

Not All Evidence Is Equal: Source-Sensitive Calibration in Large Language Models

Sercan Karakaş
University of Chicago
skarakas@uchicago.edu

Yusuf Şimşek
Fırat University
ysimsek@firat.edu.tr

Abstract

Large language models are increasingly used in settings where information must be evaluated not only for content, but also for reliability. This paper introduces a human-grounded evaluation of source-sensitive calibration in Turkish and English.¹ Across hearer-side uptake tasks, we manipulate source reliability while holding propositional content constant, and test whether humans and LLMs adjust action-readiness judgments according to source trustworthiness. Turkish human participants show robust source sensitivity: they are substantially more willing to act on information attributed to high-trust sources than to low-trust sources, while Turkish evidential morphology itself does not independently affect uptake once the source is overtly specified. LLMs show weaker calibration. In Turkish, models partially register explicit source reliability, but the effect is much smaller than in humans and there is no reliable effect of the *-DI/-mlş* contrast. In English, models show small effects of both source reliability and the lexical marker *apparently*, but no robust interaction between them. These results suggest that current LLMs encode source cues only partially and often fail to convert them into human-like judgments about when information is reliable enough to guide action.

1 Introduction

Recent progress in large language models (LLMs) has shifted evaluation beyond text generation toward evidence summarization, question answering, claim assessment, and decision support under uncertainty (OpenAI et al., 2024; Wang et al., 2023; Huang et al., 2024; Vashurin et al., 2025; Ramos et al., 2026). In these settings, models must recover not only propositional content, but also where information comes from, how trustworthy the source is,

and whether the evidence is reliable enough to support belief, recommendation, or action. This distinction is central to human communication: hearers do not treat official announcements, expert reports, direct observations, and informal rumors as equally action-worthy, even when the proposition is held constant (Hovland and Weiss, 1951; Koenig et al., 2004).

This raises a key question for LLM evaluation: do models use source reliability as a stable, decision-relevant variable, or merely as a surface cue? Existing benchmarks have advanced the evaluation of language understanding, factuality, reasoning, and robustness (Hendrycks et al., 2021; Lin et al., 2022; Srivastava et al., 2023; Liang et al., 2023), but they typically test whether a model gives the correct answer, not whether it conditions its judgment on the epistemic status of the input. This is crucial for agentic and decision-support systems: treating a low-credibility rumor as nearly as action-worthy as an institutional report is a failure of evidence-based calibration, not just factual accuracy. Source-sensitive calibration is therefore an action-grounding problem: models should be evaluated on how strongly they believe, recommend, or act on information given its source (Kim et al., 2024).

We introduce a controlled evaluation of source-sensitive calibration through hearer-side uptake: the degree to which information is treated as reliable enough to guide action. Our human-grounded design follows prior work showing that robust ML benchmarks require human judgments and carefully documented data collection, especially when the target behavior involves meaning, preference, uncertainty, commonsense, or decision-relevant reasoning (Christiano et al., 2017; Kiela et al., 2021; Gebru et al., 2021; Krumdick et al., 2025, among others). The benchmark manipulates source reliability while holding propositional content constant. Participants and models read attributed sen-

¹All data and code are available at: <https://github.com/ysufsimsek/Not-All-Evidence-Is-Equal-Source-Sensitive-Calibration-in-Large-Language-Models/tree/main>

tences and answer action-oriented questions on a 1–7 scale, with higher ratings indicating stronger behavioral uptake. This design allows us to test whether humans and LLMs convert explicit source-reliability cues into calibrated judgments about whether information is reliable enough to guide action.

Turkish is particularly useful for this evaluation because it grammatically marks evidential distinctions in the past domain through *-DI* and *-mİş* (Slobin and Aksu, 1982; Aksu-Koç, 1988; Johanson, 2000; Karakaş and Şimşek, 2026a). Broadly, *-DI* is the unmarked past form and is commonly associated with direct evidence, whereas *-mİş* can convey reportative or inferential meanings, although its interpretation is context-dependent. English, by contrast, typically expresses such meanings through optional lexical devices such as *reportedly*, *apparently*, or *according to X* (Aikhenvald, 2004). This typological difference allows us to examine whether action-readiness is shaped by the reliability of an explicitly named source, by evidential form, or by both. Existing Turkish benchmarks (Yüksel et al., 2024; Bayram et al., 2025a; Başar et al., 2025; Karakaş and Şimşek, 2026b), including TurkishMMLU, TR-MMLU, and TurBLiMP, primarily evaluate general language understanding, classification, and grammatical competence (Karakaş, 2026a). In contrast, the present work directly manipulates source reliability and measures its effect on action-readiness, rather than production alone (Karakaş and Şimşek, 2026a), in both Turkish and English. This design allows us to test whether humans and LLMs calibrate their judgments according to the epistemic quality of the source.

For English, the Turkish materials are translated while preserving the same 2×2 structure: Source Reliability is manipulated as High Trust versus Low Trust, and Evidential Form contrasts an unmarked/simple form with the lexical marker *apparently*. Because English lacks the Turkish *-DI/-mİş* morphological contrast, this manipulation is intended as a functional parallel rather than a semantic equivalent. In particular, *apparently* can signal indirect evidence and reduced certainty or speaker commitment, which partially overlaps with reportative uses of Turkish *-mİş*, although the two forms differ in their broader semantic and pragmatic distributions (Kornfilt, 1997; Göksel and Kerslake, 2005; Aikhenvald, 2004). The English experiment therefore tests whether source-sensitive uptake is

similarly modulated when evidential meaning is expressed lexically rather than morphologically.

Turkish and English also differ in their NLP resource profiles: English is high-resource and central to LLM training and evaluation, whereas Turkish is comparatively underrepresented in multilingual NLP (Joshi et al., 2020; Toraman, 2024; Bayram et al., 2025b). Because typological and resource differences cannot be independently disentangled in the present design, we treat the two languages as parallel probes of source-sensitive calibration under different evidential systems rather than as a controlled causal comparison. This distinction is especially relevant for multilingual and language-adapted LLMs, whose behavior may vary with pretraining coverage, tokenization, and adaptation data (Joshi et al., 2020; Toraman, 2024; Bayram et al., 2025b; Karakaş, 2026b).

Thus, this paper makes three contributions. First, we introduce a human-grounded evaluation framework for source-sensitive calibration based on 1–7 action-readiness judgments. Second, we provide a controlled Turkish human baseline for how hearers use source reliability when deciding whether information is reliable enough to guide action. Third, we evaluate Turkish and English LLM uptake behavior to test whether models convert explicit source cues into appropriately cautious action-guiding judgments across different evidential systems. We argue that source-grounded reasoning should be evaluated as a calibration problem: models should be tested not only on whether they recognize source cues, but also on whether they use those cues to determine when information is reliable enough to support action.

The remainder of this paper is structured as follows. Section 2 presents the experimental framework and datasets used to evaluate source-sensitive calibration. Section 3 reports the Turkish human hearer-side uptake experiment, establishing a human baseline for the effects of source reliability and evidential morphology. Section 4 evaluates LLMs on the same task, first in Turkish, including an expanded 200-item evaluation, and then in English using a lexical evidential contrast. Section 5 discusses the implications of these findings for source-sensitive calibration, evidentiality, and human-grounded LLM evaluation. Finally, Section 6 summarizes the main conclusions, followed by limitations, ethical considerations, and supplementary materials in the appendix.

2 Experimental Overview

We operationalize source-sensitive calibration using a 2×2 factorial design crossing *Source Reliability* and *Evidential Form*. Source Reliability has two levels: High Trust and Low Trust, defined by whether the attributed source is well positioned to know or assess the proposition at issue, for example through relevant expertise, direct access to the information, or an institutional role (Giannakidou and Mari, 2021, 2024, 2026). High-Trust items therefore use sources with a strong epistemic basis for the relevant proposition, whereas Low-Trust items use informal or unofficial sources without comparable access or authority. The target proposition is held constant across source conditions.

Each critical item consists of a proposition, an attributed source, an evidential form, and an action-oriented question. For example, one scenario states that classes were cancelled and attributes this either to the institution’s official page or to an unofficial social-media account. Respondents then rate how likely they would be to go to class given the information. Evidential Form contrasts *-DI* with *-mİş* in Turkish and an unmarked/simple form with *apparently* in English. The English contrast is intended as a functional parallel rather than a semantic equivalent: *apparently* can signal indirect evidence and reduced certainty or speaker commitment, partially overlapping with reportative uses of *-mİş*. Full examples of all four conditions in both languages are provided in Appendix A. The dependent measure is a 1–7 action-readiness rating indicating willingness to act as if the proposition were true. Where necessary, responses are directionally recoded so that higher values consistently represent stronger uptake. We predict higher ratings for High-Trust than for Low-Trust sources and test whether this effect varies with Evidential Form.

We use three evaluation settings. Dataset 1 is the Turkish human-facing set, containing 10 scenarios realized across the four factorial conditions, yielding 40 critical items, together with 30 fillers for a total of 70 items. Fillers were included to reduce the transparency of the experimental manipulation and were excluded from the factorial analyses. Dataset 2 expands the materials to 50 scenarios, yielding 200 Turkish critical items—100 High-Trust and 100 Low-Trust items distributed across the two evidential conditions—plus the same 30 fillers. This larger set is used for the LLM-only evaluation to test whether the observed pat-

terns generalize beyond the smaller human-facing dataset. The third setting uses an English translation of Dataset 2 with the same factorial structure. The materials were translated by the authors and independently reviewed by two native speakers of American English for grammaticality, naturalness, and clarity. Because Turkish and English differ both in how evidential meaning is expressed and in their NLP resource profiles, we treat the two evaluations as parallel tests rather than as a controlled causal comparison between languages.

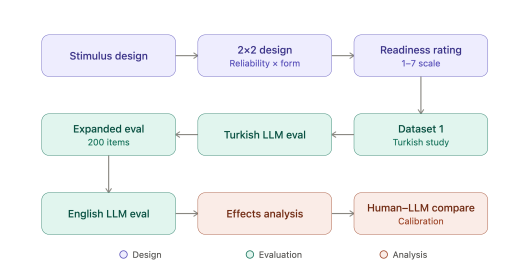


Figure 1: Overview of the experimental pipeline, from stimulus design and the 2×2 manipulation of Source Reliability and Evidential Form to the Turkish human and LLM evaluations, expanded evaluation, English LLM evaluation, and final effect-based calibration analyses.

3 Human hearer-side uptake experiment

The human hearer-side experiment received approval from the University of Chicago Institutional Review Board (IRB26-0198) prior to data collection. Seventy-eight participants completed the study. After excluding two test submissions and two non-native Turkish speakers, the final analysis included 74 native speakers of Turkish, all at least 18 years old. The study was implemented online using PCIBEX/PENNCONTROLLER (Zehr and Schwarz, 2018). Participants read fully inflected Turkish sentences followed by an action-oriented question and responded on a 1–7 scale. The design crossed Source Reliability (High Trust vs. Low Trust) with Evidential Form (*-DI* vs. *-mİş*). Each participant completed 40 experimental trials: 10 scenarios crossed with two source types and two evidential forms. Fillers were excluded, yielding 2,960 experimental responses, with 740 observations in each cell.

Because the raw scale did not have the same direction across all scenarios, ratings were recoded so that higher values consistently indicated stronger behavioral uptake. For action-promoting items, the original 1–7 rating was retained. For action-

inhibiting items, responses were reverse-coded as 8 – rating, following standard practice for aligning item polarity in survey and psychological research (Chyung et al., 2018; Suárez-Alvarez et al., 2018).

Uptake results. Human participants showed a robust source-reliability effect. Table 1 summarizes the mean uptake ratings by source condition and evidential form.

Source condition	-DI	-mIs
High trust	5.91	5.93
Low trust	4.22	4.25

Table 1: Human hearer-side uptake ratings. Higher scores indicate stronger behavioral uptake, i.e., greater willingness to act as if the proposition is true.

High-Trust contexts elicited substantially stronger behavioral uptake than Low-Trust contexts, whereas evidential morphology had little independent effect. Mean uptake was 5.91 for High-Trust+-DI and 5.93 for High-Trust+-mIs, compared with 4.22 for Low-Trust+-DI and 4.25 for Low-Trust+-mIs. Thus, participants strongly distinguished reliable from unreliable sources, but did not assign different uptake to -DI and -mIs once the source was overtly specified.

A participant-level repeated-measures analysis confirmed this pattern. High-Trust contexts increased uptake by 1.69 points on the 1–7 scale, 95% CI [1.34, 2.03], $t(73) = 9.70$, $p = 9.15 \times 10^{-15}$. In contrast, the evidential morphology effect was not reliable: -mIs differed from -DI by only 0.02 points, 95% CI [-0.10, 0.05], $t(73) = -0.66$, $p = 0.514$. The Source $\times$ Evidential interaction was also absent, $t(73) = 0.11$, $p = 0.913$. Robustness checks and reaction-time analyses are reported in Appendix B.

4 LLM hearer-side uptake experiment

We next evaluated whether LLMs show human-like hearer-side uptake in the same Turkish 2 $\times$ 2 design. Models read fully inflected Turkish sentences and answered an action-oriented question on a 1–7 scale. The prompt instructed each model to return only a single number between 1 and 7. Across all model runs, generation settings were held constant: temperature was set to 0.1, top- p to 1, do_sample to true, and the maximum number of generated tokens to 50. Each item was presented three times, yielding 120 experimental responses

per model before exclusions. Fillers were excluded from all analyses. As in the human analysis, ratings were directionally recoded where necessary so that higher scores consistently indicated stronger behavioral uptake. The prompt is given in Appendix D.

Model selection. For the Turkish experiments, we prioritized Turkish-focused or Turkish-adapted models, alongside a small number of general-purpose multilingual models for comparison. For English, we retained these models and added general-purpose counterparts where available, including Gemma-2-9B-IT alongside Gemma-2-9B-IT-TR. Full model names and grouping details are provided in Appendix C.

Uptake results. Across the eight included models, LLMs showed only a weak numerical effect of source reliability, much smaller than the human effect. At the model level, the mean High–Low source effect was +0.35 points, 95% CI [-0.00, 0.71], $t(7) = 2.35$, $p = 0.051$. Thus, the source effect is marginal even across models and remains much smaller than the human source effect of +1.69 points. Models also tended to assign high action-readiness even to Low-Trust information, suggesting weak source-sensitive calibration. Full model-by-model uptake results are reported in Appendix E. As in the human uptake experiment, there was no reliable independent effect of evidential morphology. Collapsing across source conditions, the mean morphology effect across the eight models was only +0.02 points, 95% CI [-0.15, 0.19], $t(7) = 0.29$, $p = 0.781$. Thus, the models do not show evidence of using the -DI/-mIs contrast as an independent cue for action-guiding uptake once the source is overtly specified. The Source $\times$ Evidential interaction was likewise not reliable.

4.1 Expanded 200-item LLM hearer-side uptake evaluation

To test whether the smaller LLM uptake results generalize to a broader item set, we ran an enlarged LLM-only hearer-side experiment with 200 trust-manipulated Turkish sentences. The dataset was expanded from 10 to 50 scenarios using the same construction procedure and epistemic-authority criteria as Dataset 1: for each additional scenario, we defined an action-relevant proposition and question, paired the proposition with High-Trust and Low-Trust sources differing in their epistemic standing with respect to that information, and crossed this manipulation with Evidential Form. The design

therefore crossed 50 scenarios with Source Reliability (High Trust vs. Low Trust) and Evidential Form (*-DI* vs. *-mIs*), with each item repeated three times. This yielded 600 experimental responses per model before exclusions: 150 responses in each of the four Source $\times$ Evidential cells. Fillers were excluded from all analyses. As before, ratings were directionally recoded where necessary so that higher scores consistently indicated stronger behavioral uptake. All eight evaluated models produced valid 1–7 experimental responses for nearly all trials. Seven models produced complete data, with 600/600 valid experimental responses. Kizagan-E4B had 597/600 valid responses; the three missing responses all came from the same High-Trust+*-DI* item. We retained Kizagan in the analysis because the missingness was minimal.

Uptake results. The expanded experiment shows a statistically clearer source-trust effect than the smaller LLM uptake set, but the effect remains small relative to humans. Model-averaged uptake was 5.51 for High-Trust+*-DI* and 5.54 for High-Trust+*-mIs*, compared with 5.23 for Low-Trust+*-DI* and 5.24 for Low-Trust+*-mIs*. Treating model as the unit of generalization, the mean High–Low source effect was +0.29 points, 95% CI [0.12, 0.47], $t(7) = 3.95$, $p = 0.0056$. Thus, LLMs do show reliable sensitivity to explicit source reliability in the expanded set. However, the effect is much smaller than the human source effect of +1.69 points. Table 2 reports the model-by-model results for the expanded Turkish uptake evaluation.

Evidential morphology and interaction. The enlarged item set provides no evidence that LLMs use the *-DI/-mIs* contrast as an independent cue for behavioral uptake. Across models, the mean morphology effect was only +0.017 points, 95% CI [–0.072, 0.106], $t(7) = 0.45$, $p = 0.665$. The morphology contrast was also null within each source condition: within High Trust, *-mIs* exceeded *-DI* by only +0.026 points ($p = 0.662$), and within Low Trust the difference was only +0.008 points ($p = 0.803$). The Source $\times$ Evidential interaction was likewise absent, interaction = +0.017, 95% CI [–0.106, 0.141], $t(7) = 0.33$, $p = 0.749$.

Calibration and response-style artifacts. Although the source effect is reliable in the expanded set, response-style artifacts remain substantial. The source effect is positive in all eight models, but its magnitude varies widely. The

strongest effects appear in GPT-OSS-20B-Astro (+0.58), Gemma-2-9B-IT-TR (+0.49), Qwen3-32B (+0.49), and Gemma-3-27B-IT (+0.31), whereas the weakest effects appear in Magibu-11B-v0.8 (+0.01), Novus-7B-TR (+0.13), and WiroAI-TR-8B (+0.13). Using the 50 scenario sets as paired items, the source effect is significant for five models: GPT-OSS-20B-Astro ($p = .0011$), Gemma-2-9B-IT-TR ($p = .0045$), Qwen3-32B ($p < .001$), Gemma-3-27B-IT ($p = .0228$), and Kizagan-E4B ($p = .026$). In contrast, Magibu-11B-v0.8 ($p = .965$), Novus-7B-TR ($p = .593$), and WiroAI-TR-8B ($p = .080$) do not show reliable within-model effects.

Several models are also compressed toward the high end of the transformed uptake scale: transformed scores of 7 account for 85.3% of Magibu-11B-v0.8 responses, 83.7% of Novus-7B-TR responses, and 73.7% of Kizagan-E4B responses. The cell means show large differences in overall scale use across models. For example, WiroAI-TR-8B gives relatively low uptake scores overall, with means around 3.79–3.95 across conditions, whereas Novus-7B-TR stays near the high end, with means around 6.21–6.43. This means that even when models show a positive High–Low source effect, their absolute action-readiness ratings are not directly comparable without considering scale use. The expanded experiment therefore shows that LLMs encode explicit source reliability to some extent, but this sensitivity is uneven across models and remains under-calibrated relative to humans.

4.2 English LLM hearer-side uptake evaluation

We next tested whether the uptake pattern generalizes to English, where the evidential contrast is expressed lexically rather than morphologically. The English materials were translated from the Turkish scenarios by the authors while preserving the same 2×2 design: Source Reliability (High Trust vs. Low Trust) crossed with Evidential Form (regular tense vs. *apparently* + tense). The translations were then independently reviewed by two native speakers of American English for grammaticality, naturalness, clarity, and preservation of the intended source-reliability contrast. Each item was repeated three times, yielding 600 experimental responses per model before exclusions: 150 responses in each Source $\times$ Evidential cell. Fillers were excluded from all analyses. As in the Turkish uptake analy-

Model	H--DI	H-- <i>mls</i>	L--DI	L-- <i>mls</i>	Source eff.	Source <i>p</i>	Morph. eff.
Gemma-2-9B-IT-TR	4.74	4.75	4.27	4.24	+0.49	.0045	-0.01
Gemma-3-27B-IT	5.27	5.39	5.01	5.02	+0.31	.0228	+0.06
GPT-OSS-20B-Astro	6.08	6.06	5.43	5.55	+0.58	.0011	+0.05
Kizagan-E4B	6.03	5.73	5.73	5.63	+0.20	.026	-0.20
Magibu-11B-v0.8	6.00	6.28	6.12	6.15	+0.01	.965	+0.15
Novus-7B-TR	6.37	6.43	6.21	6.33	+0.13	.593	+0.09
Qwen3-32B	5.69	5.71	5.27	5.15	+0.49	< .001	-0.05
WiroAI-TR-8B	3.91	3.95	3.79	3.81	+0.13	.080	+0.03
Model-level mean	5.51	5.54	5.23	5.24	+0.29	.0056	+0.02

Note. Higher scores indicate stronger behavioral uptake. Source effect is High Trust minus Low Trust. Morphology effect is *-mls* minus *-DI*, collapsed across source conditions. Source *p* values are scenario-level paired tests within each model. Kizagan-E4B had 597/600 valid experimental responses; all other models had 600/600. The model-level mean treats model as the unit of generalization.

Table 2: Expanded 200-item LLM hearer-side uptake results. Cell values are mean uptake scores after directional recoding.

ses, ratings were directionally recoded where necessary so that higher scores consistently indicated stronger action in accordance with the reported proposition. The English analysis included eleven usable models with recoverable numeric responses.

Uptake results. The English experiment showed statistically reliable but small effects of both source reliability and lexical evidential marking. Model-averaged uptake was 5.00 for High-Trust+regular tense and 4.80 for High-Trust+*apparently*, compared with 4.74 for Low-Trust+regular tense and 4.55 for Low-Trust+*apparently*. Thus, High-Trust sources produced higher action-consistency than Low-Trust sources, and regular-tense sentences produced higher action-consistency than sentences with *apparently*. However, both effects were small in magnitude.

Treating model as the unit of generalization, the mean High–Low source effect was +0.252 points, 95% CI [0.163, 0.342], $t(10) = 6.28$, $p < .001$. The regular-minus-*apparently* effect was also reliable but small, +0.195 points, 95% CI [0.067, 0.323], $t(10) = 3.40$, $p = .0068$. The Source $\times$ Evidential interaction was not reliable, +0.016 points, 95% CI [−0.133, 0.165], $t(10) = 0.24$, $p = .819$. Thus, unlike the Turkish morphology contrast, the English lexical marker *apparently* produced a small but reliable reduction in action-consistency. However, this effect did not depend on source reliability. Table 3 reports the model-by-model English uptake results.

Lexical evidential marking and interaction. The English results differ from the Turkish uptake results in one respect: the lexical marker *apparently*

produced a small but statistically reliable reduction in action-consistency. Across models, regular-tense items received higher aligned ratings than *apparently* items by about 0.20 points on the 1–7 scale. This suggests that models register *apparently* as a weak commitment-reducing or action-weakening cue. However, the effect is modest, and it is not interactively integrated with source reliability. The Source $\times$ Evidential interaction was close to zero and not reliable across models, indicating that the effect of *apparently* did not robustly depend on whether the attributed source was high-trust or low-trust.

Across-item tests yielded the same qualitative pattern. Averaging over models and treating item as the unit of analysis, the source effect was reliable, +0.252 points, 95% CI [0.144, 0.360], $t(49) = 4.68$, $p < .001$. The regular-minus-*apparently* effect was also reliable, +0.195 points, 95% CI [0.117, 0.273], $t(49) = 5.05$, $p < .001$. The interaction was again absent, +0.016 points, 95% CI [−0.105, 0.136], $t(49) = 0.26$, $p = .794$.

Calibration and response-style artifacts. As in the Turkish LLM uptake experiments, the English results show detectable source sensitivity but weak calibration. The source effect is positive for all eleven usable models, but its magnitude varies substantially. The strongest source effects appear in Gemma-3-27B-IT (+0.49), Gemma-4-E4B-IT (+0.42), Gemma-2-9B-IT-TR (+0.33), and Qwen3-32B (+0.32), whereas the weakest effects appear in Kizagan-E4B, Magibu-11B-v0.8, and Novus-7B-TR, each around +0.09. The *apparently* effect is also model-dependent: Qwen3-32B, Magibu-11B-v0.8, GPT-OSS-20B-

Model	H-Reg.	H-App.	L-Reg.	L-App.	Source eff.	Reg. > App.	Interaction
Gemma-2-9B-IT	4.45	4.26	4.17	4.01	+0.27	+0.18	+0.03
Gemma-2-9B-IT-TR	4.59	4.45	4.31	4.07	+0.33	+0.19	-0.09
Gemma-4-E4B-IT	5.27	5.24	4.90	4.77	+0.42	+0.08	-0.10
GPT-OSS-20B-Astro	6.01	5.27	5.51	5.41	+0.18	+0.42	+0.63
Kizagan-E4B	6.42	6.37	6.35	6.26	+0.09	+0.07	-0.05
Magibu-11B-v0.8	3.50	3.15	3.52	2.95	+0.09	+0.46	-0.21
Novus-7B-TR	5.03	5.17	4.91	5.10	+0.09	-0.17	+0.05
Qwen3-32B	5.52	5.14	5.26	4.77	+0.32	+0.44	-0.11
WiroAI-TR-8B	4.29	4.03	4.03	3.82	+0.23	+0.23	+0.05
Gemma-3-27B-IT	5.30	5.13	4.85	4.60	+0.49	+0.21	-0.08
Llama-3.1-8B-Instruct	4.61	4.55	4.33	4.31	+0.26	+0.04	+0.05
Model-level mean	5.00	4.80	4.74	4.55	+0.25	+0.20	+0.02

Note. Higher scores indicate stronger action-consistency after directional recoding. H = High Trust; L = Low Trust; Reg. = regular tense; App. = *apparently* + tense. Source effect is High Trust minus Low Trust, collapsed across evidential form. Reg. > App. is regular tense minus *apparently*, collapsed across source condition. The interaction is the difference-of-differences (H-Reg. – H-App.) – (L-Reg. – L-App.). The model-level mean treats model as the unit of generalization.

Table 3: English LLM hearer-side uptake results across eleven usable models. Cell values are mean aligned action-consistency scores.

Astro, and WiroAI-TR-8B show clearer regular-minus-*apparently* differences, while Llama-3.1-8B-Instruct shows almost no *apparently* effect and Novus-7B-TR shows a small effect in the opposite direction.

Overall, the English evaluation suggests that LLMs weakly track both explicit source reliability and English lexical evidential marking. However, the effects remain small: approximately 0.25 points for source reliability and 0.20 points for *apparently* on a 1–7 scale. Moreover, the absence of an interaction suggests that models treat source reliability and *apparently* largely as additive cues, rather than integrating them into a richer pragmatic judgment about whether information is reliable enough to guide action.

5 Discussion

This paper introduced a human-grounded evaluation of source-sensitive calibration in LLMs. The central question was not whether models can recognize evidential markers in isolation, but whether they use information about source reliability to calibrate action-guiding judgments. Across Turkish human data, Turkish LLM evaluations, and English LLM evaluations, the results show a clear asymmetry: humans strongly adjust uptake according to source reliability in Turkish, whereas LLMs show only weak, uneven, and model-dependent calibration. The main contribution is therefore both empirical and methodological. Empirically, the results show that current LLMs only partially convert source information into action-readiness. Methodologically, the paper proposes a controlled way to

evaluate whether models distinguish information that is merely presented from information that is reliable enough to guide action. The Turkish human results establish the strongest baseline for source-sensitive uptake. Participants were substantially more willing to act on information attributed to high-trust sources than to low-trust sources, even though the propositional content was held constant. This supports the view that hearers do not treat communicated propositions as context-free pieces of content. Instead, they evaluate information through epistemic vigilance: who provided the information, how credible that source is, and whether the evidence is strong enough to support belief or action (Sperber et al., 2010; Boulat, 2015, 2023). In this sense, uptake is not simply semantic interpretation. It is a pragmatic and epistemic judgment about whether a proposition should guide behavior.

At the same time, evidential morphology did not independently affect Turkish hearers’ action-readiness once the source was overtly specified. Participants gave nearly identical uptake ratings for *-DI* and *-mİş* within the same source condition, suggesting that explicit source reliability was the dominant cue in this design. Since all critical items contained attribution frames such as *X’e göre* ‘according to X’, hearers primarily relied on the credibility of the named source rather than on the *-DI/-mİş* contrast when deciding whether to act. This result cautions against treating *-mİş* as a simple marker of low trust or reduced actionability. In the present data, *-mİş* did not lower uptake relative to *-DI* when the same source was explicitly provided. Instead, action-readiness appears to depend on how evidential morphology, overt attribution,

and source credibility jointly determine epistemic commitment.

More theoretically, this pattern is compatible with analyses in which Turkish evidentiality contributes more than purely temporal or aspectual information. Although *-mİş* is historically and semantically related to perfect/anterior aspect, its interpretation is systematically connected to the speaker’s evidential basis and epistemic relation to the proposition (Slobin and Aksu, 1982; Aksu-Koç, 1988). On such an analysis, *-mİş* need not itself encode that the proposition is unreliable; rather, it can constrain the type or accessibility of evidence on which commitment to the proposition is based. The reliability of that evidence can then be supplied by the discourse context and, in the present experiment, by the explicitly named source. The absence of a morphology effect once source identity is fixed is therefore not evidence that evidential morphology is semantically vacuous. Instead, it suggests that evidential marking and source reliability contribute at different levels: evidential morphology characterizes the evidential or modal basis for the proposition, whereas contextual information determines the quality of that basis and consequently the degree of uptake. This division of labor is more naturally compatible with evidential/modal accounts of Turkish *-mİş* than with a view in which its contribution is exhausted by anterior aspect alone, and it aligns with commitment- and veridicality-based approaches in which evidential expressions regulate the speaker’s epistemic relation to *p* (Giannakidou and Mari, 2021, 2024; Boscaro et al., 2025; Giannakidou and Mari, 2026). At the same time, because the present experiment was not designed to directly contrast modal and aspectual semantic analyses, this conclusion should be understood as a theoretical implication of the pattern rather than as decisive evidence against an anterior-based account.

The Turkish LLM results show partial cue sensitivity but weak calibration. In the expanded 200-item evaluation, the models showed a reliable High-Trust versus Low-Trust effect, indicating that they are not completely blind to explicit source reliability. However, the effect was much smaller than the human effect and varied substantially across models. Several models also showed response-scale compression, assigning high action-readiness even to low-trust information. This pattern suggests that models can often detect source-reliability cues at a surface level, but do not reliably trans-

late those cues into appropriately cautious action-guiding judgments. In other words, the models exhibit some source sensitivity, but not human-like source-sensitive calibration.

This distinction between cue sensitivity and calibration is central to the methodology. A model may produce slightly lower ratings for low-trust sources and still remain poorly calibrated if its absolute ratings are too high, if its source effect is small relative to humans, or if its responses are dominated by scale-use artifacts. The relevant failure is therefore not simply failure to notice the source phrase. It is failure to adjust downstream commitment strongly enough. This matters for LLM evaluation because many applications require models to decide how much weight to place on source-dependent information. In summarization, question answering, recommendation, tool use, and decision support, the risk is not only that a model may produce false content, but that it may treat weakly grounded information as action-worthy (Kadavath et al., 2022; Pan et al., 2024; Hwang et al., 2025; Kim et al., 2024; Steyvers et al., 2025).

The English results extend this conclusion to a typologically different evidential system. Unlike Turkish, English does not encode the *-DI/-mİş* contrast in verbal morphology. The English version therefore tested whether models are sensitive to source reliability when evidential meaning is expressed lexically through *apparently*. The models showed a small but reliable source-reliability effect, and they also assigned slightly lower action-consistency ratings to sentences with *apparently* than to regular-tense sentences. This suggests that models represent *apparently* as a weak commitment-reducing or action-weakening cue. However, the effect was small, and the Source $\times$ Evidential interaction was not reliable. Thus, models did not robustly integrate lexical evidential marking with source reliability. They treated source type and *apparently* largely as additive cues, rather than combining them into a richer pragmatic judgment about whether the information was reliable enough to guide action.

The Turkish–English comparison clarifies what the benchmark may reveal. In Turkish, explicit source reliability dominates uptake, while the *-DI/-mİş* contrast contributes little once the source is overtly specified. In English, *apparently* produces a small independent reduction in action-consistency, but this effect does not depend on whether the source is high-trust or low-trust. This contrast sug-

gests that the action-guiding force of evidential marking depends partly on how evidential meaning is encoded. A grammatical evidential such as *-mİş* may not independently reduce uptake when an attribution phrase already specifies the source. A lexical marker such as *apparently*, by contrast, may more transparently signal reduced certainty or weaker speaker commitment. Yet in both languages, the absence of a robust interaction suggests that LLMs do not fully integrate source reliability and evidential form in a human-like pragmatic model of uptake.

This methodology also separates shallow recognition from action-relevant use, thereby aligning with the broader goal of rethinking progress and evaluation in NLP. The benchmark evaluates not only task performance, but also whether models generalize in a trustworthy and action-relevant way when the same information is attributed to sources of different reliability. A model may recognize that an unofficial social media account is less reliable than an institutional source, but the crucial question is whether this recognition changes how strongly the model treats the proposition as action-worthy. This distinction is difficult to observe in standard accuracy-based benchmarks, where the model’s task is usually to select or generate a contentful answer. In contrast, the present task asks whether the same content receives different degrees of uptake depending on its source. This makes source-sensitive calibration directly measurable.

The framework is also portable. Turkish provides a morphologically explicit test case through the *-Dİ/-mİş* contrast, while English provides a lexically marked comparison through *apparently*. Other languages and evidential systems could be tested using the same logic: hold propositional content constant, manipulate source reliability and evidential marking, and measure whether humans and models adjust action-readiness accordingly. This makes the benchmark useful not only for Turkish or English, but for broader multilingual evaluation of uncertainty, evidentiality, and pragmatic reasoning.

More broadly, this work contributes to human-grounded NLP evaluation by comparing model behavior against a human baseline for source-sensitive uptake. Instead of evaluating model outputs in isolation, we ask whether LLMs use information about the source in ways that meaningfully affect belief, recommendation, and action. This provides a useful perspective for evaluating reasoning under uncertainty, factuality, and trustworthi-

ness.

The theoretical contribution is to connect evidentiality, source credibility, and action-guiding commitment in a single evaluation framework. The results show that source-grounded reasoning is not merely recognizing evidential markers or attribution phrases, but using source information to calibrate how strongly a proposition should guide behavior. Humans show robust source-sensitive uptake, whereas LLMs show weak and uneven calibration. They often register that not all evidence is equal, but do not reliably translate this into human-like judgments about when information is reliable enough to act on.

6 Conclusion

This paper evaluated source-sensitive calibration in humans and LLMs across Turkish and English through hearer-side uptake and action-readiness judgments. The human results show robust source sensitivity: Turkish speakers are substantially more willing to act on information attributed to high-trust sources than to low-trust sources. At the same time, Turkish evidential morphology does not independently modulate uptake once the source is overtly specified. LLMs show weaker and less stable calibration. In Turkish, models partially register explicit source reliability, but the effect is much smaller than in humans and there is no reliable effect of the *-Dİ/-mİş* contrast. In English, models show small effects of both source reliability and *apparently*, but no reliable interaction between them. To summarize, the results show that source-grounded reasoning should be evaluated as a calibration problem, not only as an accuracy or marker-recognition problem. Models used for decision support must distinguish information that is merely available from information that is reliable enough to guide action.

Limitations

This study has four limitations. First, all critical items include explicit source phrases. This design isolates source reliability, but may reduce the independent contribution of evidential morphology in Turkish; future work should test morphology-only contexts and contexts where the source is inferred rather than overtly named. Second, uptake is measured through final 1–7 action-readiness ratings, which capture decision outcomes but not the online process of source integration. Future work could

combine this paradigm with reaction times, eye-tracking, or process-level measures. In addition, the human baseline is currently limited to Turkish. The English evaluation tests LLMs on translated materials, but does not include an English human comparison group. Future work should collect English human uptake judgments to test whether the small *apparently* effect observed in models corresponds to human pragmatic behavior. Finally, the LLM results are limited to the models tested here. As multilingual and Turkish-capable models improve, the benchmark should be rerun. At the same time, response-format failures, ceiling effects, and scale compression are themselves informative evidence of weak calibration, since they show that some models struggle to use rating scales in ways that support reliable source-sensitive evaluation.

Ethical Considerations

The human experiment received approval from the University of Chicago Institutional Review Board (IRB26-0198) prior to data collection. Participants were adult native speakers of Turkish and completed the study online after providing informed consent. The task involved ordinary language judgments about source reliability and action-readiness; it did not ask participants to disclose sensitive personal information. Responses were analyzed in aggregate, and no personally identifying information is reported. Participation was voluntary, and participants were compensated for their time.

The LLM evaluations use constructed stimuli rather than real user data. The benchmark is intended to evaluate source-sensitive calibration, not to provide deployment-ready judgments about whether specific real-world claims should be trusted or acted upon. Because source labels such as institutional, expert, informal, or unofficial can reflect social and cultural assumptions, future versions of the benchmark should test a wider range of sources, domains, and participant populations to avoid overgeneralizing from the present materials. Before release, the dataset was checked to ensure that it does not contain names, contact information, or other personally identifying information. The materials consist of constructed experimental stimuli about ordinary source-reliability scenarios and were also checked for offensive or harmful content. Human responses are reported only in aggregate.

The released artifacts, including constructed stimuli and prompts, are provided for research and

evaluation purposes. The data and evaluation materials are released under a CC BY-NC 4.0 license. These artifacts are intended to support reproducibility and further research on source-sensitive calibration, and should not be used as a deployment-ready tool for deciding whether real-world claims should be trusted or acted upon.

AI assistance was used for language-level support, including spell checking, wording suggestions, and troubleshooting LaTeX issues. All scientific content, experimental design, analyses, interpretations, and final writing decisions were reviewed and verified by the authors.

References

- Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.
- Ayhan A. Aksu-Koç. 1988. *The Acquisition of Aspect and Modality: The Case of Past Reference in Turkish*. Cambridge University Press.
- 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.
- Ali Bayram. 2025. Magibu-11b: A Turkish-Native Multilingual Vision-Language Model with Optimized Tokenization. <https://huggingface.co/magibu/magibu-11b-v0.8>. Developed by Magibu AI Research.
- M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Banu Diri, Savaş Yıldırım, and Öner Aytaş. 2025a. *Setting standards in Turkish NLP: TR-MMLU for large language model evaluation*. *Preprint*, arXiv:2501.00593.
- M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, and Savaş Yıldırım. 2025b. *Tokenization standards and evaluation in natural language processing: A comparative analysis of large language models on Turkish*. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.
- Abdullah Bezir, Furkan Burhan Türkay, and Cengiz Asmazoğlu. 2024. *WiroAI/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 Susana Rodríguez Rosique and Jordi M. Antolí Martínez, editors, *The Grammar of Interaction: Epistemicity, Information Management*

- and Discourse in Language Use, volume 46 of *IVITRA Research in Linguistics and Literature*, pages 98–131. John Benjamins Publishing Company.
- Kira Boulat. 2015. [Hearer-oriented processes of strength assignment: A pragmatic model of commitment](#). *Belgian Journal of Linguistics*, 29(1):19–40.
- Kira Boulat. 2023. *The Pragmatics of Commitment*. Peter Lang, Berlin.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Seung Youn Chyung, Julie R. Barkin, and Jennifer A. Shamsy. 2018. [Evidence-based survey design: The use of negatively worded items in surveys](#). *Performance Improvement*, 57(3):16–25.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Communications of the ACM*, 64(12):86–92.
- Anastasia Giannakidou and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. University of Chicago Press, Chicago.
- Anastasia Giannakidou and Alda Mari. 2024. [The human and the mechanical: Logos, veridicality judgment, and GPT models](#). *Intellectica*, 2024/2(81):37–54.
- Anastasia Giannakidou and Alda Mari. 2026. *Modal Sentences*. Key Topics in Semantics and Pragmatics. Cambridge University Press.
- Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge, London.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Proceedings of the International Conference on Learning Representations*.
- Carl I. Hovland and Walter Weiss. 1951. [The influence of source credibility on communication effectiveness](#). *Public Opinion Quarterly*, 15(4):635–650.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *Preprint*, arXiv:2311.05232.
- Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. [Retrieval-augmented generation with estimation of source reliability](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34279–34303, Suzhou, China. Association for Computational Linguistics.
- Lars Johanson. 2000. [Turkic indirectives](#). In Lars Johanson and Bo Utas, editors, *Evidentials: Turkic, Iranian and Neighbouring Languages*, pages 61–88. Mouton de Gruyter.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- 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, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Sercan Karakaş. 2026a. [Clause-internal or clause-external? Testing Turkish reflexive binding in adapted versus chain-of-thought large language models](#). In *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic 10*, pages 172–186. Linguistic Society of America.
- Sercan Karakaş. 2026b. [Plausibility as commonsense reasoning: Humans succeed, large language models do not](#). In *Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics*, pages 309–328, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).
- Sercan Karakaş and Yusuf Şimşek. 2026a. [Benchmarking source-sensitive reasoning in Turkish: Humans and LLMs under evidential trust manipulation](#). In *Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics*, pages 329–346, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).
- Sercan Karakaş and Yusuf Şimşek. 2026b. [Do we still need fine tuning? Turkish sentiment analysis in the era of large language models](#). In *2026 34th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124. Association for Computational Linguistics.
- Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. “I’m not sure, but...”: Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.
- Alican Kiraz. 2026. Kızagan-E4B: A Turkish Reasoning Model Fine-Tuned from Gemma 4 E4B. <https://huggingface.co/AlicanKiraz0/Kizagan-E4B-Turkish-Reasoning-Model>.
- Melissa A. Koenig, Fabrice Clément, and Paul L. Harris. 2004. [Trust in testimony: Children’s use of true and false statements](#). *Psychological Science*, 15(10):694–698.
- Jaklin Kornfilt. 1997. *Turkish*. Descriptive Grammars. Routledge, London.
- Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. [No free labels: Limitations of LLM-as-a-judge without human grounding](#). *Preprint*, arXiv:2503.05061.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic Evaluation of Language Models](#). *Transactions on Machine Learning Research*.
- 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*, pages 3214–3252.
- neuralwork. [gemma-2-9b-it-tr](#). Hugging Face model card. Accessed: 2026-01-20.
- NovusResearch. [Novus-7b-tr_v1](#). Hugging Face model card. Accessed: 2026-01-20.
- ocaklisemih. [gpt-oss-20b-Turkish-astrology-gguf](#). Hugging Face model card. Accessed: 2026-01-20.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. [Not all contexts are equal: Teaching LLMs credibility-aware generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863. Association for Computational Linguistics.
- Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. [EuroLLM-22B: Technical report](#). *Preprint*, arXiv:2602.05879.
- Dan I. Slobin and Ayhan A. Aksu. 1982. [Tense, aspect and modality in the use of the Turkish evidential](#). In Paul J. Hopper, editor, *Tense-Aspect: Between Semantics and Pragmatics*, pages 185–200. John Benjamins.
- Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. [Epistemic vigilance](#). *Mind & Language*, 25(4):359–393.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. [Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models](#). *Transactions on Machine Learning Research*.
- Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. [What large language models know and what people think they know](#). *Nature Machine Intelligence*, 7(2):221–231.
- Javier Suárez-Alvarez, Ignacio Pedrosa, Luis M. Lozano, Eduardo García-Cueto, Marcelino Cuesta, and José Muñiz. 2018. [Using reversed items in likert scales: A questionable practice](#). *Psicothema*, 30(2):149–158.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert

- Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- 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, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Cagri Toraman. 2024. [Adapting open-source generative large language models for low-resource languages: A case study for Turkish](#). In *Proceedings of the 4th Workshop on Multilingual Representation Learning*, pages 30–44. Association for Computational Linguistics.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. [Benchmarking uncertainty quantification methods for large language models with LM-Polygraph](#). *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. [A survey on large language model based autonomous agents](#). *Preprint*, arXiv:2308.11432.
- 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, and 41 others. 2025. [Qwen3 Technical Report](#). *Preprint*, arXiv:2505.09388.
- Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. [TurkishMMLU: Measuring massive multitask language understanding in Turkish](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.
- Jérémy Zehr and Florian Schwarz. 2018. [PennController for internet based experiments \(IBEX\)](#).

A High- and Low-Trust Evidence Contexts

Condition	Sentence and question
High Trust + - <i>DI</i>	Kurumun resmi sayfasına göre dersler iptal edildi. <i>Bu habere göre derse gitme ihtimaliniz nedir? (1–7)</i>
High Trust + - <i>mİş</i>	Kurumun resmi sayfasına göre dersler iptal edilmiş. <i>Bu habere göre derse gitme ihtimaliniz nedir? (1–7)</i>
Low Trust + - <i>DI</i>	Resmi olmayan bir sosyal medya hesabına göre dersler iptal edildi. <i>Bu habere göre derse gitme ihtimaliniz nedir? (1–7)</i>
Low Trust + - <i>mİş</i>	Resmi olmayan bir sosyal medya hesabına göre dersler iptal edilmiş. <i>Bu habere göre derse gitme ihtimaliniz nedir? (1–7)</i>

Table 4: Turkish source-reliability contrast.

Condition	Sentence and question
High Trust + simple form	According to the institution's official page, classes were cancelled. <i>According to this news, how likely are you to go to class? (1–7)</i>
High Trust + <i>apparently</i>	According to the institution's official page, apparently, classes were cancelled. <i>According to this news, how likely are you to go to class? (1–7)</i>
Low Trust + simple form	According to an unofficial social media account, classes were cancelled. <i>According to this news, how likely are you to go to class? (1–7)</i>
Low Trust + <i>apparently</i>	According to an unofficial social media account, apparently, classes were cancelled. <i>According to this news, how likely are you to go to class? (1–7)</i>

Table 5: English source-reliability contrast.

B Robustness checks and reaction times for the human hearer-side uptake experiment

Robustness checks and reaction times. The source-reliability effect remained robust in non-parametric and scenario-level analyses. A Wilcoxon test confirmed the effect ($p = 1.03 \times 10^{-10}$), and the scenario-level High–Low difference was 1.69 points across the 10 scenarios, $t(9) = 17.61$, $p = 2.78 \times 10^{-8}$. Neither the evidential morphology effect nor the interaction was reliable across scenarios.

Response-time analyses provided weaker evidence. Low-Trust contexts were numerically slower than High-Trust contexts, but after controlling for sentence word count, the effect was only marginal: approximately 5.8% slower, $p = 0.083$. Sentence length reliably predicted RT, whereas evidential morphology and the interaction did not. Thus, the rating data provide strong evidence for source-sensitive uptake, while the RT data provide only suggestive evidence for increased deliberation in Low-Trust contexts. Overall, Turkish hearers strongly use explicit source reliability when converting linguistic information into action-guiding judgments, while the *-DI* vs. *-mİş* contrast does not independently modulate uptake once the source is overtly specified.

C Models

The models evaluated in this study consist of Turkish-focused, Turkish-adapted, and general-purpose multilingual large language models. Because the Turkish and English uptake evaluations include partially overlapping model sets, we report them separately below.

Models used in the Turkish uptake evaluations.

The Turkish hearer-side uptake experiments included the following eight models:

- Gemma-2-9B-IT-TR ([neuralwork](#))
- Gemma-3-27B-IT ([Team et al., 2025](#))
- GPT-OSS-20B-Astro ([ocaklisemih](#))
- Kizagan-E4B ([Kiraz, 2026](#))
- Magibu-11B-v0.8 ([Bayram, 2025](#))
- Novus-7B-TR ([NovusResearch](#))
- Qwen3-32B ([Yang et al., 2025](#))
- WiroAI-TR-8B ([Bezir et al., 2024](#))

Models used in the English translated uptake evaluation. The English translated uptake experiment included eleven usable models:

- Gemma-2-9B-IT ([Team et al., 2024](#))
- Gemma-2-9B-IT-TR ([neuralwork](#))
- Gemma-4-E4B-IT ([Team et al., 2026](#))
- GPT-OSS-20B-Astro ([ocaklisemih](#))
- Kizagan-E4B ([Kiraz, 2026](#))
- Magibu-11B-v0.8 ([Bayram, 2025](#))
- Novus-7B-TR ([NovusResearch](#))
- Qwen3-32B ([Yang et al., 2025](#))
- WiroAI-TR-8B ([Bezir et al., 2024](#))
- Gemma-3-27B-IT ([Team et al., 2025](#))
- Llama-3.1-8B-Instruct ([Grattafiori et al., 2024](#))

Model grouping. For the Turkish experiments, we prioritized Turkish-focused or Turkish-adapted models, including Gemma-2-9B-IT-TR ([neuralwork](#)), GPT-OSS-20B-Astro ([ocaklisemih](#)), Kizagan-E4B ([Kiraz, 2026](#)), Magibu-11B-v0.8 ([Bayram, 2025](#)), Novus-7B-TR ([NovusResearch](#)), and WiroAI-TR-8B ([Bezir et al., 2024](#)), while Gemma-3-27B-IT ([Team et al., 2025](#)) and Qwen3-32B ([Yang et al., 2025](#)) provide general-purpose multilingual comparison points. This selection emphasizes models expected to have strong Turkish capabilities, which is important for sensitivity to Turkish source expressions and evidential morphology.

For the English evaluation, we retained the Turkish-oriented models and expanded the model set with general-purpose multilingual models and counterparts where available. In particular, Gemma-2-9B-IT ([Team et al., 2024](#)) was included alongside the Turkish-adapted Gemma-2-9B-IT-TR ([neuralwork](#)), together with Gemma-4-E4B-IT ([Team et al., 2026](#)) and Llama-3.1-8B-Instruct ([Grattafiori et al., 2024](#)). This broader set provides a comparison between language-adapted and more general model variants.

Computational budget. All model evaluations were run in Google Colab using a G4 GPU runtime. Runtime varied across models. The shortest runtime was observed for Llama-3.1-8B-Instruct, which completed in approximately 33 seconds. The longest runtime was observed for Qwen3-32B, which completed in approximately 21 minutes and 35 seconds. The remaining models completed within this runtime range.

D Prompt for the LLM Hearer-Side Uptake Experiment (3 runs)

Prompt (TR)

Sana bazı cümleler verilecektir.
Her cümleden sonra, o cümleyle ilgili bir soru göreceksin.
Görevin, verilen soruya göre 1 ile 7 arasında bir değerlendirme yapmaktır.

1 = çok düşük
7 = çok yüksek

Cümle: "{cumle}"
Soru: {soru}

Sadece 1 ile 7 arasında tek bir sayı ver.

SADECE CEVAP ÜRET VE AÇIKLAMA ASLA YAPMA. SORUYU TEKRAR ETME.

Cevap:

Figure 2: Turkish prompt template used in the LLM hearer-side uptake experiment.

Prompt (EN)

You will be given some sentences.
After each sentence, you will see a question related to that sentence.
Your task is to give a rating between 1 and 7 according to the question.

1 = very low
7 = very high

Sentence: "{sentence}"
Question: {question}

Give only a single number between 1 and 7.

PRODUCE ONLY THE ANSWER. NEVER GIVE AN EXPLANATION. DO NOT REPEAT THE QUESTION.

Answer:

Figure 3: English prompt template used in the LLM hearer-side uptake experiment.

E Model-by-model LLM uptake results

Table 6 reports the model-by-model hearer-side uptake results for the eight included LLMs.

Table 6: LLM hearer-side uptake results across eight models. Source effect is High Trust minus Low Trust. Morphology effect is $-mI\text{§}$ minus $-DI$, collapsed across source conditions.

Model	Valid	Source	<i>p</i>	Morph.
Gemma-2-9B-IT-TR	120/120	+1.27	< .001	-0.07
Gemma-3-27B-IT	120/120	+0.48	< .001	-0.12
GPT-OSS-20B-Astro	120/120	+0.32	.002	+0.15
Novus-7B-TR	120/120	+0.08	.858	+0.48
Qwen3-32B	120/120	+0.53	.001	-0.13
WiroAI-TR-8B	120/120	+0.05	.568	-0.02
Magibu-11B-v0.8	120/120	+0.00	–	+0.00
Kizagan-E4B-TR	118/120	+0.09	.505	-0.12
Model-level mean	–	+0.35	.051	+0.02

Note. Higher scores indicate stronger behavioral uptake. Source *p* values are item/round-level tests within each model. The model-level mean tests whether the effect generalizes across the eight included models. Magibu-11B-v0.8 produced maximal transformed uptake across all conditions, leaving no meaningful within-model variance for a significance test. Kizagan-E4B-Turkish-Reasoning produced mostly raw 7 responses, so its results should be interpreted cautiously as evidence of poor response-scale calibration rather than genuine source-sensitive uptake.

F Instructions for Human Participants in the Hearer-Side Uptake Experiment

The experiment was administered in Turkish. Screenshots of the original Turkish instructions and items are provided in the supplementary materials. For readability, we provide English translations of the participant-facing instructions below.

F.1 Consent and General Information

Thank you in advance for your participation and interest in this study!

This study is part of a research project conducted by the researchers listed below.

- Researchers: [Sercan Karakaş and Yusuf Şimşek]
- Institutions: [The University of Chicago and Fırat University]

If you would like to contact the research team, you can reach us at the email addresses listed above.

This study takes approximately 15–20 minutes to complete when done carefully and without interruption. Before the study begins, you will be asked to read the instructions and answer example questions. Participation is entirely voluntary.

To prevent bot use and multiple submissions in online experiments, participation may be considered invalid if a response appears to have been generated by a bot, for example, if the task does not seem to have been understood or if it is completed unusually quickly.

To continue, please click the “Next” button below.

F.2 Demographic Information

Please fill in the following information:

- First name:
- Last name:
- Age:
- Is your native language Turkish?
- Gender:
- Education level:

For questions with a drop-down menu, participants selected the appropriate response from the available options.

To continue, please click the “Next” button below.

F.3 Task Instructions

In this experiment, you will be shown several sentences.

After each sentence, you will see a question related to that sentence.

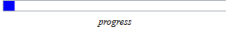
Your task is to provide a rating from 1 to 7 based on the question.

You will make your selections by clicking the numbered boxes.

- 1 = I do not trust it at all / very low
- 7 = I completely trust it / very high

You will first see a few practice questions. After that, the main experiment will begin.

G Participant-facing instruction screens



progress

Bu deneyde size bazı cümleler gösterilecektir.
Her cümleden sonra, o cümleyle ilgili bir soru göreceksiniz.
Göreviniz, verilen soruya göre **1 ile 7 arasında** bir değerlendirme yapmaktır.
Seçimlerinizi numaralı kutulara tıklayarak yapacaksınız.
1 = hiç güvenmiyorum / çok düşük
7 = tamamen güveniyorum / çok yüksek
Önce birkaç alıştırma sorusu göreceksiniz. Ardından asıl deney başlayacaktır.

[Alıştırmalara Geç](#)

Figure 4: Instruction screen shown to participants before the human uptake experiment.



progress

"Kalabalık ortamlarda bulunmak bazı insanlar için yorucu olabilir."

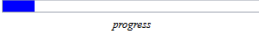
Bu ifadeye göre kalabalık ortamlardan kaçınma ihtimaliniz nedir? (1–7)

1 = çok düşük ihtimal 7 = çok yüksek ihtimal

1 2 3 4 5 6 7

[Devam](#)

Figure 5: Example practice trial shown before the main experiment.



progress

Alıştırmalar tamamlandı!

Şimdi asıl deney kısmına geçiyoruz.
Bu bölümde yanıtlarınız kayıt altına alınacaktır.

Hazırsanız 'Başla' tuşuna basın.

[Başla](#)

Figure 6: Transition screen shown after the practice trials and before the main experiment.