Benchmarking Source-Sensitive Reasoning in Turkish: Humans and LLMs under Evidential Trust Manipulation

Sercan Karakaş[1], **Yusuf Şimşek**[2]
[1]University of Chicago, skarakas@uchicago.edu
[2]Fırat University, ysimsek@firat.edu.tr

Abstract

This paper investigates whether *source trustworthiness* shapes Turkish evidential morphology and whether large language models (LLMs) track this sensitivity. We study the past-domain contrast between *-DI* and *-mIş* in controlled cloze contexts where the information source is overtly external, while only its perceived reliability is manipulated (High-Trust vs. Low-Trust). In a human production experiment, native speakers of Turkish show a robust trust effect: High-Trust contexts yield relatively more *-DI*, whereas Low-Trust contexts yield relatively more *-mIş*, with the pattern remaining stable across sensitivity analyses. We then evaluate 10 LLMs in three prompting paradigms (open gap-fill, explicit past-tense gap-fill, and forced-choice A/B selection). LLM behavior is highly model- and prompt-dependent: some models show weak or local trust-consistent shifts, but effects are generally unstable, often reversed, and frequently overshadowed by output-compliance problems and strong base-rate suffix preferences. The results provide new evidence for a trust-/commitment-based account of Turkish evidentiality and reveal a clear human–LLM gap in source-sensitive evidential reasoning.

Keywords: Turkish evidentiality, trustworthiness, LLMs, cloze task, human–LLM comparison, benchmarks

1. Introduction

Evidentiality refers to the linguistic encoding of information source, that is, how a speaker indicates the basis on which a proposition is presented (e.g., direct perception, inference, or report) (Willett, 1988; Dendale and Tasmowski, 2001; de Haan, 2001; Plungian, 2001; Aikhenvald, 2004; Boye, 2012; Ünal and Papafragou, 2020). In many languages, evidential markers systematically signal whether the speaker witnessed an event, inferred it from available evidence, or learned it from someone else. Because evidentiality lies at the interface of semantics, pragmatics, and discourse, it has been central to research on speaker commitment, reliability, and perspective in natural language.

A useful perspective on evidentiality comes from language acquisition and "thinking for speaking": as children learn a language, they learn to attend to distinctions that the language regularly encodes (Brown and Lenneberg, 1954; Slobin, 1996). In evidential languages, this can make knowledge source (e.g., direct experience, inference, report) especially salient early on, since speakers must repeatedly track and express it in discourse. For Turkish, this suggests that evidential morphology is not just grammatical, but part of a broader cognitive routine linking source monitoring, memory, and speaker stance (Aksu-Koç and Slobin, 1986). Turkish offers a particularly important case for evidentiality research because past-domain morphology is closely tied to distinctions in information source and epistemic stance (Aksu-Koç and Slobin, 1986; Underhill, 1976; Johanson, 2003; Göksel and Kerslake, 2005; Kornfilt, 1997). A well-known contrast is between the suffix *-DI*, traditionally associated with direct evidence or stronger speaker commitment, and the suffix *-mIş*, which is commonly associated with indirect evidence, such as inference or report (Aksu-Koç and Slobin, 1986; Underhill, 1976; Johanson, 2003; Gül, 2009; Şener, 2011; Göksel and Kerslake, 2005).

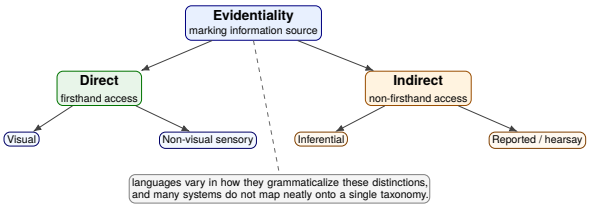

Figure 1: Information-source marking, adapted from typological overviews of evidential systems (Willett, 1988; Aikhenvald, 2004).

A minimal illustration is given in (1)–(2): in (1), the use of *-DI* is classically associated with a context in which the speaker directly witnessed the event, whereas in (2), the use of *-mIş* suggests that the speaker did not directly witness the event but instead inferred it or learned it from another source. At the same time, a substantial body of work notes that the actual distribution and interpretation of these forms are sensitive to discourse context, information structure, and speaker stance, so the traditional labels are best understood as useful descriptive generalizations rather than rigid one-to-one mappings between form and meaning

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 329–346
16 May 2026.

(Gül, 2009; Şener, 2011).

(1) Ahmet Ali'yi gör-dü.
Ahmet Ali-ACC see-PST.DIR
'Ahmet saw Ali (witnessed/direct evidence).'

(2) Ahmet Ali'yi gör-müş.
Ahmet Ali-ACC see-PST.INDIR
'Ahmet apparently/reportedly saw Ali (hearsay/inference).'

The *-DI*/*-mIş* contrast offers a well-studied case for investigating how languages encode information source, speaker stance, and pragmatic inference (Aksu-Koç and Slobin, 1986; Johanson, 2003; Göksel and Kerslake, 2005; Gül, 2009; Şener, 2011; Kandemirci et al., 2023; Ataman-Devrim et al., 2025), and it is also central to developmental work on how children learn to track and express knowledge source in discourse (Aksu-Koç and Slobin, 1986; Slobin, 1996). In addition, evidential distinctions are highly relevant for NLP tasks such as translation, summarization, and source-sensitive interpretation.

1.1. Why evidentiality matters for LLM research?

Evidentiality is also increasingly relevant to LLM research, because many of the core reliability problems in modern text generation are, in effect, problems about *source status*: whether a claim is grounded in direct input evidence, retrieved documents, parametric memory, inference, or unsupported generation. Recent work on factuality and hallucination has shown that fluent model outputs may be unfaithful to available evidence or false in world knowledge terms, motivating finer-grained analyses of how models represent and communicate the basis of their claims (Maynez et al., 2020; Lin et al., 2022; Manakul et al., 2023; Min et al., 2023; Bang et al., 2025; Liu et al., 2026). In parallel, work on calibration and self-evaluation suggests that models can sometimes track uncertainty or answerability, but this ability is uneven and sensitive to task format and available context (Kadavath et al., 2022). At the same time, frontier and open models are improving rapidly in reasoning, long-context processing, multilinguality, and agentic performance, as reflected in recent reports (Singh et al., 2025; Yang et al., 2025). These gains make source-sensitive evaluation more, not less, important: as models become stronger and more fluent, the central question is increasingly not only whether they can produce plausible language, but whether they can reliably represent the evidential basis of what they say.

From this perspective, evidentiality offers a linguistically grounded framework for studying how systems encode (or fail to encode) distinctions among *witnessed*, *inferred*, and *reported* information. This is especially important in settings that explicitly require provenance, such as retrieval-augmented generation and citation-based QA, where models must connect claims to supporting evidence and where attribution quality itself has become a major evaluation target (Lewis et al., 2020; Nakano et al., 2022; Gao et al., 2023; Li et al., 2024). More broadly, evidential categories provide a principled way to ask whether LLMs merely optimize surface plausibility or whether they track source-sensitive epistemic distinctions that are central to human communication. Besides, this line of work is also important because direct *human–LLM* comparisons for Turkish remain relatively scarce, especially for source-sensitive semantic/pragmatic phenomena such as evidentiality. Much of the recent Turkish LLM literature has focused on model development, benchmarking, and engineering issues (Bayram et al., 2025; Er et al., 2025; Isbarov et al., 2025; Umutlu et al., 2025; Wicaksono et al., 2025; Bayram et al., 2026; Karakaş and Şimşek, 2026; Toraman et al., 2026; Uğur et al., 2026), rather than tightly controlled comparisons between human judgments and model preferences on linguistically targeted contrasts. Emerging Turkish-focused human–LLM studies suggest that such comparisons can reveal substantial mismatches between surface-form preferences and human-like contextual reasoning in several linguistic domains (Keleş and Dinçtopal Deniz, 2024, 2025; Karakaş, 2026).

2. Related Work

Turkish evidentiality: traditional descriptive and grammatical accounts. Turkish has long occupied a central place in the evidentiality literature because its past-domain morphology is closely tied to distinctions in information source and speaker stance (Underhill, 1976; Aksu-Koç and Slobin, 1986; Johanson, 2003; Kornfilt, 1997; Göksel and Kerslake, 2005). A traditional descriptive contrast distinguishes *-DI* and *-mIş*: *-DI* is often introduced as the form associated with direct evidence (or stronger speaker commitment), while *-mIş* is associated with indirect evidence, especially inferential and reportative uses (Underhill, 1976; Johanson, 2003; Göksel and Kerslake, 2005; Aksu-Koç and Slobin, 1986). This contrast has also been influential in acquisition and discourse-based work, where Turkish evidential morphology is treated as a system that helps speakers track and communicate how information was obtained (Aksu-Koç and Slobin, 1986).

At the same time, the Turkish literature shows that the traditional mapping is not a simple one-to-one encoding. A central debate is whether *-DI* is a true direct evidential, or primarily a past

tense form whose "directness" and stronger commitment effects arise through discourse-pragmatic inference in canonical contexts (Gül, 2009; Şener, 2011). Likewise, analyses of *-mIş* emphasize sensitivity to evidential source, discourse configuration, and speaker perspective, rather than a single fixed meaning (Gül, 2009; Johanson, 2003). For this reason, recent work often treats *direct*/*indirect* as useful descriptive labels while allowing context-sensitive meaning and pragmatic strengthening.

For the present study, we adopt Giannakidou and Mari's framework on veridicality, nonveridicality, commitment, and evidential bias (Giannakidou, 1998, 2006; Giannakidou and Mari, 2016, 2018, 2021). This framework is useful because it separates notions often conflated in evidentiality research: (i) information source, (ii) speaker commitment, and (iii) the structure of the speaker's information state. It also motivates treating source trustworthiness as a factor shaping evidential interpretation and commitment (Boscaro et al., 2025).

Let $M = M_i(w, t)$ be speaker i's information state at world w and time t. A compact way to encode the relevant distinction is to contrast veridical and nonveridical states:

(3) $\textsc{Veridical}_M(p)$ iff all worlds in M are p-worlds: $\forall w \in M\,[p(w)]$.

(4) $\textsc{Nonveridical}_M(p)$ iff M contains both a p-world and a $\neg p$-world: $\exists w \in M\,[p(w)]$ and $\exists v \in M\,[\neg p(v)]$.

The formula in (3) states that the speaker's information state supports only p-worlds (full support for p), while the formula in (4) states that the information state remains compatible with both p and $\neg p$ possibilities. In this perspective, evidential and modal expressions can encode an evidence-based bias toward p without requiring full veridical commitment.

This distinction is useful for Turkish evidentiality because it allows us to ask not only which source type is invoked, but also whether that source is treated as strong enough to support a more veridical commitment profile. In our trust-based design, source trustworthiness is therefore modeled as a factor that can shift speakers between stronger commitment and weaker, evidence-biased interpretations, rather than as a purely descriptive source label.

3. Methods

To examine how Turkish evidential morphology responds to contextual *trustworthiness*, we used two complementary experimental paradigms with two corresponding datasets. First, we designed a controlled cloze-style task that can be administered to both human participants and LLMs. Second, we constructed a larger dataset for LLM-only evaluation in order to test generalization over a broader set of contexts. In all tasks, the target is a *cloze* (fill-in-the-blank) completion: participants or models infer and produce the missing final word of a sentence from the preceding context. This shared format provides a direct measure of context-sensitive morphological preferences in production and enables human–model comparison within a single task setup.

3.1. Dataset 1 for Human and LLM Comparison

Dataset 1 consists of 60 original items, partitioned into three categories: high-trust, low-trust, and filler sentences. All items were manually authored by the researchers. High-trust sources correspond to announcement channels that are generally considered reliable in everyday life, whereas low-trust sources typically consist of channels that are difficult to trust. These sources are provided in Appendix A. During construction, we aimed to make the target completion as uniquely and unambiguously recoverable as possible, thereby minimizing cases with multiple equally plausible completions, while also balancing potential confounds such as lexical frequency and sentence length across conditions. In the cloze task, participants were asked to complete the final blank with the most natural and contextually appropriate continuation.

(5) Belediyenin SMS uyarısına göre sular .....
'According to the municipality's SMS alert, the water .....'

(6) Yan binadaki teyzenin dediğine göre sular .....
'According to the lady next door, the water .....'

Across conditions, the source of information is kept *overtly external* through explicit attribution frames (e.g., *X'e göre* 'according to X'; cf. (5)–(6)). The manipulation targets the *perceived reliability* of that external source: High-Trust items present the proposition as coming from institutionally authoritative channels (e.g., an official municipal SMS notification), whereas Low-Trust items present it as coming from informal or weakly accountable channels (e.g., a statement by a neighbor). To verify that this manipulation isolates trust rather than idiosyncratic item properties, we ran a separate norming study with a broad participant pool. Participants consistently rated High-Trust sources as more credible than Low-Trust sources, and items that failed to show the intended separation were revised or removed.

3.2. Dataset 2 for LLM-Only Evaluation

Dataset 2 is an expanded item set containing all experimental items from Dataset 1 together with additional items, and was designed exclusively for large language model evaluation. It preserves the same structural properties and manipulation logic as Dataset 1, while increasing coverage and statistical stability by substantially expanding the number of scenarios and lexical realizations. The expanded set contains 200 items in total. Dataset 2 was *not* administered to human participants; instead, it was designed to test model behavior on a broader item base without increasing participant burden, and to assess whether patterns observed in the human-scale dataset replicate and generalize when the trust manipulation is instantiated across a wider range of contexts.

4. Experiments

4.1. Human Experiment

The human experiment received prior approval from the University of Chicago Institutional Review Board (IRB26-0198). All participants provided informed consent before participation. We recruited 75 unique participants who self-identified as native speakers of Turkish, were at least 18 years old, and were residing in Turkey at the time of participation. The experiment was implemented and administered online using the PCIbex/PennController platform (Zehr and Schwarz, 2018). The main study produced 4,500 trials in total: 1,500 High-Trust trials, 1,500 Low-Trust trials, and 1,500 filler trials. For each trial, we recorded the participant's completion response. Our primary analyses target the critical High-Trust vs. Low-Trust contrast, so we analyze 3,000 trials after excluding fillers. All responses were analyzed in de-identified form. This balanced design allows a controlled test of how the trust manipulation affects production preferences.

4.2. LLM Experiments

In this study, LLMs were tested using two different datasets and ten models (Team et al., 2025; neuralwork; NovusResearch; mradermacher; Yang et al., 2025; Trendyol; TURKCELL; Bezir et al., 2024; ocaklisemih; cenkersisman). and three prompts.The models evaluated in this study consist of Turkish-focused and multilingual large language models. gpt2-turkish-10M, WiroAI-turkish-LLM-8B, Turkcell-LLM-7B, Trendyol-LLM-7B, Gemma-2-9B-IT-TR, and Novus-7B-TR are models that have been either directly trained on Turkish or fine-tuned with Turkish data. In contrast, Gemma-3 and Qwen3-32B are trained on large-scale multilingual data and are capable of processing Turkish through their multilingual capacities. Some models, on the other hand, are built on a multilingual foundation and later fine-tuned with Turkish instruction data. This diversity allows us to compare the effects of Turkish-focused and multilingual training approaches on sensitivity to evidentiality morphology. During the study, the models were asked to fill in the blank format that was created (5)–(6). Three prompts were used to perform this task; these prompts are given in Appendix B. While the models were expected to generate the word in the space indicated by the space in the first two prompts, in the third prompt two different words were given as options, one of which is the word conjugated with "-DI" (7), and the other is the word conjugated with "-mIş" (8). In the third prompt, the models were expected to choose the most appropriate word between these two options according to the flow of the sentence. While generating the model outputs, the temperature was set to 0.1 and the top-p to 1; an additional max token setting of 8 was added to the third prompt.

(7) Getir-il-di
bring-PASS-PST.DIR
'was brought (-DI form)'

(8) Getir-il-miş
bring-PASS-PST.INDIR
'apparently/ reportedly was brought (-mIş form)'

5. Human experiment results

5.1. Production: trust robustly shifts *-DI* vs. *-mIş*

Table 1 shows strict-coding counts for critical trials (High+Low; fillers excluded). Descriptively, High-trust contexts produce more DI completions, whereas Low-trust contexts produce more mIş completions. The Other category is similar across conditions (High: 28.4%; Low: 29.5%), suggesting that the trust manipulation primarily redistributes responses *within* the evidential morphology space rather than increasing off-target productions.

Condition	DI	mIş	Other
High trust	**727** (48.5%)	347 (23.1%)	426 (28.4%)
Low trust	548 (36.5%)	**509** (33.9%)	443 (29.5%)

Table 1: Strict coding counts by condition (critical trials only; $n = 1500$ per condition). Percentages are within condition.

Figure 2 plots the *within-condition* proportions of DI vs. mIş after excluding Other. In the High-trust condition, DI $= 727/(727 + 347) = 0.677$ and mIş $= 0.323$. In the Low-trust condition, DI $= 548/(548 + 509) = 0.518$ and mIş $= 0.482$.

Two robustness checks yield the same qualitative pattern. (i) Under lenient last-token coding (al-

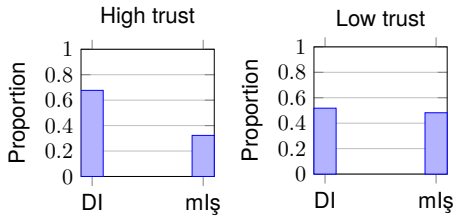

Figure 2: Human production (strict coding): within-condition proportions of *-DI* vs. *-mIş* among DI/mIş responses only.

lowing two-word VP completions), the association remains strong ($\chi^2(1) = 52.685$, $p = 3.92 \times 10^{-13}$; OR $= 1.859$, 95% CI $[1.571, 2.199]$). (ii) A content-controlled stratified analysis with matched content keys gives a pooled OR $= 2.029$, 95% CI $[1.654, 2.490]$ (CMH $\chi^2(1) = 46.789$, $p = 7.91 \times 10^{-12}$), with no strong evidence of heterogeneity across strata ($p = 0.137$).

6. The First Experiment with LLMs

Experiment I evaluates LLM behavior in a prompting-based gap-fill task using 200 trust-manipulated Turkish cloze items. Each item was presented to the model three times, and the final label was determined by majority vote across the three outputs. Table 2 summarizes the model-level results on the prompted task, reporting (i) the percentage of *usable* outputs (i.e., generations classifiable as DI or mIş), (ii) run-level DI/mIş counts in High- and Low-Trust conditions, (iii) the within-condition DI share, and (iv) Fisher exact tests for the High-vs.-Low trust contrast. A first important result is that the amount of usable evidential data varies substantially across models, ranging from 64.5% (Gemma-3-27B-IT) to 0.0% (Trendyol-LLM-7B). This means that, in the prompting setup, a large portion of outputs may be off-target and therefore cannot be interpreted as an evidential choice.

Among models with substantial usable DI/mIş outputs, the trust effect is not uniform. Gemma-3-27B-IT shows a clear and statistically reliable shift in the expected direction: the DI share decreases from 98.5% in High-Trust contexts to 88.9% in Low-Trust contexts (ΔDI $= +9.6$ pp, $p = 7.44 \times 10^{-5}$), indicating relatively more mIş in Low-Trust items. Gemma-2-9B-IT-TR shows a similar but smaller significant pattern (98.8% vs. 91.9%; ΔDI $= +6.9$ pp, $p = 0.006$). In contrast, Orbita-v0.1 shows a significant effect in the *opposite* direction, with a higher DI share in Low-Trust than in High-Trust contexts (81.5% vs. 67.5%; ΔDI $= -14.0$ pp, $p = 0.018$).

The remaining models mostly show null, unstable, or weak effects. Qwen3-32B, Novus-7B-TR, and WiroAI-TR-8B exhibit small and non-significant differences ($p = .371$, $.556$, and $.696$, respectively). GPT-OSS-20B-Astro shows a larger numerical reverse shift (ΔDI $= -20.4$ pp) but does not reach conventional significance ($p = .058$). For GPT2-TR-10M and Turkcell-LLM-7B, all usable outputs are DI, so no informative DI/mIş comparison can be made; Trendyol-LLM-7B yields no usable outputs. Overall, Experiment I suggests that in an open-generation prompting setup, trust-sensitive evidential behavior is highly model-dependent and often overshadowed by prompting/compliance limitations and a strong general bias toward DI.

7. The Second Experiment with LLMs

Experiment II evaluates LLM behavior in an *explicit past-tense* generation setup using the same 200 trust-manipulated Turkish cloze items. Unlike Experiment I, the prompt here explicitly instructed models to produce a *past-tense* completion. As in Experiment I, each item was sampled three times. To obtain a single interpretable evidential label per item, we assigned a final item-level label by majority vote across the three outputs (`tur1-tur3`) among outputs classifiable as DI or mIş (ties excluded).

Table 3 summarizes the model-level results for the explicit-past prompting condition, reporting (i) the percentage of *usable* items (i.e., items receiving a majority label classifiable as DI or mIş), (ii) item-level DI/mIş counts in High- and Low-Trust conditions, (iii) the within-condition DI share, and (iv) Fisher exact tests for the High-vs.-Low trust contrast. A first key result is that usability remains strongly model-dependent even with explicit past-tense instructions, ranging from 96.0% (Gemma-3-27B-IT) and 91.0% (Gemma-2-9B-IT-TR) to 0.0% (Trendyol-LLM-7B). Thus, explicit past-tense prompting substantially improves the proportion of usable evidential outputs for some models, but not for all.

Among models with substantial usable DI/mIş outputs, Gemma-3-27B-IT shows the clearest and statistically reliable trust-sensitive pattern in the predicted direction. Its DI share decreases from 97.9% in High-Trust contexts to 84.5% in Low-Trust contexts (ΔDI $= +13.4$ pp, $p = 1.51 \times 10^{-3}$), while the number of mIş labels increases from 2 items (High) to 15 items (Low). This indicates a relative increase in mIş under Low-Trust contexts. Gemma-2-9B-IT-TR also shows high usability and a small numerical shift in the expected direction (100.0% vs. 97.8%; ΔDI $= +2.2$ pp), but the difference is not statistically significant ($p = .243$). Orbita-v0.1 shows a larger numerical shift in the expected direction (ΔDI $= +18.1$ pp), but its low usability (22.0%) and non-significant test ($p = .327$) limit the strength of this result.

The remaining models mostly show null, unsta-

Model	Usable	H-DI	H-MIŞ	L-DI	L-MIŞ	DI% (H)	DI% (L)	ΔDI (pp)	p
Gemma-3-27B-IT	64.5	195	3	168	21	98.5	88.9	+9.6	7.44×10^{-5}
Gemma-2-9B-IT-TR	47.2	158	2	113	10	98.8	91.9	+6.9	0.006
Orbita-v0.1	40.3	83	40	97	22	67.5	81.5	-14.0	0.018
Qwen3-32B	24.7	54	3	81	10	94.7	89.0	+5.7	0.371
Novus-7B-TR	21.3	79	1	46	2	98.8	95.8	+2.9	0.556
WiroAI-TR-8B	16.8	54	5	40	2	91.5	95.2	-3.7	0.696
GPT-OSS-20B-Astro	11.3	29	11	26	2	72.5	92.9	-20.4	0.058
GPT2-TR-10M	2.8	10	0	7	0	100.0	100.0	+0.0	–
Turkcell-LLM-7B	1.2	5	0	2	0	100.0	100.0	+0.0	–
Trendyol-LLM-7B	0.0	0	0	0	0	–	–	–	–

Table 2: Experiment I (prompting-based gap-fill) LLM results. **Usable** = percentage of all generations classifiable as DI or MIŞ (i.e., DI/MIŞ%). DI% is computed as $\mathrm{DI}/(\mathrm{DI}+\mathrm{MI})$ within each trust condition. ΔDI is High minus Low (percentage points). p values are Fisher exact tests on {High, Low}×{DI, MIŞ}.

Model	Usable	H-DI	H-MIŞ	L-DI	L-MIŞ	DI% (H)	DI% (L)	ΔDI (pp)	p
Gemma-3-27B-IT	96.0	93	2	82	15	97.9	84.5	+13.4	1.51×10^{-3}
Gemma-2-9B-IT-TR	91.0	92	0	88	2	100.0	97.8	+2.2	0.243
Orbita-v0.1	22.0	19	6	11	8	76.0	57.9	+18.1	0.327
Qwen3-32B	32.5	24	3	35	3	88.9	92.1	-3.2	0.686
Novus-7B-TR	28.0	27	6	21	2	81.8	91.3	-9.5	0.449
WiroAI-TR-8B	24.5	24	3	20	2	88.9	90.9	-2.0	1.000
GPT-OSS-20B-Astro	27.5	17	11	19	8	60.7	70.4	-9.7	0.573
GPT2-TR-10M	4.5	3	0	6	0	100.0	100.0	+0.0	–
Turkcell-LLM-7B	2.0	3	0	1	0	100.0	100.0	+0.0	–
Trendyol-LLM-7B	0.0	0	0	0	0	–	–	–	–

Table 3: Experiment II (explicit past-tense generation) LLM results. Each item was generated three times (`tur1–tur3`), and the final item-level label was determined by majority vote among outputs classifiable as DI or MIŞ (ties excluded). **Usable** = percentage of all items (out of 200) receiving a majority DI or MIŞ label. DI% is computed as $\mathrm{DI}/(\mathrm{DI}+\mathrm{MI})$ within each trust condition. ΔDI is High minus Low (percentage points). p values are Fisher exact tests on {High, Low}×{DI, MIŞ} using item-level majority labels.

ble, or reverse-direction patterns. Qwen3-32B, Novus-7B-TR, WiroAI-TR-8B, and GPT-OSS-20B-Astro all show non-significant High-vs.-Low differences ($p = .686$, $.449$, 1.000, and $.573$, respectively), and several exhibit a numerical reverse shift (i.e., a higher DI share in Low-Trust than in High-Trust contexts). For GPT2-TR-10M and Turkcell-LLM-7B, all usable majority labels are DI, so no informative DI/MIŞ comparison can be made. Trendyol-LLM-7B produces no usable items. Overall, Experiment II shows that explicit past-tense prompting can improve evidential compliance and reveal a trust-sensitive DI/MIŞ pattern in some models (most clearly Gemma-3-27B-IT), but the effect remains highly model-dependent and is weak, unstable, or absent in most others.

7.1. Exploratory Head-Level Modulation of Trust-Cue Attention

As an exploratory mechanistic analysis, we examined attention-weight redistribution in *Gemma-2-9B-IT-TR* under the explicit past-tense prompting condition. We focused on the pre-generation decision state (`q = last_prompt`) and quantified attention directed to the trust-cue span, operationalized as the token sequence immediately preceding *göre*. Across heads, the strongest condition sensitivity was confined to a small subset of mid-to-late layers, with the single best layer at $L = 25$ and the highest-scoring layers concentrated in the range $L = 22$–29. However, the aggregate pairwise contrast was weak and directionally unstable (Δ_{pair} mean $= -0.0179$, 95% bootstrap CI $[-0.0513, 0.0170]$; $P(\Delta_{\text{pair}} > 0) = 46\%$). Under this operationalization, the results do not provide strong evidence that trust is robustly encoded via systematic reallocation of attention mass to the cue span in *Gemma-2-9B-IT-TR*; accordingly, we treat the attention maps as qualitative diagnostic evidence rather than as a stable mechanistic signature. The corresponding visualizations are provided in Appendix C.

Applying the same analysis to *Gemma-3-27B-IT* yielded a larger and more internally consistent

Model	DI% High	DI% Low	Δpp	Acc%	Abstain%	Fisher p	Holm p	3-run same%
Gemma-2-9B-IT-TR	60.0	44.0	16.0	58.0	0.0	0.033	0.268	97.5
Gemma-3-27B-IT	78.9	66.3	12.6	56.0	3.5	0.054	0.381	97.5
GPT-OSS-20B-Astro	94.9	84.0	10.9	55.3	0.5	0.019	0.173	92.0
Orbita-v0.1	91.0	84.0	7.0	53.5	0.0	0.199	0.994	96.0
Turkcell-LLM-7B	97.0	90.0	7.0	53.3	0.5	0.082	0.491	81.0
WiroAI-TR-8B	15.2	11.0	4.2	52.3	0.5	0.408	1.000	91.0
Qwen3-32B	81.0	77.0	4.0	52.0	0.0	0.603	1.000	95.0
Novus-7B-TR	99.0	98.0	1.0	50.5	0.0	1.000	1.000	99.0
Trendyol-LLM-7B	0.0	0.0	0.0	50.0	0.0	1.000	1.000	100.0
GPT2-TR-10M	0.0	–	–	0.0	99.5	–	–	99.5

Table 4: Experiment III (A/B forced-choice prompting) combined model results using item-level majority vote over three runs. DI% High/Low are computed after excluding abstentions (invalid outputs and tie/no-majority items). Δpp = DI% High minus DI% Low. Acc% uses the target mapping High→*-DI*, Low→*-mIş*. Fisher exact tests are computed on High/Low × {DI, mIş} after excluding abstentions; Holm p values correct for multiple testing across models. 3-run same% is the proportion of items with identical outputs across all three replicates.

pair-level contrast. Condition-sensitive modulation was again localized to a restricted subset of mid-to-late-layer heads, but with a clearer concentration in higher layers (best layer $L = 31$; top layers $L = 31, 33, 36, 37$). The aggregate pairwise effect was reliably non-zero under the present metric definition (Δ_{pair} mean $= -0.0983$, 95% bootstrap CI $[-0.1151, -0.0817]$; $P(\Delta_{\text{pair}} > 0) = 9\%$), indicating systematic condition-dependent redistribution of attention toward the trust-cue span. We interpret this as evidence that *Gemma-3-27B-IT* exhibits a more coherent internal sensitivity to trust-related contextual cues than *Gemma-2-9B-IT-TR* in this prompting setup. At the same time, this interpretation should remain qualified, since raw span-level attention mass can be affected by surface differences such as cue-length asymmetries across conditions. The relevant maps are shown in Appendix C.

8. The Third Experiment with LLMs

Experiment III evaluates LLM behavior in a *forced-choice* prompting setup, in which models are explicitly asked to choose between two candidate completions: a form (option A) and a form (option B). This design is more constrained than Experiments I–II because it removes open-ended lexical generation and directly targets the evidential choice itself. The same 200 trust-manipulated Turkish items are used (100 High-Trust, 100 Low-Trust), and each item is queried three times (`tur1-tur3`). Final item-level labels are determined by majority vote across the three runs. Outputs other than `A/B`, as well as ties/no-majority cases, are treated as abstentions and excluded from evidential-rate and Fisher-test calculations.

Table 4 reports per-model condition-wise DI rates (DI% High and DI% Low), the High-minus-Low difference (Δpp), target-mapping accuracy (High→, Low→), and abstention rate over all 200 items. Table 4 reports per-model Fisher exact tests (High-/Low × DI/MIS), Holm-corrected p values across models, and 3-run agreement (*same*%), which measures response stability across the three replicates.

A first important result is that most usable models show a *directionally trust-consistent trend*: DI% is higher in High-Trust than in Low-Trust contexts. The largest shift is observed for neuralwork/gemma-2-9b-it-tr (Δpp $= 16.0$), followed by google/gemma-3-27b-it (12.6 pp) and ocaklisemih/gpt-oss-20b-turkish-astrology-gguf (10.9 pp). Several other models also show smaller positive shifts (e.g., mradermacher/Orbita-v0.1-i1-GGUF, Turkcell-LLM-7b-v1, WiroAI/wiroai-turkish-llm-8b, Qwen/Qwen3-32B, NovusResearch/Novus-7b-tr$_v1$).

Trendyol-LLM-7b-base-v1.0 shows no shift (0.0 vs. 0.0). This pattern suggests that, under direct A/B comparison, many models weakly track the trust manipulation in the predicted direction.

At the same time, the results also show that *base-rate preferences remain strong* and often dominate behavior. Several models overwhelmingly prefer one option across both conditions (e.g., very high DI rates for Novus, Turkcell, Orbita, and GPT-OSS; near-zero DI rates for Trendyol and GPT2-TR-10M), which compresses the effect of trust and limits target-mapping accuracy. Consistent with this, Acc% values remain close to chance for most models (roughly 50–58%), even when the direction of the trust effect is correct. In other words, models can show a small High>Low DI shift without demonstrating robust context-sensitive evidential selection at the item level.

Statistical testing reinforces this conclusion. Some models yield nominally small Fisher exact p values (e.g., $p = .019$ for GPT-OSS and $p = .033$ for Gemma-2-9B-IT-TR; Table 4), but *none* remains significant after Holm correction across models. Thus, Experiment III provides evidence for a weak, directionally consistent trend in several systems, but not for a robust per-model trust effect after correction for multiple comparisons.

Finally, response stability across the three replicates is generally high (3-run same% often above 90%), indicating that model choices in this A/B paradigm are relatively deterministic once the prompt format is fixed. This is methodologically useful: compared with open-ended prompting, the forced-choice setup reduces output-format variability and improves comparability across models. However, the combination of high replicate agreement and near-chance Acc% for many systems suggests that the models are often making stable choices driven by global option preferences rather than reliably applying the trust cue in a human-like way.

9. Discussion

This study makes theoretical and methodological contributions to Turkish evidentiality and its evaluation in LLMs. Human results show a clear trust effect: more credible external sources favor *-DI*, while less credible sources favor *-mIş* (Section 5; Table 1; Figure 2). In contrast, current LLMs do not reliably reproduce this pattern, and their apparent evidential behavior varies with prompt format, output compliance, and response constraints (Tables 2, 3, 4).

The paper's central theoretical contribution is a direct experimental test of a *trustworthiness*-based perspective on Turkish evidentiality, inspired by the Giannakidou–Mari framework. To our knowledge, this is a novel application of that framework to Turkish *-DI*/*-mIş*. Rather than treating evidential morphology only as a categorical source-type marker, we test whether the credibility of an explicitly external source shifts evidential choice; the human data show that it does.

This matters because the source is explicitly external in both conditions ((e.g., X'e göre 'according to X')); the manipulation changes only the source's trustworthiness. The observed shift therefore supports an analysis in which Turkish evidential morphology is sensitive not just to source *type*, but also to source *quality* and the speaker's resulting commitment profile. In Giannakidou–Mari terms, the pattern is consistent with differences in support strength for the prejacent: more trustworthy sources favor stronger commitment (and relatively more *-DI*), whereas less trustworthy sources favor weaker, evidence-weighted commitment (and relatively more *-mIş*) (Giannakidou and Mari, 2016, 2018, 2021; Boscaro et al., 2025).

Our findings do not by themselves settle the long-standing question of whether *-DI* is lexically a direct evidential or a past tense form whose "directness" effects arise pragmatically. However, they do provide an important constraint on the debate. Because the manipulation keeps source attribution explicit and varies trustworthiness instead, the human pattern is difficult to reduce to a simple source-present/source-absent contrast. The results are more naturally captured by accounts that allow context-sensitive interactions among evidential morphology, speaker commitment, and discourse-pragmatic reasoning. In this sense, the present data strengthen the case for treating Turkish evidential choice as part of a broader inferential and commitment-sensitive system, rather than a purely rigid source-labeling mechanism.

What the LLM results may suggest. Across the three LLM experiments, the main result is a dissociation between *surface task performance under prompting constraints* and *human-like trust-sensitive evidential reasoning*. In Experiment I (open gap-fill prompting), many models produce a high proportion of unusable outputs, and even among usable outputs the trust effect is highly model-dependent. In Experiment II (explicit past-tense prompting), usability improves substantially for some models (especially Gemma variants), and Gemma-3-27B-IT shows the clearest trust-consistent shift, but the pattern remains weak, absent, or unstable in most models. In Experiment III (A/B forced-choice), response stability becomes high and several models show directionally trust-consistent trends, yet effects are generally modest, target-mapping accuracy remains near chance for many systems, and no per-model effect survives Holm correction. Thus, these results suggest that many LLMs can exhibit *format-dependent traces* of the expected pattern, but they do not reliably reproduce the human sensitivity to source trustworthiness as a stable, cross-paradigm property. This is precisely the kind of mismatch that a linguistically targeted, human–LLM comparison can reveal: a model may appear "reasonable" in one prompt format while still failing to encode the underlying contextual factor in a robust way. A related question is why explicit past-tense prompting changes model performance. We suggest that the main reason is task specification: in the open gap-fill setting, models must jointly infer a plausible lexical completion, tense, and evidential morphology, which leaves the task relatively underspecified. By explicitly requiring a past-tense completion in Experiment II, the prompt narrows the response space and makes

the *-DI*/*-mIş* contrast more directly relevant to the decision. In this sense, the improvement for some models is consistent with recent work showing that more specific instructions can reduce prompt sensitivity relative to underspecified prompts, even if they do not solve the deeper representational problem (Pecher et al., 2026). At the same time, our results also align with recent findings that morphological generalization remains difficult for LLMs in agglutinative languages such as Turkish, especially when models must produce the appropriate inflection rather than merely recognize it (Ismayilzada et al., 2025).

The paper also contributes a reusable evaluation paradigm for Turkish source-sensitive semantics/pragmatics by combining: (i) a human cloze dataset, (ii) an expanded LLM-only dataset, and (iii) three prompting regimes that vary constraint level (open generation, explicit tense generation, forced choice). This multi-paradigm design makes it possible to separate at least three sources of variation that are often conflated in LLM evaluations: *compliance/formatting behavior*, *base-rate morphological preference*, and *context sensitivity to the trust manipulation*. In particular, the contrast between Experiments I–III shows that forcing the response space can improve comparability and stability, but it can also expose strong default biases that limit genuine contextual adaptation. More broadly, the study argues for evaluating Turkish LLMs on controlled linguistic contrasts with human baselines, rather than relying only on broad benchmarks. A useful point of comparison is *TurBLiMP*, the first Turkish benchmark of linguistic minimal pairs, which evaluates LMs on 16 grammatical phenomena and has significantly expanded linguistically informed evaluation for Turkish (Başar et al., 2025). Our study is complementary to that line of work. Whereas TurBLiMP focuses on minimal-pair judgments over a broad range of grammatical phenomena, it does not target evidentiality. This matters because evidentiality is not only a morphosyntactic contrast in Turkish, but also a source-sensitive semantic–pragmatic phenomenon tied to speaker commitment, information source, and discourse reasoning. Our results therefore extend Turkish LLM evaluation into a domain that is central to Turkish grammar but not captured by existing minimal-pair benchmarks. For evidentiality and related phenomena, this is especially important because the relevant behavior is not merely lexical accuracy but context-dependent mapping between linguistic form and epistemic/discourse structure.

10. Conclusion

This paper investigated how *source trustworthiness* shapes Turkish evidential choice in the *-DI*/*-mIş* contrast, and whether current LLMs track this sensitivity. In a controlled human cloze experiment, we found a robust and replicable trust effect: participants produced relatively more *-DI* in High-Trust contexts and relatively more *-mIş* in Low-Trust contexts. This provides new empirical support for a trust-/commitment-based perspective on Turkish evidentiality and shows that the Giannakidou–Mari framework offers a productive theoretical lens for Turkish data.

Across three LLM experiments (open gap-fill prompting, explicit past-tense prompting, and forced-choice A/B prompting), model behavior was substantially more variable and strongly dependent on prompt format, output compliance, and base-rate suffix preferences. While some models showed weak or local trust-consistent trends, the effect was not robust across models or paradigms. Overall, the results reveal a clear human–LLM gap in source-sensitive evidential reasoning and motivate future work on more faithful evaluation and modeling of evidential and commitment-related meaning in Turkish.

Ethics Statement

This work investigated the interaction between contextual source trustworthiness and Turkish evidential morphology (*-DI* vs. *-mIş*) using both human cloze responses and LLM-based preference scoring. Human data were collected under informed-consent procedures and analyzed only at the aggregate level; no personally identifying information is reported, and all examples are presented in anonymized form. The human-subjects component of this study received Institutional Review Board (IRB) approval.

For the computational component, model outputs are treated as task-conditioned behavioral responses rather than as evidence of human-like semantic or cognitive representations. Because LLM preferences can be sensitive to evaluation protocol design (including prompting format and tokenization effects), our findings should not be overgeneralized to broad claims about model competence or human evidential reasoning. Finally, by focusing on Turkish, this study helps address the English-centric bias of much computational psycholinguistics and LLM evaluation.

Limitations

This study has several limitations. First, the strongest behavioral evidence comes from the human experiment on Dataset 1, whereas the larger Dataset 2 was used only for LLM evaluation; future work should test whether the same human

trust effect scales to the expanded item set. Second, LLM results are partly constrained by prompt compliance and output-format errors, especially in open-generation settings, which can reduce the amount of usable evidential data and make model comparisons harder to interpret. Third, although the forced-choice setup improves control and stability, it may introduce response biases (e.g., option or format preferences) that are not specific to evidential reasoning. The exploratory attention analysis is also limited in scope since it was conducted only for two relatively successful models from Experiment II and is intended as a preliminary qualitative diagnostic rather than a definitive mechanistic account.

Acknowledgements

We thank Anastasia Giannakidou for valuable input on the semantic and pragmatic theory and the interpretation of the results, Ming Xiang for suggestions on the human experimental design, and Christopher Potts and Craig Thorburn for their comments on future directions. We are also grateful to the anonymous reviewers for their helpful feedback and suggestions. Any remaining errors are our own.

11. Bibliographical References

Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.

Ayhan Aksu-Koç and Dan I. Slobin. 1986. A psychological account of the development and use of evidentials in Turkish. In Wallace Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*, pages 159–167. Ablex Publishing Corporation, Norwood, NJ.

Merve Ataman-Devrim, Gaye Soley, and Ayhan Aksu-Koç. 2025. Children's understanding of source reliability and knowledge generalizability from grammatical cues: Evidence from Turkish. *Journal of Child Language*, pages 1–23. First published online 11 December 2025.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: LLM hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16495–16510, Suzhou, China. Association for Computational Linguistics.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, and Savaş Yıldırım. 2025. Tokenization standards for linguistic integrity: Turkish as a benchmark. *arXiv preprint arXiv:2502.07057*.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2026. Tokens with meaning: A hybrid tokenization approach for Turkish.

Abdullah Bezir, Furkan Burhan Türkay, and Cengiz Asmazoğlu. 2024. Wiroai turkish llm 8b. Hugging Face model card.

Marie Boscaro, Anastasia Giannakidou, and Alda Mari. 2025. Evidence type and trustworthiness: The view from social media. In *The Grammar of Interaction*, pages 98–131.

Kasper Boye. 2012. *Epistemic Meaning: A Crosslinguistic and Functional-Cognitive Study*. De Gruyter Mouton, Berlin.

Roger W. Brown and Eric H. Lenneberg. 1954. A study in language and cognition. *Journal of Abnormal and Social Psychology*, 49(3):454–462.

cenkersisman. gpt2-turkish-10m. Hugging Face model card. Accessed: 2026-01-20.

Ferdinand de Haan. 2001. The place of inference within the evidential system. *International Journal of American Linguistics*, 67(2):193–219.

Patrick Dendale and Liliane Tasmowski. 2001. Introduction: evidentiality and related notions. *Journal of Pragmatics*, 33(3):339–348.

Yakup Abrek Er, Ilker Kesen, Gözde Gül Şahin, and Aykut Erdem. 2025. Cetvel: A unified benchmark for evaluating language understanding, generation and cultural capacity of LLMs for Turkish. *arXiv preprint arXiv:2508.16431*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Anastasia Giannakidou. 1998. *Polarity Sensitivity as (Non)veridical Dependency*. John Benjamins, Amsterdam/Philadelphia.

Anastasia Giannakidou. 2006. Only, emotive factive verbs, and the dual nature of polarity dependency. *Language*, 82(3):575–603.

Anastasia Giannakidou and Alda Mari. 2016. Epistemic future and epistemic must: Nonveridicality, evidence, and partial knowledge. In Joanna Błaszczak, Anastasia Giannakidou, Dorota Klimek-Jankowska, and Krzysztof Migdalski, editors, *Mood, Aspect, Modality Revisited: New Answers to Old Questions*, pages 75–124. The University of Chicago Press, Chicago.

Anastasia Giannakidou and Alda Mari. 2018. A unified analysis of the future as epistemic modality: The view from greek and italian. *Natural Language & Linguistic Theory*, 36(1):85–129.

Anastasia Giannakidou and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought*. University of Chicago Press, Chicago.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge, London and New York.

Demet Gül. 2009. Semantics of Turkish evidential -(i)miş. In S. Ay, Ö. Aydın, İ. Ergenç, S. Gökmen, S. İşsever, and D. Peçenek, editors, *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6–8, 2008*, pages 177–186. Harrassowitz, Wiesbaden.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Lars Johanson. 2003. Evidentiality in Turkic. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *Studies in Evidentiality*, pages 273–290. John Benjamins, Amsterdam/Philadelphia.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Birsu Kandemirci, Anna Theakston, Ditte Boeg Thomsen, and Silke Brandt. 2023. Does evidentiality support source monitoring and false belief understanding? a cross-linguistic study with Turkish- and english-speaking children. *Child Development*, 94(4):889–904.

Sercan Karakaş. 2026. Clause-internal or clause-external? Testing Turkish reflexive binding in adapted versus chain of thought large language models. *arXiv preprint arXiv:2602.00380*.

Sercan Karakaş and Yusuf Şimşek. 2026. From lemmas to dependencies: What signals drive light verbs classification? In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 220–227, Rabat, Morocco. Association for Computational Linguistics.

Onur Keleş and Nazik Dinçtopal Deniz. 2024. A comparative study with human data: Do LLMs have superficial language processing? In *2024 32nd IEEE Conference on Signal Processing and Communications Applications (SIU)*.

Onur Keleş and Nazik Dinçtopal Deniz. 2025. When men bite dogs: Testing good-enough parsing in Turkish with humans and large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 219–231, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Jaklin Kornfilt. 1997. *Turkish*. Routledge, London.

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. *arXiv preprint arXiv:2005.11401*.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association*

for Computational Linguistics: ACL 2024, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. 2026. Agenthallu: Benchmarking automated hallucination attribution of LLM-based agents. *arXiv preprint arXiv:2601.06818*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

mradermacher. Orbita-v0.1-i1-GGUF. Hugging Face model card. Accessed: 2026-01-20.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

neuralwork. gemma-2-9b-it-tr. Hugging Face model card. Accessed: 2026-01-20.

NovusResearch. Novus-7b-tr$_v$1. Hugging Face model card. Accessed: 2026-01-20.

ocaklisemih. gpt-oss-20b-Turkish-astrology-gguf. Hugging Face model card. Accessed: 2026-01-20.

Branislav Pecher, Michal Spiegel, Robert Belanec, and Jan Cegin. 2026. Revisiting prompt sensitivity in large language models for text classification: The role of prompt underspecification. *arXiv preprint arXiv:2602.04297*.

Vladimir A. Plungian. 2001. The place of evidentiality within the universal grammatical space. *Journal of Pragmatics*, 33(3):349–357.

Nilüfer Şener. 2011. *Semantics and Pragmatics of Evidentials in Turkish*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch

Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

Dan I. Slobin. 1996. From "thought and language" to "thinking for speaking". In John J. Gumperz and Stephen C. Levinson, editors, *Rethinking Linguistic Relativity*, pages 70–96. Cambridge University Press, Cambridge.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Cagri Toraman, Ahmet Kaan Sever, Ayşe Aysu Cengiz, Elif Ecem Arslan, Görkem Sevinç, Sarp Kantar, Mete Mert Birdal, Yusuf Faruk Güldemir, Ali Buğra Kanburoğlu, Sezen Felekoğlu, Birsen Şahin Kütük, Büşra Tufan, Elif Genç, Serkan Coşkun, Gupse Ekin Demir, Muhammed Emin Arayıcı, Olgun Dursun, Onur Gungor, Susan Üsküdarlı, Abdullah Topraksoy, and Esra Darıcı. 2026. TurkBench: A benchmark for evaluating Turkish large language models. In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 126–154, Rabat, Morocco. Association for Computational Linguistics.

Trendyol. Trendyol-LLM-7b-base-v1.0. Hugging Face model card. Accessed: 2026-01-20.

TURKCELL. Turkcell-LLM-7b-v1. Hugging Face model card. Accessed: 2026-01-20.

Özgür Uğur, Mahmut Göksu, Mahmut Çimen, Musa Yılmaz, Esra Şavirdi, Alp Talha Demir, Rumeysa Güllüce, İclal Çetin, and Ömer Can Sağbaş. 2026. Mecellem models: Turkish models trained from scratch and continually pre-trained for the legal domain. *arXiv preprint arXiv:2601.16018*.

Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Ercenur Ünal and Anna Papafragou. 2020. Relations between language and cognition: Evidentiality and sources of knowledge. *Topics in Cognitive Science*, 12(1):115–135.

Robert Underhill. 1976. *Turkish Grammar*. MIT Press, Cambridge, MA.

Darmawan Wicaksono, Hasri Akbar Awal Rozaq, and Nevfel Boz. 2025. Emotion recognition for low-resource Turkish: Fine-tuning BERTurk on TREMO and testing on xenophobic political discourse. *arXiv preprint arXiv:2505.12160*.

Thomas Willett. 1988. A cross-linguistic survey of the grammaticization of evidentiality. *Studies in Language*, 12(1):51–97.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Jeremy Zehr and Florian Schwarz. 2018. Penncontroller for internet based experiments (ibex).

A. Source Categories

Condition	Turkish	English
High	Resmi Gazete	Official Gazette
High	Kurumun resmi sayfası	Official website of the institution
High	Hastanenin astığı duyuru	Hospital notice board announcement
High	Belediyenin SMS uyarısı	Municipality SMS alert
High	Resmi mail	Official email
Low	Anonim telegram grubu	Anonymous Telegram group
Low	Yan apartmanda oturan komşu	Neighbor living in the adjacent building
Low	Diğer şirketteki çalışan	Employee at another company
Low	Yanlış haberler paylaşan Instagram grubu	Instagram group sharing false news
Low	Sokakta geçen biri	A random passerby on the street

Table 5: Examples of high- and low-trust source categories used in the experiment.

B. Prompt Templates

B.1. Turkish Prompts

Prompt 1:

Prompt 1 (TR)

Aşağıdaki cümlede "_____." ile belirtilmiş bir boşluk var.
Bu boşluğu cümlenin yapısının akışına en uygun fiille tamamla.

Cevap verirken kurallar:
- Sadece boşluğa gelecek kısmıyaz.
- Cümleyi tekrar etme.
- Açıklama yapma.
Tek kelime cevap ver ve fiil kullanarak tamamla
- Cevap vermeden geçme.

Cümle:
`{prefix}` _____.

Cevap:

Prompt 2:

Prompt 2 (TR)

Aşağıdaki cümlede "_____." ile belirtilmiş bir boşluk var.
Bu boşluğu cümlenin yapısının akışına ve geçmiş zamana göre tamamlanabilecek en uygun kelimeyle tamamla.

Cevap verirken kurallar:
- Sadece boşluğa gelecek kısmıyaz.
- Cümleyi tekrar etme.
- Açıklama yapma.
- Cevap vermeden geçme.

Cümle:
`{prefix}` _____.

Cevap:

Prompt 3:

Prompt 3 (TR)

Cümledeki boşluğu doldurmak için doğru seçeneği seç.

Cümle: `{sentence}`

Seçenekler:
A) `{di}`
B) `{mis}`

Kurallar:
- Sadece A ya da B yaz.
- Açıklama yapma.

Cevap:

B.2. English Prompts

Prompt 1:

Prompt 1 (EN)

There is a blank indicated by "_____." in the sentence below.
Fill this blank with the verb that best fits the flow and structure of the sentence.

Rules for your answer:
- Write only what should go in the blank.
- Do not repeat the sentence.
- Do not provide an explanation.
Answer with a single word and complete it using a verb.
- Do not skip answering.

Sentence:
`{prefix}` _____.

Answer:

Prompt 2:

Prompt 2 (EN)

There is a blank indicated by "_____." in the sentence below.
Fill this blank with the most appropriate word that can complete the sentence according to the flow/structure and past tense.

Rules for your answer:
- Write only what should go in the blank.
- Do not repeat the sentence.
- Do not provide an explanation.
- Do not skip answering.

Sentence:
`{prefix}` _____.

Answer:

Prompt 3:

Prompt 3 (EN)

Choose the correct option to fill the blank in the sentence.

Sentence: `{sentence}`

Options:
A) `{di}`
B) `{mis}`

Rules:
- Write only A or B.
- Do not provide an explanation.

Answer:

C. Exploratory Attention Maps

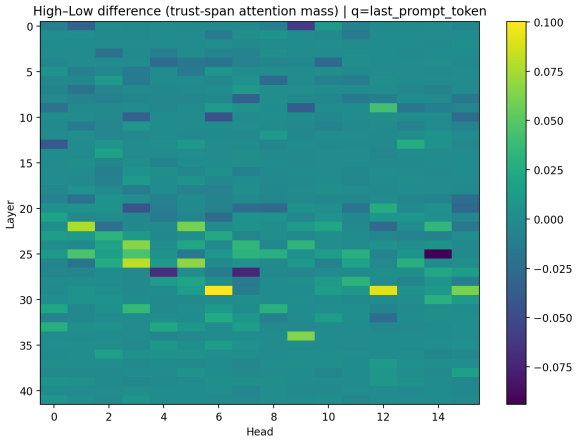

Figure 3: Gemma-2-9B-IT-TR (Experiment II): high–low difference in trust-span attention mass (query token: `last_prompt_token`). The pair-level effect is weak and inconsistent (Δ_{pair} mean $= -0.0179$, 95% bootstrap CI $[-0.0513, 0.0170]$; $P(\Delta_{\text{pair}} > 0) = 46\%$), so we treat this as a qualitative diagnostic rather than strong evidence of robust trust encoding.

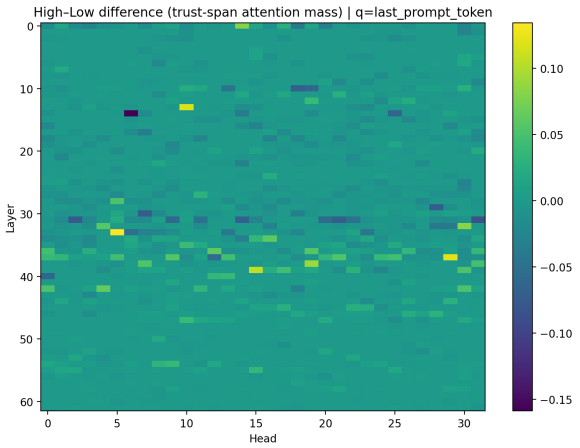

Figure 4: Gemma-3-27B-IT (Experiment II): high–low difference in trust-span attention mass (query token: `last_prompt_token`). The map shows localized modulation concentrated in a small cluster of mid-to-late layers and heads, consistent with the exploratory analysis in Section 7.