analyze GPT-5-mini separately in Section 5.3 because its default reasoning settings occasionally produce structured-output failures.

Sample size. We run a 50-item pilot for the first prompt tour and full evaluations on 500 held-out senses for the second and third tours. The pilot includes nine Gemini runs to enable matched comparisons between prompt versions. The final tour uses the retrieved few-shot setup for all main models, with an additional ablation over prompting setups for GPT-5-mini.

5 Results

5.1 Frame Prediction across Models and Setups

Table 2 reports exact match, macro-F1, and micro-F1 for the second prompt tour, where exact match requires the full predicted role set to match the gold set. Across all models, retrieved few-shot prompting is the strongest setup, indicating that semantically similar annotated senses provide useful grounding for both lexical choice and argument structure. The best overall result is obtained by Gemini 2.5 Flash with retrieval (EM = 0.628). Figure 1 confirms this ranking across all metrics. The smaller Qwen model is consistently behind the API models, especially on cases involving ARG2, which require more fine-grained sense distinctions.

Averaging over non-reasoning models, retrieval improves exact match (0.591) over static few-shot (0.552) and zero-shot (0.542). Gains are more stable at the full evaluation size than in the pilot, where retrieval and zero-shot were nearly tied. Error analysis shows that remaining mistakes are structurally plausible rather than random, as shown in Figure 2: most errors collapse three-role frames into $\{ARG0, ARG1\}$ or confuse unergative ARG0-only cases with two-role frames.

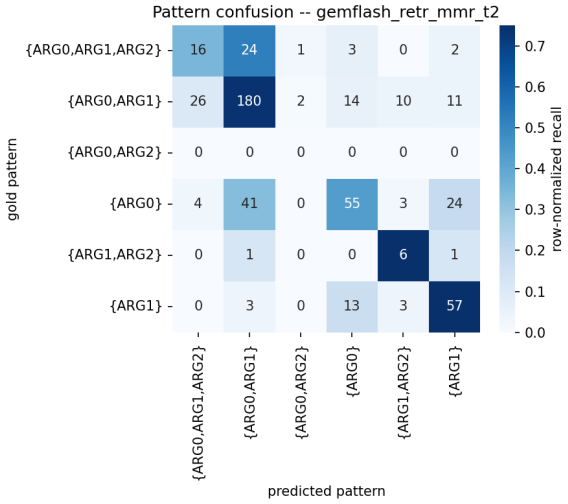

Figure 2: Pattern confusion for the best t2 prediction run (`gemini-2.5-flash` with retrieved few-shot prompting).

5.2 Prompt Iteration across Three Tours

We next analyze how prompt revisions change error behavior. For tours 1 and 2, we compare nine matched Gemini configurations (same model and setup before/after revision). While tour 1 uses a smaller pilot sample, the direction of changes is highly consistent across runs.

Effect of targeted edits. The t2 revision successfully fixes its intended issue: ARG0-only accuracy increases by 21.7 points across all runs, with consistent directionality and a confidence interval excluding zero.

Side effects. However, the same change redistributes errors elsewhere in the label space. Performance drops for multi-argument patterns, and

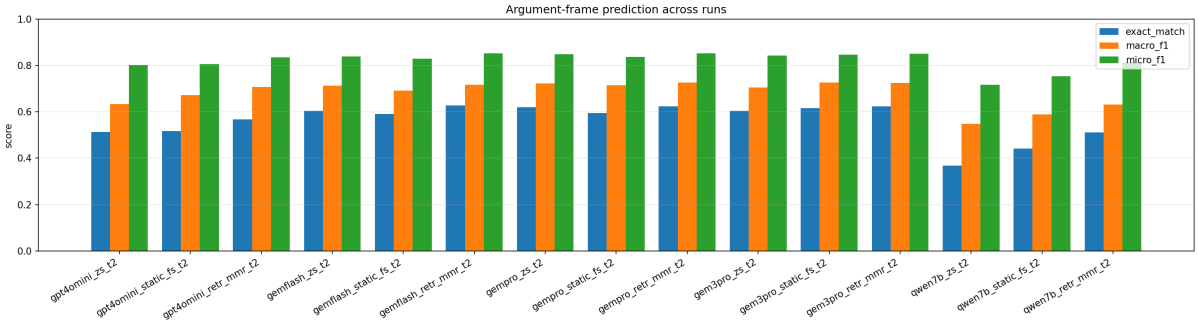

Figure 1: Main t2 benchmark across models and prompting setups. Retrieved few-shot prompting gives the strongest run within every model family, while the Qwen-7B runs lag most clearly on the strict exact-match metric.

Model	Setup	EM	macro-F1	micro-F1	ARG0 leak	ARG2 miss
gemini-2.5-flash	retr_mmr	**0.628**	0.717	0.853	0.202	0.574
gemini-2.5-pro	retr_mmr	0.624	0.726	0.853	0.083	0.481
gemini-3-pro-preview	retr_mmr	0.624	0.724	0.851	0.119	0.537
gemini-2.5-pro	zs	0.620	0.722	0.849	0.131	0.500
gemini-3-pro-preview	static_fs	0.616	0.726	0.847	0.107	0.444
gemini-2.5-flash	zs	0.604	0.713	0.838	0.214	0.500
gemini-3-pro-preview	zs	0.604	0.705	0.843	0.143	0.574
gemini-2.5-pro	static_fs	0.594	0.715	0.838	0.119	0.444
gemini-2.5-flash	static_fs	0.590	0.692	0.829	0.190	0.556
gpt-4o-mini	retr_mmr	0.568	0.707	0.835	0.119	0.556
gpt-4o-mini	static_fs	0.518	0.673	0.805	0.107	0.574
gpt-4o-mini	zs	0.514	0.633	0.802	0.119	0.778
qwen2.5-7b-instruct	retr_mmr	0.512	0.632	0.814	0.179	0.833
qwen2.5-7b-instruct	static_fs	0.442	0.589	0.754	0.119	0.796
qwen2.5-7b-instruct	zs	0.368	0.547	0.717	0.060	0.833

Table 2: Argument-frame prediction at t2, n=500. Sorted by EM.

ARG2 misses increase by 30.9 points. This indicates that correcting one linguistic distinction can shift decision boundaries for others in a closed label space.

Tour 3 correction. Tour 3 addresses this regression by clarifying that morphological cues are *sufficient but not necessary* for ARG2 licensing. As shown in Table 4, all Gemini models improve (up to +2.4 EM), while GPT-4o-mini degrades slightly and Qwen remains mostly unchanged. This suggests that the correction helps primarily within the Gemini family.

Overall, prompt improvements are not monotonic: they tend to shift errors between related classes rather than uniformly improving all metrics.

5.3 Reasoning-Model Ablation: gpt-5-mini

To extend OpenAI-family coverage, we evaluate `gpt-5-mini` under the t3 prompt. It improves over GPT-4o-mini in all three prompting setups, with exact-match gains of 1.2–3.8 points and substantially higher ARG0-only accuracy. However, the ablation also exposed a structured-output issue: with the default token budget and reasoning mode, 15–20% of outputs fail JSON parsing because hidden reasoning can consume the completion budget before the required JSON is emitted. Setting `reasoning_effort` to minimal and increasing `max_completion_tokens` eliminates parse failures and makes decoding roughly three times faster, so reasoning-model results should be reported only after explicit output-budget calibration. The full ablation table is provided in Appendix A.

5.4 Judge Reliability and Provider-Family Clustering

We apply four judges to the 500 predictions from the best second-tour run (Gemini 2.5 Flash with retrieved few-shot prompting). Table 5 and Figure 3 report pairwise quadratic-weighted Cohen's κ on the 496 cases with valid scores from all judges.

Asymmetric cohesion. The effect is asymmetric across families. Gemini judges are much more internally consistent (Gemini–Gemini $\kappa = 0.701$ vs OpenAI–OpenAI $\kappa = 0.376$, about $1.9\times$ higher).

Pattern	t1 mean	t2 mean	$\Delta_{t_2-t_1}$	95% CI (bootstrap)	sign consistency
{ARG0} unergative (TARGETED)	0.157	0.375	$+0.217$	$[+0.166, +0.261]$	9/9 ↑
{ARG1} unaccusative	0.833	0.775	-0.058	$[-0.104, -0.009]$	6/9 ↓
{ARG0,ARG1} agent–target	0.815	0.716	-0.099	$[-0.128, -0.078]$	9/9 ↓
{ARG1,ARG2} passive/recip.	1.000	0.694	-0.306	$[-0.375, -0.236]$	9/9 ↓
{ARG0,ARG1,ARG2} triadic	0.644	0.430	-0.215	$[-0.318, -0.104]$	8/9 ↓
ARG2 miss rate	0.204	0.512	$+0.309$	$[+0.239, +0.377]$	9/9 worse
ARG0 leakage rate	0.124	0.146	$+0.022$	$[-0.027, +0.076]$	5/9 worse

Table 3: Paired per-pattern and diagnostic-rate deltas, t1→t2, across the nine Gemini runs (3 models × 3 setups). Bootstrap 5000-resample 95% CIs. Bold rows are statistically significant regressions or gains (CI excludes 0).

Model	EM t2	EM t3	Δ
gemini-2.5-flash	0.628	**0.652**	**+0.024**
gemini-3-pro-preview	0.624	0.634	+0.010
gemini-2.5-pro	0.624	0.628	+0.004
qwen2.5-7b-instruct	0.512	0.514	+0.002
gpt-4o-mini	0.568	0.554	−0.014

Table 4: Paired t2→t3 EM deltas on the RETR_MMR setup (n=500 each). All three Gemini variants gain; gpt-4o-mini regresses; Qwen-7B is flat on aggregate EM but loses -0.087 on the {ARG0,ARG1,ARG2} triadic pattern. The Gemini-vs-OpenAI signature parallels the inter-judge κ stratification of Table 5.

Judge 1	Judge 2	κ_{quad}
Within-family pairs		
gemini-2.5-pro	**gemini-3-flash-preview**	**0.701**
gpt-4o-mini	**gpt-5-mini**	**0.376**
Cross-family pairs		
gemini-3-flash-preview	gpt-5-mini	0.351
gemini-2.5-pro	gpt-5-mini	0.303
gemini-3-flash-preview	gpt-4o-mini	0.266
gemini-2.5-pro	gpt-4o-mini	0.239

Table 5: Pairwise quadratic-weighted Cohen's κ across four judges (n=496). The Gemini–Gemini pair reaches "substantial" agreement (Landis and Koch, 1977); both within-family OpenAI and all cross-family pairs sit in the "fair" band (0.21–0.40), but within-family OpenAI (0.376) is consistently above the cross-family range (0.24–0.35). Families therefore differ in internal cohesion: Gemini is about 1.9× as internally consistent as OpenAI under this metric.

Notably, the OpenAI pair is closer to the strongest cross-family agreement (0.35) than to the Gemini pair. This is partly explained by score distributions: `gpt-4o-mini` is the strictest judge ($\bar{s} = 3.53$), Gemini judges are more lenient ($\bar{s} = 4.47, 4.49$), and `gpt-5-mini` is intermediate ($\bar{s} = 4.03$), yielding correspondingly intermediate agreement patterns.

Implication. Judge agreement is not a single scalar property of "provider family." High within-family agreement can reflect strong calibration consistency, while cross-family disagreement can reflect scale differences rather than semantic disagreement. For evaluation, adding a second judge from the same provider may add little new information, whereas cross-family pairing provides a more independent signal at similar cost.

5.5 Committee-Filtered Precision

We next evaluate whether LLM judges can identify a high-precision subset of predictions. Table 6 reports the main coverage–precision frontier; Appendix Figure 5 visualizes the same trade-off. The committee analysis uses the original three-judge panel (`gpt-4o-mini`, `gemini-2.5-pro`, `gemini-3-flash-preview`); the fourth judge (`gpt-5-mini`) is reserved for Section 5.4 to isolate within-family OpenAI agreement and is excluded here to keep the comparison consistent with earlier results.

Committee filtering. The strongest filter requires both GPT-4o-mini and Gemini 2.5 Pro to assign the maximum score. This retains 41.3% of items while improving exact match by 11.9 points over the full set, more than twice the gain of the best single-judge filter. Adding the third same-family Gemini judge yields only a 0.3-point change, indicating that cross-family diversity is more important than committee size.

Score spread as a diagnostic. Judge disagreement is also informative, but not monotonically so. Figure 4 reports exact match by three-judge score spread. Flagging cases with a spread of at least two captures roughly one third of the data but selects a lower-accuracy subset overall. Large disagreements often correspond to genuinely ambiguous or polysemous predicates rather than clear errors, while moderate disagreement is a better indicator

Filter strategy	N	Coverage	EM on subset	ΔEM vs. full
Full set (no filter)	496	100.0%	0.627	—
Single-judge filters				
gpt-4o-mini ≥ 4	277	55.8%	0.682	+0.055
gemini-2.5-pro ≥ 4	394	79.4%	0.673	+0.046
Two-judge committees (cross-family)				
Either gpt or pro ≥ 4 (union)	419	84.5%	0.659	+0.032
Both gpt and pro ≥ 4 (intersection)	252	50.8%	0.706	+0.079
Both gpt and pro $= 5$ (unanimous high)	205	41.3%	**0.746**	**+0.119**
Three-judge committees				
≥ 2 of 3 say ≥ 4 (majority)	389	78.4%	0.663	+0.036
Unanimous ≥ 4	244	49.2%	0.709	+0.082
Unanimous $= 5$	199	40.1%	**0.749**	**+0.122**
Spread-based diagnostic				
Spread ≥ 2 (route to human)	161	32.5%	0.596	−0.031

Table 6: Coverage–precision frontier for filter strategies. Bootstrap 95% CI for the (unanimous $= 5$ EM $-$ full EM) lift: $[+0.046, +0.193]$.

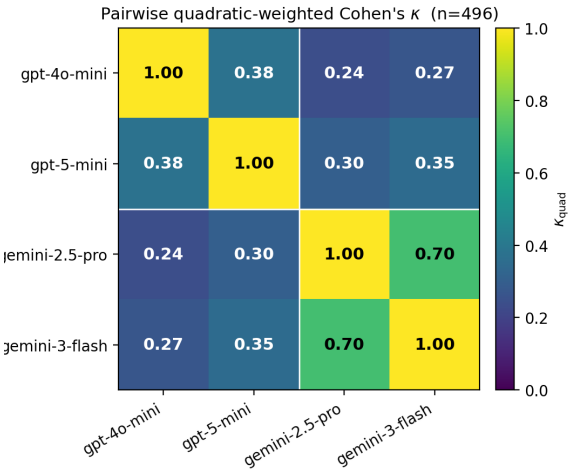

Figure 3: Four-judge agreement matrix. The two within-family blocks (OpenAI top-left, Gemini bottom-right) are both denser than the off-diagonal cross-family block, but the two within-family blocks differ in magnitude: Gemini–Gemini is 0.70, OpenAI–OpenAI is 0.38. The asymmetry visualises "provider-family clustering" as a real but non-symmetric effect rather than a uniform property.

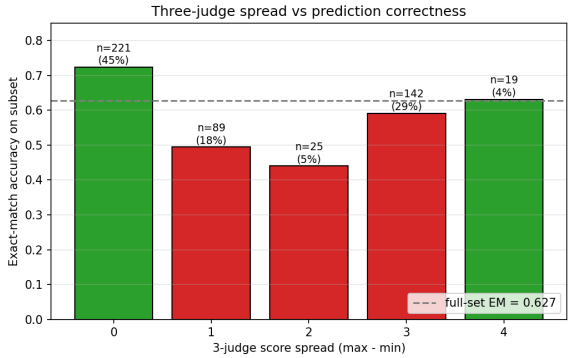

Figure 4: Exact-match accuracy by three-judge score spread. Moderate spread is the clearest warning signal; maximum spread is rare and not itself the lowest-accuracy bucket.

of likely mistakes.

5.6 Qualitative Case Analysis

Representative prediction and judge outputs are provided in Appendix Table 9. The examples illustrate the main mechanisms behind the aggregate results: t2 repairs ARG0-only errors, ARG2 is under-generated after the morphology-focused revision, t3 recovers some lexically licensed ARG2 cases, and same-family judges can accept the same wrong frame.

6 Discussion

Why does the unergative guidance help? ARG0-only senses are challenging because a single participant must be recognized as agent-like rather than defaulting to ARG1. In Turkish, additional material is often expressed as an oblique phrase (e.g., constructions similar to "laugh at X" or "cry for X"), which models may incorrectly treat as a core ARG1 argument. The revised prompt addresses this by introducing a simple diagnostic: check whether the verb sense can take a true accusative-marked theme. If not, the sole animate participant is more likely to be ARG0. This concrete test is easier to apply reliably than an abstract definition of unergativity.

Why does the morphology guidance hurt ARG2? The morphology block is linguistically

correct, but it changes how the model weights evidence. Listing suffixes that often affect valency appears to encourage the model to treat morphology as a near-exhaustive signal for argument structure. This is harmful for ARG2, since some predicates license a third role lexically rather than through overt derivational morphology. For example, verbs like *açıklamak* 'explain' can express a three-participant relation (someone explains something to someone), and constructions such as *farklı olmak* 'be different from' also introduce a third participant without the listed morphological markers. The third-tour clarification corrects this by stating that morphology is evidence for ARG2 but not a requirement.

Provider family vs. scale. The judge results show that large API models are not interchangeable, but the structure is more nuanced than a single "family clustering" effect. Within-family agreement is always higher than cross-family agreement, yet internal cohesion differs substantially across families: Gemini judges agree strongly ($\kappa = 0.70$), while OpenAI judges are much less consistent ($\kappa = 0.38$). Within OpenAI, `gpt-5-mini` behaves as an intermediate scorer, with mean score and pairwise κ values between `gpt-4o-mini` and the Gemini models, suggesting that reasoning-tuned successors can shift scoring scales within the same provider ecosystem. The practical implication remains that cross-family judge pairs are preferable, but the benefit of adding a second same-family judge depends on how internally calibrated that family is.

When to consult humans. The most useful trigger for human review is not low confidence or any form of judge disagreement. Large disagreements often reflect stable differences between judge families on genuinely ambiguous senses. Instead, moderate disagreement is more informative, as it tends to identify cases near the decision boundary. For annotation support systems, this suggests that human effort should be focused on near-miss cases rather than all instances where judges differ due to scale or calibration effects.

Why this matters beyond Turkish. Our dataset is Turkish, but the two main findings are not Turkish specific. Any closed semantic taxonomy that depends on linguistically subtle distinctions — predicate–argument structure in other morphologically rich languages, fine-grained relation extraction, sense disambiguation, frame assignment — should expect prompt revisions to redistribute rather than purely add accuracy. The unergative-versus-unaccusative split that hurts us in Turkish has direct analogues in, for example, Hungarian, Finnish, or Basque, where the same single-participant verb can be agent-like or undergoer-like depending on the sense. Likewise, any pipeline that uses LLM judges as a quality gate should expect intra-family judge pairs to share calibration biases that an inter-family pair would expose; this is a property of how the judges were trained, not of which language they are judging. The practical recommendation — pair across families before adding within a family — therefore travels with the methodology, not with the dataset.

7 Conclusion

We presented an iterative prompt-engineering and LLM-judge committee pipeline for Turkish PropBank argument-frame prediction. Three findings extend prior work: (i) within-family inter-judge agreement consistently exceeds cross-family agreement, but families differ markedly in internal cohesion ($\kappa = 0.70$ for two Gemini judges vs. $\kappa = 0.38$ for two OpenAI judges across four judges and 496 paired predictions), so provider-family effects are real but not symmetric; (ii) a two-judge cross-family committee matches three-judge unanimous-5 precision at higher coverage, supporting committee design by family diversity over committee size; and (iii) targeted prompt blocks redistribute rather than purely add accuracy in closed-set taxonomies, motivating systematic paired evaluation of prompt revisions.

For practitioners, the most actionable result is the committee design. If an LLM judge is used as a quality gate, pair two judges from different providers and treat a third same-family judge as redundant. The compute saved on that third judge is better spent either on routing uncertain cases to a human annotator or on adding a fourth model from a yet different ecosystem; in our experiments, the marginal precision gained by a same-family addition was within rounding error of zero.

Future work will integrate LoRA fine-tuning, extend judge benchmarking to local models, and apply the spread-based diagnostic in a deployed Turkish PropBank annotation interface.

8 Limitations

Sample-size confound in paired t1→t2. The pilot t1 evaluation uses n=50 whereas t2/t3 use n=500; the paired delta therefore mixes prompt change and sample change. We mitigate this by reporting directional consistency (9/9 across patterns) and bootstrap CIs but cannot fully isolate the prompt effect without an additional n=500 run under the t1 prompt.

Single fine-tuned baseline. We do not parameter-efficiently fine-tune (e.g., via LoRA (Hu et al., 2021)) a local model on PropBank. Whether domain adaptation of a small open model would outperform the API baselines on multi-label PropBank set prediction — as has been reported for single-label frame tasks in related work — is an open question.

Judge coverage. Four judges are evaluated, all API-scale, drawn from two provider families (two Gemini, two OpenAI). Other API-scale judges (e.g., Anthropic Claude, Mistral) and local-model judges (Qwen, Llama, Gemma) are not tested here for the PropBank setting, so we do not know whether a third provider family or open-weights judges would behave more like the internally cohesive Gemini family, the more heterogeneous OpenAI family, or something else entirely.

Provider-family confounding. The judge comparison is observational. Provider family, training data, alignment procedure, score calibration, and model size are not independently controlled. We therefore interpret family clustering as a practical committee-design result, not as causal proof that provider identity alone determines judge behaviour.

Closed three-label set. Our task does not exercise ARG3/ARG4 or thematic role labels; results should not be interpreted as transferring to full PropBank role labelling.

t3 evaluated on a single setup only. The t3 disclaimer was evaluated across all five models in the retrieved-few-shot setup (the strongest t2 configuration), plus a three-setup gpt-5-mini ablation. We did not re-evaluate t3 on zero-shot or static few-shot for the original five models, so the model-family effect is established for retrieval-augmented prompting and remains to be verified for other setups.

9 Ethical Considerations

This work uses a lexical-semantic resource and does not introduce new human-subject data. The main ethical risk is automation bias: high agreement between LLM judges could be mistaken for correctness, causing incorrect frames to enter an annotation resource. Our results argue against fully automatic acceptance. The judge committee is useful as a precision-oriented filter, but ambiguous or disagreement-heavy cases should remain subject to expert review, especially because the resource supports downstream Turkish NLP systems. AI assistance was used for editing, and we avoid reporting individual annotator behaviour; all analyses are at the model-output and frame-pattern level.

References

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gözde Gül Şahin and Eşref Adalı. 2018. Annotation of semantic roles for the turkish proposition bank. *Lang. Resour. Eval.*, 52(3):673–706.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: a simulation framework for methods that learn from human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

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

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kemal Oflazer. 1993. Two-level description of Turkish morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

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.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *Preprint*, arXiv:1904.05255.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

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.

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 GPT-5-mini Ablation

Setup	EM	Ma-F1	ARG0-only	Triadic
zero-shot	0.552	0.663	0.472	0.261
static FS	0.530	0.681	0.457	0.413
retrieved FS	**0.584**	**0.719**	0.441	**0.457**

Table 7: GPT-5-mini under t3 with minimal reasoning and a 4096-token completion budget. All runs parse successfully; EM improves over GPT-4o-mini by 1.2–3.8 points, with much higher ARG0-only accuracy (mean 0.457 vs. 0.184).

Tour	Main instruction state	Intended target	Observed effect
t1	Baseline taxonomy with five frame patterns and a three-part lexical-frame test for ARG2.	Establish whether models can apply the closed PropBank role set without task-specific corrective rules.	Pilot runs exposed severe weakness on ARG0-only unergatives.
t2	Added an explicit unergative anti-pattern block and Turkish morphological cues for causative, passive, reciprocal, reflexive, and light-verb constructions.	Correct the frequent error where animate single-participant predicates were mapped to ARG1 instead of ARG0.	ARG0-only accuracy improved by 21.7 points, but ARG2 misses increased by 30.9 points.
t3	Added a sufficiency disclaimer: morphology can support ARG2 but is not required for ARG2.	Repair the t2 side effect where models treated the morphology list as an exhaustive ARG2 trigger list.	The best Gemini retrieved-few-shot run gained 2.4 EM points; effects did not transfer uniformly to OpenAI or Qwen.

Table 8: Prompt evolution across tours. The table separates the linguistic motivation from the observed empirical side effects.

B Prompt Templates

The three prompt tours are diffs over a common base. Table 8 summarizes what changed in each tour. Below we reproduce the t3 additions. Full t1/t2/t3 prompt files and evaluation scripts will be released in an anonymized repository.

Tour 2 — unergative anti-pattern (excerpt).

```
CRITICAL anti-pattern for unergatives
(high model error rate observed in
pilot):
- An animate participant performing
a sound, behaviour, emission, or
expressive act (anırmak, ağlaşmak,
havlamak) is UNERGATIVE — its sole
participant is ARG0, NEVER ARG1.
- Mental/emotional stative predicates
with olmak / +lAn- where the experiencer
is the intentional/animate locus are
unergative {ARG0}, NOT {ARG1}.
- Concrete test: can the verb take a
true accusative-marked direct theme (X-i
V-mek) within THIS sense?
```

Tour 3 — ARG2 sufficiency disclaimer.

```
Important: the morphological cues above
are SUFFICIENT but NOT NECESSARY for
ARG2.
- ARG2 can be licensed by class (v)
triadic causatives WITHOUT any of the
morphological markers above (açıklamak,
vermek, anlatmak).
- ARG2 can also be licensed in class
(iv) stative two-place relations with
light-verb constructions (farklı olmak,
tekelinde olmak, dayak yemek).
- Do not treat the absence of morphology
as evidence AGAINST ARG2.
```

Judge Prompt Template.

```
You are a Turkish PropBank expert
evaluating predicate argument-frame
assignment quality.
. . .
Rate the proposed frame on a 1-5 scale:
5 = perfect; 4 = mostly correct (one
role over/under); 3 = partially correct;
2 = mostly wrong; 1 = completely wrong.
Output ONLY JSON: {"score": int 1-5,
"reasoning": string}.
```

C Reproducibility

All API calls and judge calls are SHA-256-cached on (model, prompt, sampling parameters). The random seed is 42 for stratified split, sampling, and API decoding where supported. Total compute cost (t1 + t2 + t3 + judges) is approximately $25 in API spend plus ≈5 GPU-hours on an RTX 3000 Ada laptop GPU.

Repository contents. The release package contains the t1/t2/t3 zero-shot and few-shot prompt files, raw prediction CSVs, judge-score CSVs, aggregate metric CSVs, plotting scripts, and the figures used in this paper. The raw prediction files include the sense identifier, predicate, gold frame, predicted frame, parse status, model confidence, retrieved example identifiers when applicable, and the raw model response. The metric CSVs contain exact match, macro-F1, micro-F1, Hamming loss, per-label precision/recall/F1, per-label Cohen's κ, and the diagnostic ARG0/ARG2 over- and under-generation rates.

Dataset access. The experiments read from the Turkish PropBank spreadsheet and filter to verb-sense rows with at least one filled core-role field. If the underlying resource cannot be redistributed under the final venue policy, the released scripts will still reproduce preprocessing, prompt construction, parsing, metric computation, judge aggregation, and figure generation once the spreadsheet is placed at the expected path.

D Judge Committee Diagnostics

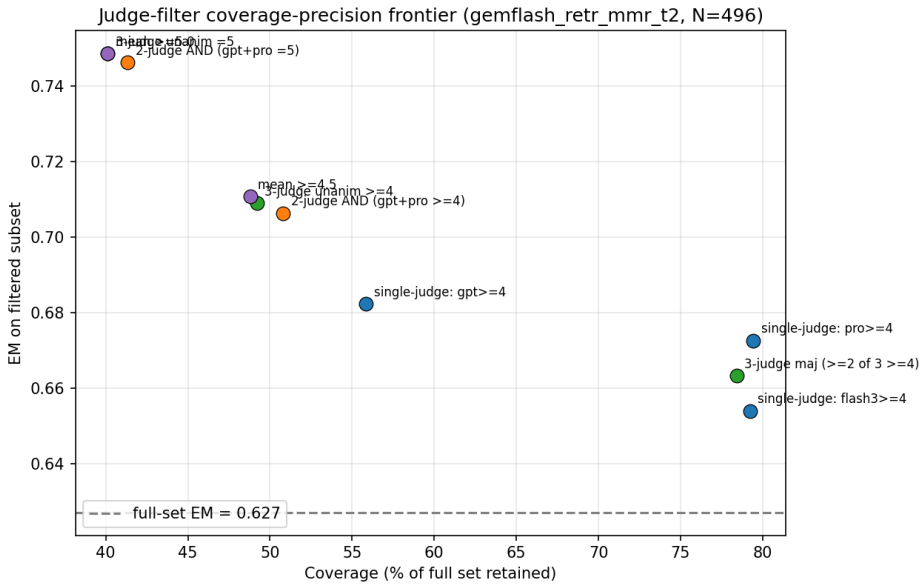

Figure 5: Coverage–precision frontier for judge filters. The best two-judge cross-family rule (GPT-4o-mini and Gemini 2.5 Pro both assign 5) is almost indistinguishable from the three-judge unanimous-5 rule, showing that the extra same-family Gemini judge adds coverage loss with negligible precision gain.

E Qualitative Examples

Phenomenon	Predicate	Gold frame	Model or judge output	Diagnostic value
Prompt repair	*gözü açılmak*	{ARG0}	t1: {ARG1}; t2: {ARG0}	The t2 unergative/stative guidance moves a single-participant sense from the dominant ARG1 error into the intended ARG0-only class.
ARG2 omission	*dikmek*	{ARG0,ARG1,ARG2}	t2: {ARG0,ARG1}	The model treats the instrument-like complement as an adjunct, a typical post-t2 error where ARG2 is under-generated.
t3 recovery	*bel bellemek*	{ARG0,ARG1,ARG2}	t2: {ARG0,ARG1}; t3: {ARG0,ARG1,ARG2}	The t3 reminder that lexical frames can license ARG2 without derivational morphology lets the model recover an intrinsic instrument role.
Judge disagreement	*açılmak*	{ARG0,ARG1}	GPT judge: 2; Gemini judges: 5/5 for {ARG0,ARG2}	The prediction is wrong by exact match, but same-family Gemini judges both accept it, illustrating why cross-family review is more useful than adding another same-family judge.

Table 9: Representative examples from cached prediction and judge outputs. They illustrate the main error mechanisms behind the aggregate prompt-iteration and judge-committee results.