Motive Attribution as a Mechanism for Social Inference in LLMs: Process Alignment Beyond Output Calibration

Anonymous ACL submission

Abstract

Large language models (LLMs) reproduce the qualitative structure of human social inferences but diverge substantially in magnitude calibration. A central open question is *why* calibration quality differs across models. We propose and test a process-level explanation: LLM social judgments are not directly learned output patterns but are mediated by implicit motive attribution. We introduce a two-task paradigm that elicits both social ratings and free-text motive attributions, enabling joint analysis of output and process alignment across eight LLMs. Our results support the motive-mediation hypothesis at two levels. Within each model, motive weighting predicts social rating position across attributes and utterance forms. Across models, the quality of the motive-to-judgment mapping, which captures how closely it mirrors the relative strength of human motive–rating associations, predicts overall calibration deviation. These findings suggest that cross-model calibration differences reflect differences in how models translate implicit motive attributions into social evaluations, rather than differences in surface-level output patterns.

1 Introduction

When a speaker reports a price as *$509.55*, a listener may infer that the speaker is well-prepared and knowledgeable; when the same speaker says *about $500*, listeners may infer either that the speaker prefers casual approximation or that they simply lack precise information, and adjust their social impression of the speaker accordingly. These graded social judgments are not arbitrary: they depend on the listener's inference about *why* the speaker chose this particular form, given the situation and the speaker's plausible knowledge state (Beltrama et al., 2022; Beltrama and Papafragou, 2023; Solt et al., 2025). Across human studies, the same form (precise or approximate) elicits different social evaluations depending on which motive listeners attribute to the speaker's choice: precision attributed to genuine knowledge enhances perceived competence, while precision attributed to pedantry reduces perceived likeability (Solt et al., 2025).

A growing body of work has investigated whether large language models (LLMs) reproduce human pragmatic and social inferences of this kind. LLMs have been shown to recover conversational implicatures (Ruis et al., 2023; Sravanthi et al., 2024; Scherrer et al., 2024), reason pragmatically about scalar expressions (Cho and Kim, 2024) and number words (Tsvilodub et al., 2025), and produce context-sensitive social judgments that align with expert human evaluations (Mittelstädt et al., 2024); see Ma et al. (2025) for a recent survey of the broader LLM-pragmatics evaluation landscape.

A parallel literature has asked whether LLMs can simulate human samples in social science experiments, reproducing population-level patterns (Argyle et al., 2023; Santurkar et al., 2023; Hewitt et al., 2024; Cui et al., 2025; Hullman et al., 2026). Across both lines of work, a consistent pattern emerges: LLMs reliably reproduce the *direction* of human effects but systematically distort their *magnitude*, often by factors of 2–10 (Hewitt et al., 2024; Cui et al., 2025). Mühlenbernd (2026) documented this dissociation for social inferences from numerical (im)precision across three frontier LLMs (see §2 and Appendix B).

These findings raise a follow-up question: *why* do models differ in calibration quality? If all models recover the direction of human social inference but only some recover its magnitude, the differences must lie not in what is learned but in how the underlying inferential process operates. Mühlenbernd (2026) provided indirect evidence for this view: prompting models to reason explicitly about speaker knowledge states and communicative motives reduced calibration deviation in models prone to magnitude inflation, suggesting that motive reasoning is on the critical path to human-like

social judgment. But that study examined motive reasoning only as a prompted intervention; whether models attribute motives unprompted and whether such profiles vary across models in ways that explain differential calibration remained untested.

We close this gap with the **motive-mediation hypothesis**: LLM social judgments are not directly learned output patterns but arise from implicit motive attribution, and the quality of this motive-to-judgment mapping determines both the position of social ratings and the magnitude of calibration deviation. This framing is grounded in empirical evidence that human social inferences from linguistic form are mediated by motive attribution (Beltrama and Papafragou, 2023; Solt et al., 2025). To test it, we introduce a two-task paradigm that elicits both social ratings (Task 1) and free-text motive attributions (Task 2) for the same speaker utterances, enabling joint analysis of output and process alignment across eight LLMs.

Contributions.

- We introduce a two-task paradigm eliciting both social ratings and free-text motive attributions from LLMs, with automated annotation validated against a human gold standard (F1 $\geq$ .83 for primary diagnostic categories).
- We define the Motive Balance Score B_M, a theoretically grounded measure of how a model's motive attribution profile maps onto social attributes, and show it predicts Task 1 ratings across eight chat-format LLMs (pooled $\rho = +.659$, $p < .001$).
- We define the Motive–Rating Mapping Coherence (Mag ρ), a measure of how closely a model's motive–rating association structure mirrors the human benchmark, and show that it predicts cross-model calibration deviation across eight chat-format LLMs ($\rho = -.81$, $p = .015$, $N = 8$).
- We report a theoretically informative null: aggregate motive frequency deviation does not predict calibration, indicating the locus of error is the inference mechanism rather than the attribution surface.

2 Background

2.1 Pragmatic Reasoning and Social Inference

In many sociolinguistic phenomena, listeners' social inferences about a speaker depend not on the chosen form per se but on their inference about *why* the speaker selected that form rather than an available alternative (Burnett, 2019; Acton, 2019; Beltrama, 2020). The choice between *about $500* and *$509.55* is one such case: the two forms are semantically distinct (the approximate covers a wider range of values), yet a speaker who knows the precise value can choose between them. The social inference attached to such a choice is not arbitrarily indexed but derives from the listener's inference about the speaker's plausible knowledge state and the situational demand for precision (Beltrama et al., 2022; Solt et al., 2025). The same form can therefore warrant different social evaluations depending on which motive listeners attribute to its use: precision attributed to genuine knowledge enhances perceived competence, while precision attributed to pedantry reduces perceived likeability. The Rational Speech Act framework (Frank and Goodman, 2012; Goodman and Frank, 2016) formalises this joint inference; its extension to social meaning follows naturally (Beltrama and Papafragou, 2023), with the listener probabilistically inferring which speaker beliefs and goals best explain the observed form choice, and grading the social evaluation accordingly.

Solt et al. (2025) provided the empirical foundation for the present work. Across six everyday dialogue scenarios in a 2×2 design crossing utterance form (precise/approximate) with contextual precision demand (high/low), participants' social ratings on six attributes (competent, well-prepared, knowledgeable, helpful, likeable, pedantic) correlated systematically with the motives they attributed in free-text responses. The specific motive–rating associations established in that study, for example a significant negative correlation between attributing approximation to the speaker's ignorance and rating the speaker as competent ($r \approx -.17$), constitute the empirical impact structure that grounds our motive-mediation framework (see Appendix A).

2.2 Calibration of LLM Social Inferences

Mühlenbernd (2026) established the calibration framework that the present study extends. Three frontier LLMs (GPT, Claude, Gemini) were benchmarked on the Task 1 social ratings of Solt et al. (2025), using two magnitude-sensitive metrics (defined in §4.3): the Effect Size Ratio (ESR), which compares the magnitude of each significant human effect to the corresponding model effect, and the Calibration Deviation Score (CDS), which summarises ESR across all benchmark effects as the

mean absolute deviation from perfect calibration.

The findings revealed a systematic dissociation: all three models reproduced the direction of every human effect, but magnitude calibration varied substantially across architectures, with baseline CDS values ranging from 0.39 (GPT) to 1.48 (Gemini). Knowledge-and-motives-aware (KMA) prompting most consistently reduced calibration deviation, positioning motive reasoning as a candidate mediator of LLM social inference. The present study addresses the follow-up questions left open by Mühlenbernd (2026): whether models attribute motives without explicit prompting, and whether their unprompted motive profiles can explain cross-model calibration differences.

2.3 The Motive-Mediation Hypothesis

We formalise the motive-mediation account introduced in Section 1 as follows. For a model $\mathcal{M}$ evaluating a speaker on attribute a given utterance form v and context c, the social rating does not arise as a direct function of (v, c) but from an implicit attribution of motives $m \in \mathcal{M}_{\text{att}}$, each of which carries an association weight with a in the model's internal representation. Under this view, calibration quality is a property of how a model translates motive attributions into social judgments, not of what motives it attributes or how often.

To operationalise this view, we introduce two complementary metrics. The first quantifies, within each model, whether motive attribution predicts social rating position. The Motive Balance Score combines a model's attribution rates with the empirical impact of each motive on human social judgment:

$$B_M(a, v) = \sum_{m \in \mathcal{M}_{\text{att}}} r_H(m, a, v) \cdot \hat{p}(m \mid v), \quad (1)$$

where $r_H(m, a, v)$ is the empirical human point-biserial correlation between motive m and attribute rating a for utterance form v (Solt et al., 2025), and $\hat{p}(m \mid v)$ is the model's attribution rate for motive m given form v. Intuitively, $B_M(a, v)$ captures the net signal that a model's motive profile generates for attribute a, weighted by the empirical strength with which each motive raises or lowers a in human judgments. Under the motive-mediation hypothesis, $B_M(a, v)$ should positively predict the model's observed Task 1 rating for a on form v, across attributes and forms.

The second metric quantifies the structural similarity between a model's motive–rating mapping and the human one. Let $r_M(m, a, v)$ denote the point-biserial correlation between motive m and the rating on attribute a for form v, computed on the model's ratings, and let $r_H(m, a, v)$ denote the corresponding human correlation. The Motive–Rating Mapping Coherence is then defined as

$$\begin{aligned} \text{Mag}\,\rho = \rho_{\text{Sp}}\big(&\{r_M(m, a, v)\}_{(m,a,v)\in\mathcal{S}}, \\ &\{r_H(m, a, v)\}_{(m,a,v)\in\mathcal{S}}\big), \end{aligned} \quad (2)$$

where ρ_{Sp} is the Spearman rank correlation and $\mathcal{S} = \{(m, a, v) : p_H(m, a, v) < .10,\ n_+ \geq 5\}$ is the set of motive–attribute–form cells with significant human correlations and at least five positive observations of motive m on form v. High Mag ρ indicates that the model's motive–rating associations preserve the relative strength structure observed in humans, even where absolute values differ. Under the motive-mediation hypothesis, models with higher Mag ρ should show lower calibration deviation on Task 1 ratings (lower CDS), since faithful preservation of the human motive–rating structure is what produces human-like ratings.

3 Hypotheses

H1 (Motive-mediated social inference, within-model). Within each model, the motive balance score $B_M(a, v)$ (Eq. 1) positively predicts observed Task 1 social ratings across attributes and utterance forms. Models that over-weight motives positively associated with an attribute (per the human benchmark) rate that attribute higher; models that under-weight them rate lower.

H2 (Motive–rating mapping quality predicts calibration, cross-model). Models whose motive–rating mapping more closely mirrors the human benchmark in relative magnitude (higher Mag ρ, Eq. 2) show lower calibration deviation (lower CDS).

Theoretically informative null. We additionally test whether motive frequency deviation (MFD, Eq. 3) predicts CDS. A null result for MFD vs. CDS supports the motive-mediation account over a simpler direct-attribution account in which more human-like motive frequencies should produce more human-like ratings.

4 Method

4.1 Human Benchmark

The human benchmark is Experiment 1 of Solt et al. (2025), which tests whether the social meaning of

Scenario	Jamie talks to an **insurance agent** [highPr] / **friend** [lowPr] who asks about the price of his bike. Jamie answers: "**The bicycle cost $496.99**" [precise] / "**about $500**" [approx].
Task 1	Based on what Jamie says, Jamie sounds: not [a] at all ○ ○ ○ ○ ○ ○ ○ very [a] where [a] stands for each of the six social attributes rated together: *competent*, *well-prepared*, *knowledgeable*, *helpful*, *likeable*, *pedantic*.
Task 2	Why do you think Jamie used "$496.99" rather than "about $500"? (Or vice versa for the approx condition.) Please explain in your own words.

Table 1: Example stimulus and task structure for the bicycle scenario. Bracketed labels indicate experimental conditions; each participant saw a single combination of context (highPr/lowPr) and form (precise/approx).

numerical (im)precision derives from pragmatic reasoning about speaker knowledge and communicative motives, or is fixed by the linguistic form itself. The study employs a 2×2 fully between-subjects matched-guise design ($N = 362$) crossing utterance form (precise vs. approximate) with contextual precision requirements (high-precision [highPr] vs. low-precision [lowPr]) across six everyday dialogue scenarios (absence, bike, cinema, conference, cooking, house). Table 1 shows a sample stimulus and the Task 1 and Task 2 structure for the bicycle scenario. Each participant saw a single scenario in a single condition and completed all tasks for that scenario.

Three tasks were collected per participant. Task 1 (Table 1) elicited 7-point Likert ratings on six social attributes (*competent*, *well-prepared*, *knowledgeable*, *helpful*, *likeable*, *pedantic*), rated together for the same speaker. Task 2 asked participants to describe in free text why the speaker had chosen the observed form rather than the alternative; responses were manually annotated into a 10-category motive taxonomy developed by Solt et al. (2025): NOTKNOW, KNEWEXACT, REALWORLD, FUSSYPERSON, NEEDED, INFONOTAVAIL, PURPOSE, INFOAVAIL, USUALAPX, and EASYSPEAKER. Detailed category descriptions and examples are provided in Appendix A. Task 3 presented a fixed list of 10 multiple-choice options for the same motivation question; Task 3 results are reported in Appendix I.

Solt et al. (2025) report ten statistically significant Task 1 effects from this design: five main effects of form and five form $\times$ context interactions, all directionally consistent with motive-mediated social inference (see Appendix A for details). From the publicly available data we additionally computed point-biserial correlations between Task 2 motive annotations and Task 1 ratings, separately by form; 25 motive–attribute–form cells reach significance ($p < .10$, $n_+ \geq 5$) and constitute the empirical impact structure that defines $r_H(m, a, v)$ in Eq. 1 (see Appendix C for the full table).

4.2 LLM Experiment

Conversation structure. Each run is a two-turn conversation mirroring the human paradigm. Turn 1 elicits a Task 1 rating (integer 1–7) on all six social attributes simultaneously, using the minimal prompt adapted from Mühlenbernd (2026) to the present scenario set. Turn 2 appends the Task 2 free-text motive attribution prompt verbatim from Solt et al. (2025), asking why the speaker chose the precise (or approximate) form rather than the alternative. Eliciting Task 2 within the same conversation as Task 1 preserves the sequential structure of the human experiment and ensures that motive attributions are conditioned on the same rating context.

Design. Six scenarios $\times$ two forms $\times$ two contexts $\times$ 30 runs per cell, yielding 720 conversations per model. Temperature was set to $\tau = 1.8$ for all models except Claude, which was run at $\tau = 1.0$ due to the Anthropic API temperature ceiling. Both settings were selected to approximate the human within-condition Task 1 rating SD (1.38 approx / 1.19 precise, computed from `experiment1.csv`; see Appendix E); $\tau = 1.0$ is the closest available setting for Claude and produces somewhat lower rating variance. Achieved within-condition SDs and mean motive counts per model are reported in Appendix E; the CDS metric operates on condition means and is not materially affected by the variance difference.

Models. Of the twelve LLMs we attempted to test, two smaller models (Llama 3.1 8B and Phi-4 mini) were excluded during data collection by runtime quality monitoring: both produced Task 2 truncation rates of $\geq 70\%$ at $\tau = 1.8$, indicating that models below ~7B parameters cannot reliably generate coherent free-text motive attributions at this temperature. The remaining ten models yielded complete datasets: GPT (GPT-4o-mini), Llama (Llama 3.3 70B), Gemini (Gemini 2.5 Pro), Claude (Claude Sonnet), Qwen (Qwen 2.5 72B), Gemma (Gemma 4), Mistral

(Mistral 7B Instruct v0.1), DeepSeek (DeepSeek-V3), Qwen-7B (Qwen 2.5 7B), and DeepSeek-R1 (DeepSeek-R1). Short names are used consistently in all tables and figures.

Format-based exclusion. Post-hoc output-format diagnostics on the ten completed datasets identified two models, Gemini and DeepSeek-R1, whose Task 2 responses were dominated by structured bullet-point templates with recurring signature phrases. Notably, these are the only two RL-trained reasoning models in our sample, an observation we discuss in §6. This formatting regime mechanically constrains within-cell motive-attribution variance and compromises Mag ρ measurement; full diagnostics are reported in Appendix F. Both models are excluded from the primary analyses on these grounds. We report results on the resulting eight chat-format LLMs (Claude, Gemma, Qwen, Llama, GPT, Mistral, DeepSeek, Qwen-7B) as primary; the full ten-model result is reported as a robustness check (§5), and descriptive analyses of the excluded models are reported in Appendix G.

Post-processing. Truncated Task 2 responses and invalid Task 1 ratings are excluded from their respective analyses; per-model exclusion rates and a sensitivity analysis for DeepSeek (which shows an elevated Task 1 exclusion rate due to JSON-format leakage) are reported in Appendix D. Free-text Task 2 responses were annotated using an LLM-based annotator validated against human gold labels (F1 $\geq$ .83 for primary diagnostic categories). The USUALAPX category is excluded from all LLM analyses due to a model-invariant annotation artefact; three further categories (REALWORLD, NEEDED, PURPOSE) show false-positive bias on precise speakers and are interpreted with caution. Full validation, tier classification, and exclusion reasoning are provided in Appendix D.

4.3 Metrics

Calibration Deviation Score (CDS). We use the CDS metric of Mühlenbernd (2026), defined as the mean absolute deviation from perfect calibration across ten benchmark effects: $\text{CDS} = \frac{1}{10}\sum_{i=1}^{10} |\text{ESR}_i - 1|$, where $\text{ESR}_i = |\Delta_M^{(i)}|/|\Delta_H^{(i)}|$ is the Effect Size Ratio comparing model and human effect magnitudes for the i-th benchmark effect. Lower CDS indicates closer alignment to human effect magnitudes overall.

Motive Frequency Deviation (MFD). MFD quantifies how far a model's overall motive attribution profile deviates from the human profile, treating the nine retained motive categories as a distribution. For each form v, let $\mathbf{p}_H^{(v)}$ and $\mathbf{p}_M^{(v)}$ denote the vectors of human and model motive attribution rates. MFD is defined as

$$\text{MFD} = 1 - \frac{1}{2} \sum_{v \in \{\text{appr,prec}\}} \rho_{\text{Sp}}\left(\mathbf{p}_H^{(v)}, \mathbf{p}_M^{(v)}\right), \quad (3)$$

the mean rank-order distance between human and model motive rate vectors, averaged over the two forms; higher values indicate greater profile-level divergence. MFD enters the H2 analysis as an alternative predictor of CDS, testing whether profile-level deviation is sufficient to explain calibration quality (see §5.2).

Motive Balance Score (B_M). $B_M(a, v)$ as defined in Eq. 1 is computed per attribute $\times$ form cell, restricted to the 25 active human motive–attribute correlations identified in Appendix C. To test H1, we compute the Spearman rank correlation between $B_M(a, v)$ values and the model's observed Task 1 mean ratings across all attribute $\times$ form cells, separately per model and pooled across models. Higher correlations indicate stronger motive-mediated prediction of social rating position.

Motive–Rating Mapping Coherence (Mag ρ). Mag ρ as defined in Eq. 2 is computed per model. To test H2, we compute the Spearman rank correlation between Mag ρ and CDS across the eight models. A negative correlation indicates that models with more human-like motive–rating mapping structures show lower calibration deviation.

5 Results

5.1 H1: Motive-Mediated Social Inference (Within-Model)

Table 2 reports Spearman and Pearson correlations between $B_M(a, v)$ and observed Task 1 mean ratings across all attribute $\times$ form cells, separately per model and pooled.

H1 is strongly supported. Six of eight models show a significant positive correlation between their motive balance score and their social ratings (pooled $\rho = +.659$, $p < .001$, robust to outliers). Mistral and Qwen-7B show positive but non-significant trends ($\rho = +.464$, $p = .151$ and $\rho = +.536$, $p = .089$ respectively), both consistent with H1 in direction. Figure 1 shows the 88

Model	N	Spearman ρ	p	Pearson r
Claude	11	+.764	.006	$+.824^{**}$
Gemma	11	+.727	.011	+.524
Qwen	11	+.818	.002	$+.748^{**}$
Llama	11	+.755	.007	$+.635^{*}$
GPT	11	+.609	.047	$+.769^{**}$
Mistral	11	+.464	.151	$+.705^{*}$
DeepSeek	11	+.782	.004	$+.643^{*}$
Qwen-7B	11	+.536	.089	+.560
Pooled	88	+.659	$< .001$	$+.611^{***}$

Table 2: Spearman and Pearson correlations between $B_M(a, v)$ and observed Task 1 mean ratings ($N = 11$ active attribute $\times$ form cells per model; $N = 88$ pooled). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Model	CDS	MFD	Mag ρ
DeepSeek	0.393	.103	.878
Claude	0.549	.152	.747
Gemma	0.590	.164	.777
Qwen	0.595	.092	.679
Llama	0.611	.077	.597
GPT	0.708	.124	.685
Qwen-7B	0.713	.146	.186
Mistral	0.719	.118	.643
ρ vs CDS		−.10	−.81
p		.82	.015

Table 3: CDS (calibration deviation, MFD (motive frequency deviation), and Mag ρ (motive–rating mapping coherence) for the eight LLMs, ordered by CDS. Cross-model Spearman ρ against CDS shown in final rows.

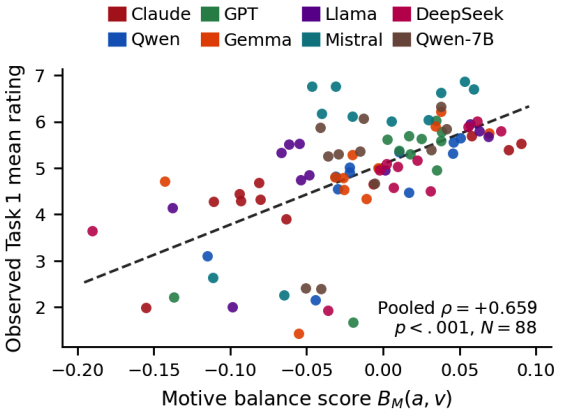

Figure 1: $B_M(a, v)$ vs. observed Task 1 mean rating across 88 attribute $\times$ form cells from eight chat-format models (colour coded). Dashed line: pooled OLS trend.

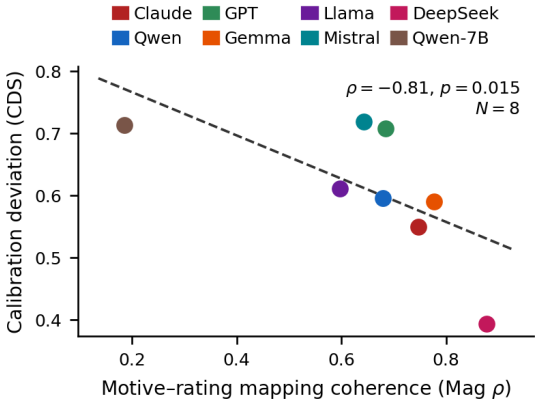

Figure 2: Motive–rating mapping coherence (Mag ρ) vs. calibration deviation (CDS) across the eight chat-format models. Dashed line: OLS trend.

data points with the pooled trend line; the positive relationship is visible within each model colour, not merely as a between-model effect.

That B_M predicts rating position across diverse attributes, both forms, and eight chat-format LLMs provides strong evidence for motive mediation in LLM social inference. A supplementary analysis of motive balance *deviation* (ΔB) against Task 1 rating deviation is reported in Appendix H: B_M predicts rating position strongly but the direction of miscalibration only weakly, indicating that the within-model motive–rating mapping is a better diagnostic of rating position than of calibration error direction.

5.2 H2 and the Null Result for Motive Frequency Deviation (Cross-Model)

Table 3 reports CDS, MFD, and Mag ρ for the eight LLMs, with cross-model Spearman correlations against CDS testing H2 (predicted negative for Mag ρ) and the null (predicted null for MFD).

H2 supported. Mag ρ negatively correlates with CDS across the eight chat-format models ($\rho = -.81$, $p = .015$, $N = 8$): models with more human-like motive–rating mapping structures are better calibrated overall (Figure 2). DeepSeek anchors the high end of the trend (highest Mag ρ, lowest CDS); Qwen-7B anchors the low end (lowest Mag ρ, second-highest CDS). The result survives every leave-one-out subset ($\rho \in [-.71, -.89]$; full LOO in Appendix H). Including the two format-excluded reasoning models as a robustness check, the relationship persists at $\rho = -.64$, $p = .048$ ($N = 10$).

The null result. By contrast, MFD shows essentially zero correlation with CDS ($\rho = -.10$, $p = .82$). This null is theoretically informative: having a more human-like motive attribution profile does not by itself produce better calibrated social judgments. The clearest illustration is the pair Claude and Mistral: Claude is the second-best

calibrated model overall but has the most deviant NOTKNOW rate among retained models (91.4% vs. human 50.8%), while Mistral is among the worst-calibrated yet has near-human NOTKNOW rate (48.3%). The locus of calibration error is therefore not the attribution surface (which motives are attributed and how often) but the inference mechanism (how those motives are translated into social ratings; full motive rates in Appendix H).

Form-exclusivity is respected. As a methodological precondition for H2, all eight chat-format models correctly respect form-exclusivity: motives logically applicable to only one form receive zero attribution for the other form (Appendix H). The motive-mediated framework therefore operates over a structurally well-formed motive space.

6 Discussion

6.1 Motive Attribution as Implicit Social Inference

The convergent evidence across both hypotheses supports a view of LLM social inference as a motive-mediated process. Six of eight chat-format models show significant within-model correlations between B_M and Task 1 ratings (H1, pooled $\rho = +.659$, $p < .001$), and the cross-model analysis reveals that the quality of the motive-to-judgment mapping, not the motive attribution profile per se, determines calibration quality (H2, $\rho = -.81$, $p = .015$).

The null result for MFD is integral to this interpretation rather than a separate finding. As the Claude / Mistral contrast in §5.2 illustrates, what separates well-calibrated from poorly-calibrated models is not how often they attribute particular motives but how those motives are mapped onto social judgments, captured by Mag ρ. This distinction between attribution surface and inference mechanism locates the source of LLM calibration differences in the inferential process itself, not in the motives that surface in free-text explanations.

This finding connects to broader questions about LLM social cognition (Binz and Schulz, 2023; Aher et al., 2023; Tsvilodub et al., 2025). Our results suggest that LLMs engage in something structurally analogous to pragmatic reasoning about speaker intent, producing motive–rating associations that mirror the human structure in relative organisation even when they diverge in absolute rates or magnitudes.

6.2 LLM Verbosity and Normalised Motive Distributions

A methodologically important finding concerns the *number* of motives per response. Humans select on average $\sim$0.9 motives per response, effectively choosing a single primary explanation. LLMs are substantially more verbose, with motive counts ranging from $0.7\times$ to $3.3\times$ the human rate (Appendix E). Direct comparison of raw motive attribution rates is therefore confounded, and we normalise rates by the ratio of human to model mean motive counts before comparing distributions (raw rates: Table 14, Appendix H).

Two systematic patterns survive normalisation. First, all eight chat-format models over-attribute REALWORLD for approximate speakers (normalised rates .10–.22 vs. human .07); for precise speakers, four models reach near-human levels (Claude, Gemma, Llama, Qwen-7B; within $\pm.05$ of human .23), while GPT shows the strongest over-attribution (.42) and Qwen, Mistral, and DeepSeek remain moderately elevated (.27–.32). Second, all eight chat-format models under-attribute NOTKNOW for approximate speakers (human .51; models .23–.48), independent of verbosity.

Together these patterns reveal a coherent tendency that we describe as the *articulable-explanation bias*. When a speaker says *about $500*, humans frequently attribute this to the speaker's lack of exact knowledge (NOTKNOW); LLMs more often attribute it to situational features (REALWORLD, NEEDED). The former is a mental state inference, reasoning about what the speaker *knows*. The latter is a situational inference, reasoning about what the situation *requires*. LLMs preferentially generate the situational class and under-produce the epistemic class.

6.3 Implications and Future Directions

Mag ρ as a process-level diagnostic. Output-level metrics (CCC, RMSE, CDS) measure how much a model's predictions deviate from human predictions at the surface. They do not localise the source of deviation. The same overall CDS could arise from qualitatively different inferential failures: a poor motive attribution profile, a poor motive-to-judgment mapping, or both. Mag ρ separates these by measuring process-level structural similarity between model and human motive–rating associations. A model with low CDS but low Mag ρ would be one that happens to land near human rat-

ings via non-human-like processes, a diagnostic distinction that output-level metrics cannot make (an empirical instance is discussed in the reasoning-models paragraph below). We propose Mag ρ as a complement to existing output-level metrics in future evaluations of LLM pragmatic and social inference.

Implications for training and fine-tuning. The articulable-explanation bias and the cross-model variation in Mag ρ both point toward concrete directions for improving LLM social inference. At the training-data level, corpora enriched with explicit epistemic state reasoning, theory-of-mind inferences, and first-person uncertainty articulation (*"I wasn't sure exactly"*, *"I figured it must be"*) could shift attribution profiles toward the human balance between epistemic and situational explanations. At the fine-tuning stage, motive attribution could enter as auxiliary supervision: models could be trained to predict motive labels alongside social judgments, with loss terms penalising departures from empirical human motive–rating correlations. More directly, Mag ρ itself could serve as a training signal or evaluation milestone during instruction-tuning or RLHF, with the optimisation target being to maximise structural similarity between the model's emerging motive–rating mapping and the human benchmark. Prompting-based interventions such as KMA are effective short-term patches but address symptoms rather than underlying limitations.

Reasoning models and templated motive attribution. A striking pattern in our data is that the two models excluded on output-format grounds (Gemini and DeepSeek-R1) are the only two RL-trained reasoning models in our sample, while the remaining eight chat-format models—spanning a comparable range of architectures and sizes—all produced free-prose Task 2 responses without bullet-point templating. With $N = 2$ the observation is suggestive rather than conclusive, but the pattern is consistent with a possible side-effect of reasoning-oriented training: the structured chain-of-thought style cultivated by RL on reasoning tasks may carry over into free-text generation, mechanically constraining within-cell motive variance and partially decoupling motive attribution from rating production. Notably, DeepSeek-R1 achieves the lowest CDS in our full ten-model sample despite only mid-range Mag ρ (Appendix G), consistent with a calibration route that operates through structured argumentation rather than fine-grained motive–rating geometry. Whether this alternative mechanism generalises across reasoning-trained models—and whether the H2 relationship reappears in reasoning models that produce free-prose rather than templated output—remains an open empirical question.

Future directions. The motive-mediation framework opens several lines of further investigation. *Causal validation*: prompting interventions that selectively suppress or amplify motive attributions and observe the predicted rating shifts would establish causal mediation. *Domain extension*: numerical precision is one of many sociolinguistic phenomena where social meaning derives from motive attribution, with politeness, register, modality, and evidentiality sharing the same form-meaning-motive structure. Moreover, adjacent work on whether LLMs detect the motives of biased *sources* (Wu et al., 2025) raises the question of how source- and speaker-motive inferences relate. *Cross-lingual generalisation*: whether the articulable-explanation bias appears symmetrically across typologically diverse languages remains an open question, particularly for those with grammaticalised epistemic marking. *Mechanistic probing*: probing analyses could test whether the behavioural patterns we observe correspond to identifiable representational structures inside the models.

7 Conclusion

We proposed and tested the motive-mediation hypothesis: that LLM social judgments arise from implicit motive attribution processes that vary systematically across models. Two findings support this account. First, in six of eight chat-format models, the motive balance score B_M significantly predicts observed social rating position (pooled $\rho = +.659$, $p < .001$). Second, across models, the quality of the motive-to-judgment mapping predicts calibration deviation ($\rho = -.81$, $p = .015$, $N = 8$). A theoretically informative null, namely that aggregate motive frequency deviation does not predict calibration, locates the source of calibration differences in the inference mechanism itself rather than the attribution surface. The motive–rating mapping coherence Mag ρ thus offers a process-level diagnostic that complements output-level calibration metrics: improving LLM social inference requires not only better motive attribution profiles, but better motive-to-judgment translation.

8 Limitations

Scope of evidence. The H2 cross-model correlation ($\rho = -.81$, $p = .015$, $N = 8$) is statistically significant and robust to single-model deletion (leave-one-out range $\rho \in [-.71, -.89]$; bootstrap 95% CI $[-1.00, -.27]$, 99.2% of 10,000 resamples negative; permutation test one-sided $p = .011$). Nevertheless, the sample of eight chat-format models remains modest, and the upper bound of the bootstrap CI approaches zero, so reviewers should treat the headline magnitude as indicative rather than definitive. Of the twelve LLMs we attempted to test, two smaller models (Llama 3.1 8B and Phi-4 mini) were excluded during data collection by runtime quality monitoring (Task 2 truncation rates $\geq 70\%$ at $\tau = 1.8$), and two further models (Gemini and DeepSeek-R1) were excluded post-hoc on output-format grounds (Appendix F). The post-hoc nature of the second exclusion is acknowledged; we report the full ten-model result as a robustness check ($\rho = -.64$, $p = .048$, $N = 10$; Appendix G). Future work could explore lower-temperature protocols for smaller models and broader sampling of mid-scale architectures, though both would require separate temperature calibration. The analysis is also restricted to one domain (numerical precision) and one language (English); whether the motive-mediation account generalises to other sociolinguistic phenomena and to typologically diverse languages remains to be tested. The human benchmark itself derives from a single experimental study (Solt et al., 2025); replication of the 25 active motive–attribute–form correlations on independent samples would strengthen the empirical foundations of the impact structure.

Methodological choices. Temperature was set to $\tau = 1.8$ for all models except Claude, which was run at $\tau = 1.0$ due to the Anthropic API ceiling. Although CDS operates on condition means and is robust to within-condition variance differences, substantive distributional differences between Claude and the other seven chat-format models cannot be ruled out. DeepSeek additionally shows an elevated Task 1 invalid-response rate (21.7% vs. $\leq 7.4\%$ for the other retained models), driven by occasional JSON-format leakage from the Task 1 prompt template into the Task 2 slot; we verified that excluding the leaked responses leaves Mag ρ effectively unchanged (within $< .01$ of the primary value .878; full sensitivity analysis in Appendix D), but the elevated rate is a methodological irregularity worth flagging. Our Task 2 annotation pipeline uses an LLM-based annotator validated against human gold labels; F1 reaches $\geq .83$ for primary diagnostic categories, but three categories (REALWORLD, NEEDED, PURPOSE) show systematic false-positive bias on precise speakers, so conclusions about those categories rely on directional rather than absolute rate comparisons (see Appendix D). The verbosity normalisation procedure in §6 represents one defensible choice among alternative normalisation schemes; the specific patterns reported as surviving normalisation could in principle be sensitive to that choice. An earlier version of this work claimed that Mag ρ values were rating-source-independent (rank-equivalent when computed against human rather than model-own ratings); on re-verification with the present pipeline that equivalence does not cleanly hold (rank correlation between the two orderings: $\rho = +.71$, not $+1.00$), and the claim has been removed. Mag ρ is therefore to be interpreted as a measure of the coherence of a model's *internal* motive-to-rating mapping, not as a rating-source-independent quantity.

Interpretation. Our evidence for the motive-mediation hypothesis is correlational. H1 shows that B_M predicts rating position within models; H2 shows that Mag ρ predicts calibration across models. Neither establishes causal mediation: we have not intervened on motive attributions and observed the resulting change in ratings. Causal mediation could be tested through prompting manipulations that selectively suppress or amplify specific motive attributions and observe whether ratings shift in the directions predicted by $r_H(m, a, v)$. More broadly, our analysis treats free-text motive attributions as observable proxies for internal processes we cannot directly access; whether the motive-mediation account corresponds to any specific internal mechanism is an open question, and cross-model variation in Mag ρ may reflect differences in training corpora, fine-tuning objectives, or instruction-tuning procedures that are confounded with architecture in the present design. The observation that the two excluded reasoning models (Gemini and DeepSeek-R1) both produced templated bullet-point outputs (§6.3, Appendix F) is based on $N = 2$ and should be treated as suggestive rather than as a generalisation about reasoning-trained models; whether the templating–calibration relationship we observe in DeepSeek-R1 (lowest CDS, only mid-range Mag ρ)

generalises to a wider set of reasoning architectures, or whether non-templated reasoning models would re-enter the chat-model Mag ρ–CDS regime, requires further empirical work.

Mechanism of the articulable-explanation bias. The articulable-explanation bias documented in §6—that LLMs preferentially generate situational over epistemic explanations for approximate utterances—has a plausible origin in training data that we cannot directly test. Articulable situational explanations are well-represented in written text: news articles, manuals, and explanatory writing routinely cite situational reasons for linguistic choices. Mental state attributions are less reliably articulated; humans frequently *infer* a speaker's epistemic state without writing it down, and where such inferences do appear in text they are often implicit rather than spelled out. The imbalance between articulable and implicit pragmatic reasoning in written text may propagate into LLM attribution profiles as a systematic preference for the former. We flag this as a speculative mechanism: it is consistent with our observations but would require direct empirical investigation (e.g., probing training corpora for the relative prevalence of epistemic vs. situational attribution language) to establish.

Ethics Statement

This work uses publicly available human data from Solt et al. (2025), collected under that study's own ethical approval procedures; no new human-subjects data was collected for the present paper.

LLM generations were obtained via standard commercial APIs (OpenAI, Anthropic, Google) and open-weight model deployments (Qwen, Llama, Mistral, Gemma, DeepSeek) at reasonable computational cost; the full set of 7,200 Task 1 and 7,200 Task 2 generations was produced within standard research-account quotas. No prompts solicited content involving real individuals, sensitive demographic attributions, or dual-use technical knowledge.

The study analyses social-inference patterns in pragmatic reasoning about numerical precision; it does not constitute or facilitate model alignment for deployment, nor does it raise sensitive-group fairness concerns. The descriptive analyses of Gemini and DeepSeek-R1 in Appendix G report aggregate output-format and calibration statistics from publicly available API responses and do not identify any specific user, prompt, or output.

We used AI assistance (Claude) for prose editing, code review, and analysis support during the preparation of this work; all scientific content and reported numbers were authored and verified by the human authors.

Data and Code Availability

All Task 1 and Task 2 data for the ten LLM model runs reported in this paper ($N = 7{,}200$ conversations), the LLM-based annotation pipeline and its validation against human gold-standard labels, and the analysis code reproducing every reported number, table, and figure are available at `https://anonymous.4open.science/r/llm-social-inference-7097/`. The human benchmark data (Solt et al., 2025) is publicly available at `https://osf.io/m4rhn/` (DOI: 10.17605/OSF.IO/M4RHN). Upon acceptance, the anonymous mirror will be replaced with a public repository under a permissive license (CC-BY 4.0 for data, Apache 2.0 for code).

References

Eric K. Acton. 2019. Pragmatics and the social life of the English definite article. *Language*, 95(1):37–65.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human studies. In *Proceedings of ICML 2023*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Andrea Beltrama. 2020. Social meaning in semantics and pragmatics. *Language and Linguistics Compass*, 14(3):e12370.

Andrea Beltrama and Anna Papafragou. 2023. We are what we say: Pragmatic violations inform speaker inferences. *Glossa Psycholinguistics*, 2(1).

Andrea Beltrama, Florian Schwarz, and Anton Benz. 2022. Being precise or being kind: How precision shapes social meaning. *Glossa*, 7(1).

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6).

Heather Burnett. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 42:419–450.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Ziyan Cui, Ning Li, Huaikang Zhou, and 1 others. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336:998.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. Working paper, New York University.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2026. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. Preprint, arXiv:2602.15785.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Justin M. Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14:27449.

Roland Mühlenbernd. 2026. Social meaning in large language models: Structure, magnitude, and pragmatic prompting. In *Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026*, pages 162–171. ELRA Language Resources Association (ELRA).

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 20827–20905.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cindy He, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Stephanie Solt, Roland Mühlenbernd, and Mariya Burbelko. 2025. Social meaning and pragmatic reasoning: The case of (im)precision. In *Proceedings of ELM 3*, pages 371–382.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097.

Polina Tsvilodub, Kanishk Gandhi, Haoran Zhao, Jan-Philipp Fränken, Michael Franke, and Noah D. Goodman. 2025. Non-literal understanding of number words by language models. In *Proceedings of the 47th Annual Conference of the Cognitive Science Society*.

Addison J. Wu, Ryan Liu, Kerem Oktar, Theodore R. Sumers, and Thomas L. Griffiths. 2025. Are large language models sensitive to the motives behind communication? In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.

A Human Benchmark: Solt et al. (2025)

Citation. Solt, S., Mühlenbernd, R., & Burbelko, M. (2025). Social meaning and pragmatic reasoning: The case of (im)precision. *Proceedings of ELM 3*, 371–382.

Design. A 2×2 between-subjects matched-guise experiment ($N = 362$): utterance form (precise/approximate) $\times$ context (highPr/lowPr) across six dialogue scenarios (absence, bike, cinema, conference, cooking, house). Each participant saw a single scenario in a single condition and completed all tasks for that scenario. Experiment 1 (base knowledge condition) is the benchmark for the present study; Experiment 2 (established knowledge condition) is not used.

Tasks. Table 1 in the main text shows an example scenario with the Task 1 and Task 2 prompt structures. Task 1: 7-point Likert ratings on six social attributes (*competent*, *well-prepared*, *knowledgeable*, *helpful*, *likeable*, *pedantic*). Task 2: free-text motive attribution, manually annotated into a 10-category taxonomy. Task 3: 10-option multiple-choice motive selection; reported in Appendix I.

Motive taxonomy (Task 2). The 10 categories with definitions and example responses follow the annotation scheme of Solt et al. (2025):

Approx	competent	well-prepared	knowledgeable	helpful	likeable	pedantic
NOTKNOW	$-.17^{*}$	$-.16^{*}$	$-.14.$	$-.17^{*}$	—	$-.17^{*}$
INFONOTAVAIL	$+.16^{*}$	$+.21^{**}$	$+.21^{**}$	$+.16^{*}$	—	—
REALWORLD	$+.12.$	$+.12.$	—	$+.14.$	—	—
PURPOSE	—	—	—	—	$+.14.$	$+.15^{*}$
EASYSPEAKER	$-.14.$	—	$-.12.$	—	$-.13.$	—
Precise						
KNEWEXACT	$+.14.$	—	$+.15^{*}$	—	—	—
NEEDED	—	$+.13.$	—	—	—	$-.20^{**}$
FUSSYPERSON	—	—	—	—	$-.24^{**}$	$+.15^{*}$
PURPOSE	—	—	—	—	$+.20^{**}$	—
REALWORLD	—	—	—	—	—	$-.14.$

Table 4: Active human motive–attribute correlations by form. $^{**}p < .01$, $^{*}p < .05$, $.\ p < .10$; — denotes $p \geq .10$ or $n_{+} < 5$. The 25 active cells define the impact weights $r_H(m, a, v)$ in the motive balance score B_M (Eq. 1).

- **NOTKNOW**: speaker did not know or was not sure about the exact value (e.g., *"She probably didn't know the exact number"*). Approx-only.
- **KNEWEXACT**: speaker knew or was confident about the exact value (e.g., *"He clearly knew the exact figure"*). Precise-only.
- **REALWORLD**: the real-world consequences of the communicated value were relevant (e.g., *"The exact amount matters for the insurance claim"*).
- **FUSSYPERSON**: speaker is pedantic, detail-oriented, or fussy (e.g., *"He sounds like the kind of person who obsesses over details"*).
- **NEEDED**: the level of precision was situationally needed or appropriate (e.g., *"Approximation is fine in this casual context"*).
- **INFONOTAVAIL**: precise information was not available to the speaker (e.g., *"The receipt wasn't to hand"*). Approx-only.
- **PURPOSE**: speaker was purposefully precise or approximate (e.g., *"She rounded it on purpose to keep things simple"*).
- **INFOAVAIL**: precise information was available to the speaker (e.g., *"He must have looked it up"*). Precise-only.
- **USUALAPX**: speaker usually uses approximate forms as a stylistic habit (e.g., *"This person tends to round numbers in conversation"*). Approx-only.
- **EASYSPEAKER**: using this form was easier or faster for the speaker (e.g., *"It's quicker to just say 'about $500' "*). Approx-only.

Five categories are form-exclusive: NOTKNOW, INFONOTAVAIL, USUALAPX, EASYSPEAKER apply only to approximate speakers, and KNEWEXACT, INFOAVAIL only to precise speakers. The remaining categories appear for both forms but typically at significantly different rates.

Key Task 1 results. Main effects of form ($p < .001$): precise $>$ approx on all attributes except likeable. Form $\times$ context interactions ($p < .001$): advantage of precise over approx is larger in highPr than lowPr for competent, well-prepared, helpful, likeable, knowledgeable. A detailed statistical analysis of these effects is provided in Solt et al. (2025).

Key Task 2 results. Dominant motive for approximate speakers: NOTKNOW (50.8%). Dominant motives for precise speakers: REALWORLD (23.2%), KNEWEXACT (21.5%), FUSSYPERSON (15.8%).

Motive–rating correlations (impact structure). Selected significant point-biserial correlations between motives and Task 1 ratings: NOTKNOW (approx) $\times$ competent: $r \approx -.17^{*}$ (ignorance attribution penalises competence); INFONOTAVAIL (approx) $\times$ competent: $r \approx +.16^{*}$ to $+.21^{**}$ (situational unavailability does not penalise); FUSSYPERSON (precise) $\times$ likeable: $r = -.24^{**}$ (pedantry attribution lowers likeability); PURPOSE (precise) $\times$ likeable: $r = +.20^{**}$ (intentionality attribution raises likeability). The full set of significant correlations defining the impact structure used in B_M (Eq. 1) is given in Appendix C.

B Mühlenbernd (2026) Summary

Citation. Mühlenbernd, R. (2026). Social meaning in large language models: Structure, magnitude, and pragmatic prompting. *Proceedings of CMCL 2026*. arXiv:2604.02512.

Core finding. LLMs reproduce the qualitative structure of human social inferences but diverge

in magnitude calibration. KMA prompting most consistently reduces CDS.

Key CDS results (MIN baseline, $\tau = 1.0$). GPT: 0.393; Claude: 0.538; Gemini: 1.480. Note: CDS values in the present study ($\tau = 1.8$) are not directly comparable due to temperature-dependent variance effects.

Role in this study. Establishes the output-level calibration baseline and the CDS metric. The motive-mediation hypothesis provides a process-level explanation for the cross-model CDS ordering observed by Mühlenbernd (2026).

C Empirical Impact Sign Table

Table 4 reports the human point-biserial correlations between motive binary variables and Task 1 ratings (computed from `experiment1.csv`), separately by form. Only cells with $p < .10$ and $n_+ \geq 5$ positive observations are treated as active and enter the B_M computation (25 active cells total: 17 for approximate speakers, 8 for precise speakers).

D Annotation Procedure and Validation

Exclusion rates. Truncated Task 2 responses and invalid Task 1 ratings are excluded from their respective analyses. Per-model exclusion rates across all ten completed model datasets are reported in Table 5. Task 2 exclusion was below 5% for every model. Task 1 exclusion was below 8% for nine of ten models; DeepSeek produced an elevated Task 1 exclusion rate of 21.7% (156 of 720 runs), discussed below.

Model	N	T1 inv.	T2 trunc.	USUALAPX%
Claude	720	0.0%	0.0%	91.4%
Gemma	720	7.4%	0.0%	93.6%
Qwen	720	0.0%	1.2%	84.2%
Llama	720	0.8%	4.9%	72.4%
GPT	720	1.1%	0.7%	89.0%
Mistral	720	0.1%	0.0%	51.1%
DeepSeek	720	**21.7%**	0.0%	84.7%
Qwen-7B	720	4.3%	0.1%	65.8%
Gemini	720	0.0%	0.0%	80.8%
DeepSeek-R1	720	0.0%	0.0%	87.5%

Table 5: Per-model exclusion rates and USUALAPX attribution rates on approximate utterances. **T1 inv.**: invalid Task 1 ratings (% of 720 runs); **T2 trunc.**: truncated Task 2 responses; **USUALAPX%**: rate of USUALAPX attribution on approx-form responses (Task 2 valid). Human reference for USUALAPX: 10.8%. Excluded reasoning models below the midrule.

DeepSeek's elevated Task 1 exclusion rate. The 21.7% Task 1 exclusion rate for DeepSeek is driven by JSON-format leakage from the Task 1 prompt template into the Task 2 generation slot: 81 of 720 Task 2 responses (11.2%) consist of JSON-formatted rating dictionaries rather than free-text motive attributions, and the same prompt-template confusion produces a higher rate of non-integer or out-of-range Task 1 ratings that fail the validity check. As a sensitivity analysis we recomputed Mag ρ for DeepSeek excluding the 81 JSON-leak Task 2 responses; the result is essentially unchanged from the primary analysis value (.878; deviation $< .01$), indicating that the leaked responses do not drive the headline Mag ρ value. The elevated Task 1 exclusion is therefore a methodological irregularity flagged for transparency rather than a substantive concern for the primary analysis; the remaining 564 valid Task 1 conversations provide adequate power for the per-cell point-biserial estimates.

Annotator. Free-text Task 2 responses were annotated using a multi-class LLM-based annotator (GPT-4o-mini) prompted with the 10-category Solt et al. taxonomy and category-specific decision rules. The annotator returns a binary vector indicating which categories apply to each response; multiple categories can be assigned per response, matching the human annotation procedure.

Validation. The annotator was validated on the full set of $N = 362$ Task 2 responses from `experiment1.csv`, for which manual gold-standard annotations are available. F1 was computed per category against the human labels. Table 6 reports F1 alongside the total positive cases assigned by humans (H+) and by the LLM annotator (L+).

Tier classification. Categories are grouped into four tiers based on F1 and inspection of error patterns: *High* (F1 $\geq$ .83, primary diagnostic categories used without further caveat); *Moderate* (F1 $\approx$ .76, used with minor caveats); *Caution* (F1 .66–.69, all three show $\sim$ 6–12pp false-positive bias on precise speakers; conclusions for these rely on directional rather than absolute rate comparisons); *Excluded* (USUALAPX, see below).

Exclusion of USUALAPX. Although USUALAPX reaches F1 = .78 on the human validation set, it is excluded from all LLM analyses due to a model-invariant annotation artefact identified in inspection.

Category	F1	H+	L+	Tier
NOTKNOW	.95	94	95	High
KNEWEXACT	.91	38	41	High
EASYSPEAKER	.88	9	7	High
FUSSYPERSON	.85	29	25	High
INFONOTAVAIL	.83	29	24	High
USUALAPX	.78	20	31	Excluded
INFOAVAIL	.76	17	20	Moderate
PURPOSE	.69	23	29	Caution
REALWORLD	.68	53	70	Caution
NEEDED	.66	40	51	Caution

Table 6: Annotator performance vs. human gold labels ($N = 362$). H+/L+ = total positive cases assigned by humans/LLM. Tier reflects how the category enters the LLM analyses.

		Task 1 SD		Motives/response	
Model	τ	**Approx**	**Precise**	**Approx**	**Precise**
Human	—	1.38	1.19	0.89	0.94
Claude	1.0	0.58	0.83	1.70	2.91
Gemma	1.8	0.90	0.93	1.14	2.69
Qwen	1.8	0.75	0.85	1.00	1.90
Llama	1.8	0.65	0.68	1.21	2.54
GPT	1.8	0.71	0.74	0.61	1.51
Mistral	1.8	0.85	0.68	1.10	1.86
DeepSeek	1.8	1.14	1.18	0.87	2.25
Qwen-7B	1.8	1.12	1.10	0.81	1.85
Gemini	1.8	0.82	0.71	1.89	3.07
DeepSeek-R1	1.8	1.11	0.93	1.66	2.75

Table 7: Within-condition Task 1 rating SD and mean motive attributions per Task 2 response, by utterance form. Claude was run at $\tau = 1.0$ (Anthropic API ceiling); all other models at $\tau = 1.8$. Excluded reasoning models below the second midrule.

The annotator triggers on generic descriptions of approximation as a linguistic convention (e.g., *"approximate numbers are common in casual speech"*) rather than attributions of habitual personal style to the specific speaker. All ten model datasets show USUALAPX attribution rates of 51–94% for approximate speakers (Table 5), compared to a human rate of 10.8%. The uniformity of this elevation across architecturally diverse models confirms an annotation artefact rather than a model behaviour. Since USUALAPX is rank 8/10 in human frequency and contributes no active cells to the impact structure (Appendix C), exclusion does not affect any substantive conclusion.

E Distributional Properties

Table 7 reports within-condition Task 1 rating SDs and mean motive attributions per Task 2 response for the eight chat-format models and the two format-excluded reasoning models, alongside the human reference values. Within-condition SD is computed as the mean across (form $\times$ context $\times$ attribute) cells of the within-cell SD across runs.

The retained chat-format models cluster between 0.58 and 1.18 on the within-condition SD scale, generally below the human reference (1.38 approx / 1.19 precise). Claude's notably lower SD reflects the API-imposed temperature ceiling of $\tau = 1.0$; the remaining seven chat-format models, run at $\tau = 1.8$, span 0.65–1.18. DeepSeek and Qwen-7B approach the human level. The two format-excluded reasoning models fall within the same range as the chat-format models on this metric.

The motives-per-response columns reveal a verbosity asymmetry: all ten models attribute substantially more motives per response than humans for precise speakers (1.51–3.07 vs. human 0.94), including some that approach human levels for approximate speakers (DeepSeek: 0.87; Qwen-7B: 0.81; both close to human 0.89), with GPT slightly below at 0.61. This form-dependent verbosity asymmetry motivates the normalisation procedure applied in §6 when comparing motive rate distributions across humans and models.

F Format-Based Exclusion: Diagnostics

This appendix documents the output-format diagnostics that motivated the exclusion of Gemini and DeepSeek-R1 from the primary cross-model analysis (§4.2). Mag ρ requires within-cell variation in attributed motives to estimate the per-cell point-biserial r_M reliably; models that produce nearly identical motive sets across all runs of a given (scenario $\times$ form $\times$ context) cell yield r_M values with little contrast, compromising the measurement. We identified the exclusion candidates from Task 2 output-format diagnostics rather than from Mag ρ itself, to avoid circularity.

Format-density metrics (exclusion criterion). Table 8 reports five output-format diagnostics computed on all non-truncated Task 2 responses from the ten completed model runs. Gemini and DeepSeek-R1 are extreme outliers across every formatting metric: they produce 3.83 and 5.52 bold-formatted text spans per response (vs. 0.57 in the most-formatted retained case, Qwen-7B; ≤ 0.13 for the other seven retained models; MAD-based modified $Z = +53.9$ and $+78.1$), generate 3.0 and 4.9 numbered-list items per response (vs. 0.29 for Qwen-7B; ≤ 0.13 elsewhere), and contain re-

Model	words	bold	nlist	phrase	top phrase
Claude	227	0.00	0.00	—	—
Gemma	67	0.00	0.00	—	—
Qwen	91	0.13	0.13	1.1%	'relevance'
Llama	114	0.00	0.00	—	—
GPT	93	0.02	0.02	0.3%	'simplicity'
Mistral	66	0.00	0.03	—	—
DeepSeek	114	0.08	0.06	0.1%	'info. retention'
Qwen-7B	77	0.57	0.29	1.4%	'accuracy'
Gemini	222	**3.83**	**3.01**	**5.0%**	'conv. norms'
DeepSeek-R1	224	**5.52**	**4.86**	**16.0%**	'efficiency'

Table 8: Task 2 output-format diagnostics across the ten completed model datasets. **words**: median word count per response; **bold**: mean count of `**bold**` text spans per response; **nlist**: mean count of numbered-list items per response; **phrase**: share of responses containing the model's most-recurrent bold-formatted header; **top phrase**: that phrase. Excluded models below the midrule. For entries with phrase share $\leq 1.5\%$ the identity of the top phrase is sensitive to tie-breaking among near-tied alternatives; the value reported is representative rather than uniquely determined.

curring signature template phrases that appear as bold-labelled headers in 5.0% and 16.0% of all responses respectively (vs. $\leq 1.7\%$ for all other models). The eight retained chat-format models produce predominantly free-prose Task 2 responses; even the most-formatted retained case (Qwen-7B at 0.57 bold spans and 0.29 numbered items per response) remains an order of magnitude below the excluded models on both metrics. The formatting gap between the excluded and retained models is more than an order of magnitude; no plausible cut-off in $[1.0, 3.0]$ bold spans per response affects the classification.

Within-cell motive uniformity. The formatting pattern correlates with high within-cell motive-attribution uniformity. Table 9 reports the mean pairwise Jaccard similarity between binary motive-attribution vectors within each (scenario $\times$ form $\times$ context) cell, averaged across cells, for the ten model datasets and the human data as a reference.

All models show higher within-cell motive uniformity than humans (0.34 approx / 0.23 precise), but the distribution within the LLM set is informative: Gemini's precise-form Jaccard of 0.75 is the highest in the sample, meaning that any two of its responses to the same cell share three-quarters of their attributed motives on average. Claude and DeepSeek-R1 reach a comparable level (0.76 and 0.61 approx; 0.73 precise each), with Claude doing so without bullet-point formatting. This suggests that high motive uniformity can arise either from canned structured templates (Gemini, DeepSeek-R1) or from consistent prose-level reasoning (Claude). The exclusion criterion is the formatting regime itself rather than within-cell uniformity per se.

Source	Jaccard (approx)	Jaccard (precise)
Human	0.34	0.23
Claude	0.76	0.73
Gemma	0.65	0.65
Qwen	0.48	0.49
Llama	0.58	0.65
GPT	0.60	0.52
Mistral	0.47	0.39
DeepSeek	0.50	0.45
Qwen-7B	0.40	0.34
Gemini	0.65	**0.75**
DeepSeek-R1	0.61	0.73

Table 9: Mean pairwise Jaccard similarity between binary motive-attribution vectors within (scenario $\times$ form $\times$ context) cells, averaged across cells, separately for approximate and precise utterance forms. Human reference computed on the Solt et al. (2025) Experiment 1 dataset. Excluded models below the second midrule.

Sample responses. Both excluded models produce highly structured Task 2 responses with bold-labelled headers and numbered sub-items, in contrast to the free-prose responses produced by all eight retained chat-format models. DeepSeek-R1's template typically opens with 4–6 numbered points (e.g. *"1. **Cognitive efficiency**: … "*, *"2. **Contextual appropriateness**: … "*), with the header *'efficiency'* appearing in 115 of 720 responses (16.0%). Gemini produces analogous structured argumentation with *'conversational norms'* as its most common recurring header. Full example responses are available in the public dataset release.

G Excluded Reasoning Models: Descriptive Analyses

This appendix reports descriptive analyses of the two reasoning models excluded from the primary cross-model analysis on output-format grounds (Gemini and DeepSeek-R1; see Appendix F for the exclusion diagnostics). We retain both models in three forms: as a ten-model robustness check on H1 and H2, as descriptive case studies of dissociations between motive–rating mapping and calibration, and as input to the future-work discussion in §6.3.

Ten-model robustness check. Both pre-specified hypotheses survive the inclusion of the two excluded models. **H1 (pooled)**: across all ten completed models, the pooled Spearman correlation between $B_M(a, v)$ and observed Task 1 mean ratings is $\rho = +.709$, $p < .001$ ($N = 110$ attribute $\times$ form cells); the per-model rank correlations for the two excluded models are $\rho = +.782$, $p = .005$ for Gemini and $\rho = +.891$, $p < .001$ for DeepSeek-R1, both significant and stronger than the 8-model median. **H2 (cross-model)**: with all ten models the Spearman correlation between Mag ρ and CDS is $\rho = -.64$, $p = .048$, $N = 10$, with bootstrap 95% CI $[-.95, -.09]$ and 98.3% of 10,000 resamples negative. The relationship is therefore significant in both the primary 8-model and inclusive 10-model analyses, though the magnitude is weaker in the inclusive analysis because the two excluded models deviate from the 8-model trend in opposite directions (see below).

Descriptive metrics. Table 10 reports the primary metrics for the two excluded models alongside their position relative to the 8-model OLS trend predicting CDS from Mag ρ. Gemini sits far above the trend line (residual $+0.72$): its CDS is more than double the value its Mag ρ would predict, reflecting systematic magnitude inflation across all benchmark effects (Table 11; mean ESR $= 2.35$, range 1.75–3.33). DeepSeek-R1 sits below the trend line (residual -0.23): its CDS is roughly 40% lower than its Mag ρ would predict, reflecting near-perfect calibration on three of five competence-related main effects (*competent*, *well-prepared*, *knowledgeable*; ESR $\in [0.93, 0.98]$) combined with extreme *pedantic* inflation (ESR $= 3.02$) and uniformly attenuated form $\times$ context interactions (ESR $\in [0.64, 1.18]$).

Dissociation in DeepSeek-R1. The combination of low CDS (0.372, the lowest in the full ten-model sample) with mid-range Mag ρ (0.659, ranked 6th of 10) is the empirical dissociation referenced in §6.3: DeepSeek-R1 achieves output-level calibration without the corresponding process-level coherence captured by Mag ρ. The per-effect breakdown (Table 11) localises the source: three main effects in the competence cluster reach near-perfect calibration (ESR $\in [0.93, 0.98]$), driving the low aggregate CDS, while the remaining seven effects deviate substantially in both directions. This pattern is consistent with templated reasoning that produces correct main-effect magnitudes on the most frequently discussed attributes (the competence cluster) without preserving the human motive–rating geometry that Mag ρ measures; the eight retained chat-format models do not show this dissociation (§5). We discuss the broader interpretation in §6.3.

Model	CDS	MFD	Mag ρ	H1 ρ
Gemini	1.352	.191	.599	$+.782^{**}$
DeepSeek-R1	.372	.200	.659	$+.891^{***}$
8-model range	.39–.72	.08–.16	.19–.88	+.46–.82

Table 10: Primary metrics for the two format-excluded models, with the range over the eight retained chat-format models for reference. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Effect	Δ_H	Gemini	DeepSeek-R1
Main effects of form			
competent	+0.88	2.31	0.98
well-prepared	+1.27	1.95	0.93
knowledgeable	+0.97	1.91	0.96
helpful	+0.90	1.82	0.75
pedantic	+0.88	2.62	**3.02**
Form × context interactions			
competent	+1.14	2.31	0.64
well-prepared	+1.07	2.55	0.64
knowledgeable	+0.63	3.33	0.74
helpful	+1.00	2.97	1.17
likeable	+0.86	1.75	1.18
mean ESR		2.35	1.10

Table 11: Per-effect ESR ($|\Delta_M|/|\Delta_H|$) breakdown for the two excluded models. Δ_H is the human effect magnitude; ESR $= 1$ indicates perfect calibration. Gemini inflates every effect; DeepSeek-R1 hits near-perfect calibration on three of five competence-cluster main effects but inflates *pedantic* by $\sim$3$\times$ and attenuates most form $\times$ context interactions.

H Supplementary Results

This appendix reports four supplementary analyses referenced in §5: H2 robustness diagnostics, form-exclusivity verification, the motive balance

deviation analysis supplementing H1, and the full motive attribution rates underlying the H2 and null discussions.

Robustness diagnostics for H2. The primary H2 correlation (Spearman $\rho = -.810$, $p = .015$, $N = 8$) is robust to single-model deletion. Table 12 reports the eight leave-one-out subsets; the magnitude varies across $\rho \in [-.71, -.89]$ and strict significance ($p < .05$) is retained in five of eight subsets, with the three marginal cases (drop Claude, drop Gemma, drop DeepSeek) reaching $p = .052$, $p = .052$, and $p = .071$ respectively. Bootstrap resampling of the 8-model set (10,000 iterations, models resampled with replacement) gives a 95% CI of $[-1.00, -.27]$ on the Spearman correlation, with 99.2% of resamples returning a negative ρ. A one-sided permutation test (10,000 shuffles of the CDS labels) gives $p = .011$; the two-sided permutation test gives $p = .021$.

Model dropped	Spearman ρ	p
Claude	$-.750$	.052
Gemma	$-.750$	.052
Qwen	$-.786$	.036
Llama	$\mathbf{-.893}$	**.007**
GPT	$-.857$	.014
Mistral	$-.857$	.014
DeepSeek	$-.714$	.071
Qwen-7B	$-.786$	.036

Table 12: Leave-one-out Spearman correlations for the H2 cross-model analysis on the 8-model primary set ($N = 7$ per LOO subset). The result is robust in direction and in significance across all eight subsets; Llama-removed yields the strongest correlation.

Form-exclusivity verification. As a methodological precondition for H2, the eight chat-format models correctly respect form-exclusivity: across all 5,731 non-truncated Task 2 responses in the primary set, precise-only motives (KNEWEXACT, FUSSYPERSON, INFOAVAIL) receive zero attribution on approximate-form responses, and approximate-only motives (NOTKNOW, INFONOTAVAIL, EASYSPEAKER) receive zero attribution on precise-form responses (per-model violation rate: 0.00% in both directions). The motive-mediated framework therefore operates over a structurally well-formed motive space.

Motive balance deviation (supplement to H1). Table 13 reports Spearman correlations between motive balance *deviation* ($\Delta B = B_M - B_H$, where B_H is the human motive balance) and Task 1 rating deviation (ΔTask 1 = LLM mean − human mean), per model and pooled across the eight chat-format models. The pooled correlation is small and not significant ($\rho = +.121$, $p = .26$, $N = 88$); per-model correlations vary widely in both magnitude and sign and none reach individual significance. This indicates that the motive balance correctly locates the *position* of ratings (the H1 result reported in §5.1) but does not predict the *direction* of calibration error in the chat-format sample. The within-model motive–rating mapping is therefore a good diagnostic of rating position but not of calibration-error direction; this dissociation parallels the H2 vs. null contrast at the per-cell level.

Model	$\rho(\Delta B, \Delta\textbf{Task1})$	p
Claude	$+.409$	.211
Gemma	$-.418$	.201
Qwen	$+.209$	.537
Llama	$+.100$	.770
GPT	$+.355$	.285
Mistral	$+.082$	.811
DeepSeek	$+.036$	.916
Qwen-7B	$-.009$	.979
Pooled	$+.121$	.262

Table 13: Spearman ρ between motive balance deviation ΔB and Task 1 rating deviation ΔTask 1, per model and pooled (pooled $N = 88$ across the 8 chat-format models).

Full motive attribution rates (supplement to §5.2). Table 14 reports raw Task 2 motive attribution rates for the six most diagnostic categories across the eight chat-format models and the human benchmark, separately by utterance form. Three patterns support the discussion in §5.2. First, all eight chat-format models correctly assign zero rates to logically impossible categories (NOTKNOW for precise speakers, KNEWEXACT for approximate speakers), confirming form-exclusivity. Second, all eight chat-format models substantially over-attribute REALWORLD for precise speakers (human: .23; models: .46–.67), a model-independent finding reframed as partly a verbosity artefact in §6. Third, NOTKNOW attribution for approximate speakers varies widely across models (.16 for GPT to .91 for Claude, vs. human .51); this variation does not correspond to calibration quality, illustrating the null result.

Form	Motive	Human	Claude	Gemma	Qwen	Llama	GPT	Mistral	DeepSeek	Qwen-7B
Approx	NOTKNOW	.508	.914	.364	.285	.629	.163	.483	.228	.294
	INFONOTAVAIL	.157	.244	.142	.144	.086	.211	.175	.189	.122
	REALWORLD	.065	.267	.128	.251	.252	.138	.175	.217	.119
	NEEDED	.076	.228	.414	.107	.199	.011	.061	.142	.028
	KNEWEXACT	.000	.000	.000	.000	.000	.000	.000	.000	.000
	FUSSYPERSON	.005	.000	.000	.000	.000	.000	.000	.000	.000
Precise	REALWORLD	.232	.608	.653	.639	.667	.666	.564	.656	.460
	KNEWEXACT	.215	.606	.258	.342	.425	.256	.397	.394	.281
	FUSSYPERSON	.158	.525	.381	.286	.411	.103	.167	.169	.409
	NEEDED	.147	.431	.508	.342	.514	.281	.281	.522	.234
	NOTKNOW	.000	.000	.000	.000	.000	.000	.000	.000	.000
	INFONOTAVAIL	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table 14: Task 2 motive attribution rates by form across the eight chat-format models and the human benchmark. Form-exclusivity constraints (zero rate for impossible categories) are respected by all models. REALWORLD inflation for precise speakers is universal across models. NOTKNOW attribution for approximate speakers varies widely without corresponding to calibration quality (see §5.2).

I Task 3 Results

Task 3 (multiple-choice motive selection) was collected for GPT and Llama only, and is reported here for completeness. Both models show near-ceiling selection rates (90–100%) for most options, eliminating meaningful form discrimination. The critical human null, EASIERUNDERSTAND selected equally for precise and approximate speakers (31.1% vs. 30.8%, $p = .957$), is violated by both models (GPT: 23.3% vs. 96.7%; Llama: 35.8% vs. 85.0%). The ceiling effect reflects over-willingness to endorse reasons rather than selective attribution; total selections per response (~ 5) are comparable to humans but distributed indiscriminately across options.