What Makes a Good Query? Measuring the Impact of Human-Confusing Linguistic Features on LLM Performance

William Watson* **Nicole Cho*** **Sumitra Ganesh** **Manuela Veloso**
J.P. Morgan AI Research
nicole.cho@jpmorgan.com

Abstract

Large Language Model (LLM) hallucinations are usually treated as defects of the model or its decoding strategy. Drawing on classical linguistics, we argue that a query's form can also shape a listener's (and model's) response. We operationalize this insight by constructing a 17-dimension query feature vector covering clause complexity, lexical rarity, and anaphora, negation, answerability, and intention grounding, all known to affect human comprehension. Using **369,837** real-world queries, we ask: *Are there certain types of queries that make hallucination more likely?* A large-scale analysis reveals a consistent "risk landscape": certain features such as deep clause nesting and underspecification align with higher hallucination propensity. In contrast, clear intention grounding and answerability align with lower hallucination rates. Others, including domain specificity, show mixed, dataset- and model-dependent effects. Thus, these findings establish an empirically observable query-feature representation correlated with hallucination risk, paving the way for guided query rewriting and future intervention studies.

1 Introduction

Large Language Models (LLMs) have transformed natural language processing, yet their propensity to hallucinate, producing plausible but factually incorrect outputs, remains a critical challenge, especially in high-stakes domains such as finance and law (Huang et al., 2025; Dahl et al., 2024; Naveed et al., 2024). The societal, financial, and legal costs of hallucinations are already evident, with multiple lawsuits emerging in response to LLM-generated errors (Milmo, 2023), underscoring the impracticality of relying on users to detect such failures. While most prior work emphasizes *reactive, post-generation* mitigations (e.g., self-verification, logit-based detection) (Lewis et al., 2021; Madaan et al., 2023) and *proactive, pre-generation* strategies (e.g., RAG) (Lewis et al., 2021; Watson et al., 2025), comparatively fewer studies take a *proactive, input-side* view beyond ambiguity detection (Zhang et al., 2024; Kuhn et al., 2023).

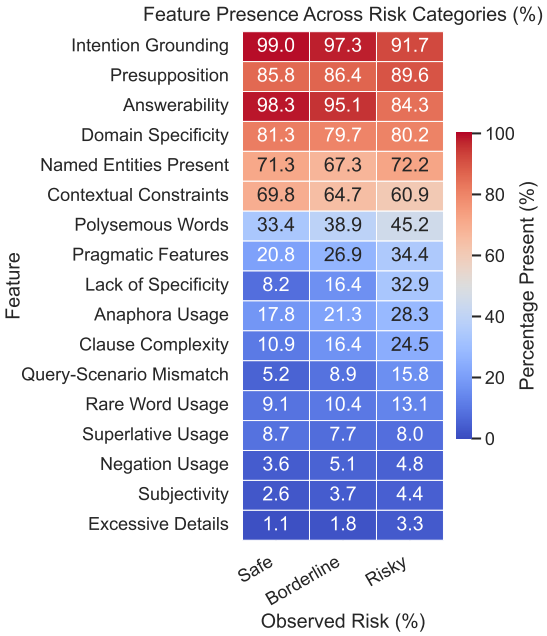

Figure 1: **Prevalence of binary linguistic features across hallucination risk categories (*Safe*, *Borderline*, *Risky*).** Warmer colors indicate higher frequency. *Lack of specificity*, *clause complexity*, and *polysemous words* show a pronounced rise from *Safe* to *Risky*.

Drawing on classical linguistics, we define a 17-dimensional query feature vector capturing structural, lexical, stylistic, and semantic aspects known to shape human comprehension and obfuscate understanding. While a few of these features have been studied for their influence on general LLM performance (Truong et al., 2023; Cho and Watson, 2025), to our knowledge there is no large-scale empirical mapping from such features to *hallucination* behavior. Following Blevins et al. (2023), we leverage an LLM to extract these features from **369,837 real-world queries** spanning **13 QA datasets** (3

*Equal Contribution

Findings of the Association for Computational Linguistics: EACL 2026, pages 4794–4827
March 24-29, 2026

scenarios, 16 configurations).

Using an semantics-preserving paraphrase neighborhood with an offline Monte Carlo correctness proxy, we provide empirical evidence of strong correlations between specific linguistic markers and observed hallucination rates, yielding a consistent "risk landscape". Features that *destabilize* interpretation (underspecification, deep clause nesting) align with higher hallucination propensity, whereas features that *tighten* semantics (clear intention grounding, answerability) align with lower risk rates. Others (e.g., domain specificity) show mixed, dataset- and model-dependent effects. Surprisingly, several linguistic features traditionally known to confuse human readers (e.g., word rarity, superlatives, complex negation) show minimal association with hallucination in LLMs, suggesting that human and model failure modes need not coincide. Our contributions are threefold:

- **Feature taxonomy and extraction:** A linguistically grounded, 17-feature representation of queries known to impact language understanding for humans. We bring this perspective to LLMs and understand whether or not these features are associated with hallucinatory behavior.
- **Risk landscape at scale:** An empirical, distributional map derived from ordinal modeling with dataset/scenario fixed effects and ECDF separations, linking query features to hallucination propensity over 369,837 queries.
- **Proactive guidance:** We highlight practical, feature-aware triage and low-effort rewrites that can complement reactive defenses.

Therefore, we advocate a practical approach to *proactively* mitigate hallucinations *before* generation by optimizing queries—rather than relying on post hoc human inspection, which is often impractical.

2 Related Work

Hallucinations in LLMs: Prior work addresses hallucinations both *proactively* and *reactively* (Ji et al., 2023a; Varshney et al., 2023; Li et al., 2024). Proactive, input-time methods (e.g., Retrieval-Augmented Generation, external tool use) enrich the context *before* decoding (Lewis et al., 2021; Schick et al., 2023; Qin et al., 2023). Reactive, post-generation methods (e.g., self-consistency, logit-based detectors) evaluate or re-rank outputs after the model decodes a response (Wang et al., 2023b; Manakul et al., 2023).

Pre-Generation Query Evaluation: More recent research has begun to address hallucination proactively by examining the input query itself (Ji et al., 2023b; Karpukhin et al., 2020). Studies have shown that query structure and semantic properties, such as polysemy, contextual nuance, and specificity, play a crucial role in shaping LLM outputs (Brown et al., 2020; Jiao et al., 2023). For example, ReLA (Zhang et al., 2021) demonstrates that sparse attention can improve both interpretability and performance without additional overhead. HalluciBot (Watson et al., 2025) further illustrated that perturbing queries can effectively estimate hallucination likelihood. In contrast, our work systematically extracts 22 linguistic features from queries and empirically analyzes their correlation with hallucination risk. This *proactive* approach lays the foundation for query pre-filtering techniques aimed at enhancing the reliability of LLM outputs.

3 Methodology

Problem setup. We study how the linguistic form of a user query modulates large language model reliability. Each query i receives an ordinal triage label $y_i \in \{0, 1, 2\}$ corresponding to $\textsc{Safe} < \textsc{Borderline} < \textsc{Risky}$. Let $x_i \in \{0, 1\}^p$ be a binary feature vector capturing human-confusing linguistic phenomena (§B), and c_i the observed covariates (dataset $d(i)$, scenario $s(i)$). We model $\Pr(y_i \mid x_i, c_i)$ to quantify (i) marginal effects of features, (ii) distributional shifts in predicted risk, and (iii) robustness under dataset shifts–*without* rewriting queries.

3.1 Linguistic features

We operationalize p=17 query-level features spanning ambiguity (*Lack of Specificity*, *Polysemous Words*, *Pragmatic Features*), referential structure (*Anaphora*), complexity (*Clause Complexity*), polarity (*Negation*), grounding (*Answerability*, *Intention Grounding*, *Contextual Constraints*), and others (§B). Detectors return structured outputs (label+rationale) via typed prompts; positive/negative 5-shot examples appear in App. G. Detector noise is treated as classical measurement error and expected to attenuate magnitudes rather than flip signs (Blevins et al., 2023).

3.2 Observed risk via semantics-preserving perturbations

Benchmark items can be memorized, biasing raw hallucination rates (Carlini et al., 2021; Nasr et al.,

	Ordinal (β-only)					Correlation			
Feature	Coef	SE	z-value	p-value	OR	ρ	τ	Adj. p	$p < 0.05$
Lack of Specificity	0.868	0.010	85.898	$<10^{-5}$	2.382	0.271	0.256	$<10^{-5}$	✓
Clause Complexity	0.568	0.010	57.363	$<10^{-5}$	1.764	0.155	0.147	$<10^{-5}$	✓
Negation Usage	0.311	0.016	19.499	$<10^{-5}$	1.364	0.028	0.026	$<10^{-5}$	✓
Excessive Details	0.247	0.026	9.668	$<10^{-5}$	1.281	0.066	0.063	$<10^{-5}$	✓
Anaphora Usage	0.214	0.009	23.827	$<10^{-5}$	1.238	0.107	0.101	$<10^{-5}$	✓
Polysemous Words	0.096	0.007	13.840	$<10^{-5}$	1.101	0.104	0.098	$<10^{-5}$	✓
Rare Word Usage	0.095	0.011	8.997	$<10^{-5}$	1.100	0.055	0.052	$<10^{-5}$	✓
Pragmatic Features	0.072	0.008	8.496	$<10^{-5}$	1.074	0.132	0.125	$<10^{-5}$	✓
Presupposition	0.056	0.010	5.565	$<10^{-5}$	1.058	0.046	0.044	$<10^{-5}$	✓
Contextual Constraints	0.044	0.007	5.812	$<10^{-5}$	1.045	-0.081	-0.077	$<10^{-5}$	✓
Parse Tree Height	0.011	0.005	2.312	0.021	1.011	-0.149	-0.121	$<10^{-5}$	✓
Named Entities Present	0.009	0.007	1.269	0.205	1.009	0.003	0.002	0.115	✗
Domain Specificity	0.003	0.009	0.396	0.692	1.003	-0.013	-0.012	$<10^{-5}$	✓
Query-Scenario Mismatch	-0.064	0.014	-4.734	$<10^{-5}$	0.938	0.153	0.145	$<10^{-5}$	✓
Superlative Usage	-0.103	0.012	-8.674	$<10^{-5}$	0.902	-0.012	-0.011	$<10^{-5}$	✓
Dependency Depth	-0.128	0.005	-24.353	$<10^{-5}$	0.879	-0.203	-0.159	$<10^{-5}$	✓
Intention Grounding	-0.168	0.023	-7.272	$<10^{-5}$	0.846	-0.159	-0.151	$<10^{-5}$	✓
Subjectivity	-0.168	0.019	-8.885	$<10^{-5}$	0.846	0.044	0.041	$<10^{-5}$	✓
Query Token Length	-0.212	0.010	-20.973	$<10^{-5}$	0.809	-0.274	-0.214	$<10^{-5}$	✓
Number of Clauses	-0.262	0.009	-28.652	$<10^{-5}$	0.769	-0.272	-0.228	$<10^{-5}$	✓
Answerability	-1.106	0.017	-63.425	$<10^{-5}$	0.331	-0.228	-0.216	$<10^{-5}$	✓

Table 1: **Results for Observed Risk analyses. Left:** Ordinal logistic regression estimates (using both binary and scaled numeric predictors. **Right:** Spearman's ρ and Kendall's τ correlation coefficients between each feature and Observed Risk, with adjusted p-values and a significance indicator. Features in *italics* (e.g., *Lack of Specificity*, *Clause Complexity*, *Query Token Length*, *Number of Clauses*, and *Answerability*) highlight particularly intriguing effects. All adjusted p-values were below 10^{-5} except for "Named Entities Present" ($p = 0.115$, not significant).

2023; Aerni et al., 2024; Watson et al., 2025). For each original query q_{orig}, we generate a local semantic equivalence class $\mathcal{N}(q_{\text{orig}})=\{q_1, \ldots, q_m\}$ by sampling paraphrases at $T=1.0$ with the instruction "PRODUCE A SEMANTICALLY INDIFFERENT BUT LEXICALLY PERTURBED VERSION OF THE QUERY." We retain the first six paraphrases whose hybrid similarity meets $s(q_{\text{orig}}, q_i) \geq 0.85$,

$$s(q_{\text{orig}}, q_i) = \lambda_{\text{bi}} \cdot \cos\big(\mathbf{e}_{\text{bi}}(q_{\text{orig}}), \mathbf{e}_{\text{bi}}(q_i)\big) + \lambda_{\text{cross}} \cdot \frac{1}{2}\left[\Pr_{\text{cross}}(q_{\text{orig}}, q_i) + \Pr_{\text{cross}}(q_i, q_{\text{orig}})\right]$$

with $(\lambda_{\text{bi}}, \lambda_{\text{cross}}) = (0.6, 0.4)$, $\mathbf{e}_{\text{bi}}$ from TEXT-EMBEDDING-3-LARGE (3,072-d), and $\Pr_{\text{cross}}$ from MS-MARCO-MINILM-L6-V2 (Reimers and Gurevych, 2019).

Empirical hallucination estimation. For each $q_i \in \mathcal{N}(q_{\text{orig}})$ we compute a convex proxy $\hat{h}(q_i) = w_0\, s_{\text{llm}} + w_1\, s_{\text{fuzz}} + w_2\, s_{\text{bleu}}$, combining a binary LLM-judge decision $s_{\text{llm}} \in 0, 1$ (semantic; Wang et al., 2023a; Liu et al., 2023b; Adlakha et al., 2024), fuzzy string similarity $s_{\text{fuzz}} \in [0, 1]$ (surface; Bachmann, 2024), and BLEU-1 $s_{\text{bleu}} \in [0, 1]$ (lexical; Papineni et al., 2002; Lin and Och, 2004; Callison-Burch et al., 2006). We use $(w_0, w_1, w_2) = (0.6, 0.3, 0.1)$, selected on a small human-labeled set by sweeping the (w_0', w_1', w_2') simplex; the ROC–AUC surface is flat for $w_0 \pm 0.2$, drops quickly with larger w_1, and is worst for BLEU-only, placing our mix on a Pareto plateau (App. C, Fig. 8). A perturbation counts as *hallucinated* if $\hat{h}(q_i) > 0.5$. Aggregating across the six paraphrases yields query-level categories: **Safe** (0/6), **Borderline** (1–3/6), **Risky** (4–6/6).

3.3 Ordinal risk model

We fit a proportional-odds (cumulative logit) model

$$\log \frac{\Pr(Y_i \leq k \mid x_i, c_i)}{\Pr(Y_i > k \mid x_i, c_i)} = \tau_k - \eta_i \qquad (1)$$

with $k \in 0, 1$, linear predictor $\eta_i = x_i^\top \beta + \alpha_{d(i)} + \gamma_{s(i)}$ and ordered cutpoints $\tau_0 < \tau_1$. Class probabilities are:

$$p_0 = \sigma(\tau_0 - \eta_i)$$
$$p_1 = \sigma(\tau_1 - \eta_i) - \sigma(\tau_0 - \eta_i)$$
$$p_2 = 1 - \sigma(\tau_1 - \eta_i)$$

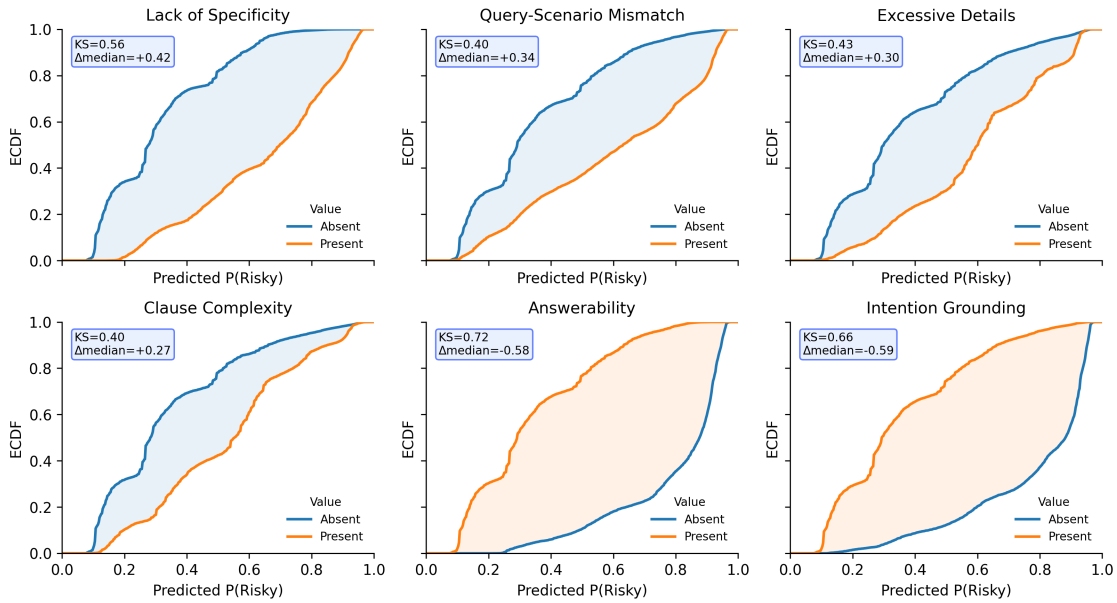

Figure 2: **ECDFs of predicted $P(\text{Risky})$ for Present vs. Absent (top six by KS).** Shaded regions indicate dominance; inset shows KS and Δmedian. Lack of Specificity, Excessive Details, Clause Complexity, and Query–Scenario Mismatch shift mass toward higher risk; Answerability and Intention Grounding shift mass lower.

optimized by NLL with ℓ_2 penalty $\lambda_{\text{reg}}\|\beta\|_2^2$ and no explicit intercept. We report:

- **Specification** S_β (feature-only): β with linguistic features only;
- **Specification** $S_{\beta,\gamma,\alpha}$ (full): β with both scenario γ and dataset α fixed effects.

Figure 4 visualizes *feature* coefficients (β; left) and *dataset–scenario* effects (α, γ; right). (We use β for features throughout, reserving α and γ for dataset/scenario.)

3.4 Metrics and diagnostics

We summarize effects at three levels:

- **Coefficients** (β) from (1) under S_β and $S_{\beta,\gamma,\alpha}$ (Fig. 4; Table 1).
- **Distributional separations**: ECDFs of predicted $P(\text{Risky})$ for *Present* vs. *Absent* groups; we report KS distance and Δmedian (Fig. 2).
- **Calibration**: reliability curves and ECE within feature strata (App. Fig. 12).

We additionally examine **length–feature interactions** by quantile-binning query length and plotting the empirical rate of a RISKY label for *Present* vs. *Absent*, by scenario (Fig. 3, App. Fig. 10). To contextualize correlational claims, we plot **propensity overlap** (Present/Absent densities; standardized mean differences) to document where comparisons are well-posed (App. Fig. 13).

Propensity modeling. For each binary linguistic feature f (treatment $T_f \in \{0, 1\}$), the *propensity score* is the probability that a query exhibits f given its other covariates. Let Z_f stack the remaining feature indicators x_{-f} together with scenario/dataset indicators (fixed effects γ, α). We fit a separate logistic model per feature, $\pi_f(z) = \Pr(T_f{=}1 \mid Z_f{=}z) = \sigma(\phi_{0f} + z^\top \phi_f)$, yielding per-item scores $\hat{\pi}_f = \pi_f(Z_f)$ used for overlap diagnostics.

3.5 Robustness

We perform Leave-One-Dataset-Out (LODO) refits of Eq. (1) and summarize the mean ± stddev of β across holds (Fig. 5). Signs and relative magnitudes remain stable, indicating that the observed "risk landscape" is not driven by any single dataset.

4 Experimental Setup

Model under test. All generations use `gpt-4o-2024-08-06` with a single prompting recipe held fixed across datasets; temperature $\tau = 1.0$ for both answering and paraphrase sampling. Detector and audit prompts (structured outputs, 5-shot positives/negatives ICL) and sampling settings are provided in (App. G).

Datasets and scenarios. We evaluate 13 QA datasets spanning three scenarios (16 total configurations; Table 4):

- **Extractive:** SQuADv2
- **Multiple Choice:** TruthfulQA, SciQ, MMLU, PIQA, BoolQ, OpenBookQA, MathQA, ARC-Easy, ARC-Challenge

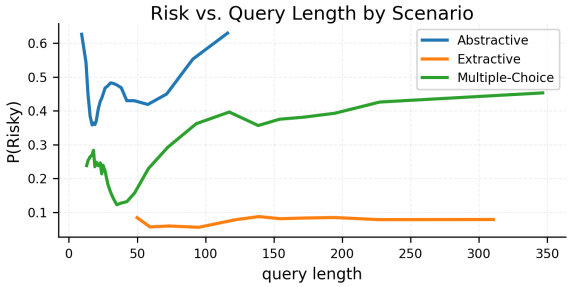

Figure 3: **Risk vs. query length by scenario.** Each curve shows the *empirical* probability of a risky output (fraction of "Risky" labels) after quantile-binning query length within a scenario ($\geq$ 50 examples per bin). Risk rises with length for **Abstractive**, remains low/flat for **Extractive**, and is intermediate for **Multiple-Choice**. *Takeaway:* longer, open-ended queries are more hallucination-prone, while extractive settings remain robust across lengths.

- **Abstractive:** SQuADv2, TruthfulQA, SciQ, WikiQA, HotpotQA, TriviaQA

In total, we analyze $N = 369{,}837$ query–response pairs. Scenario (γ) and dataset (α) enter the ordinal model as fixed effects (Figure 4).

Feature extraction. For each query we run structured detectors, producing (label $\in \{0, 1\}$, rationale) per feature. Each detector's rubric is calibrated on a 100 sample held-out set to reduce systematic bias (App. G).

Outcome construction. The triage label (SAFE/BORDERLINE/RISKY) is derived from the paraphrase set using the convex hallucination proxy $\hat{h}(\cdot) > 0.5$ threshold. We confirm that ordinal coefficients align with ECDF separations of predicted $P(\text{Risky})$ (Fig. 2). Ordinal KDE distributions per class are reported in Fig. 14.

Training details. We implement the ordinal model in PyTorch (1×NVIDIA T4), optimize NLL with Adam optimizer and ℓ_2 regularization, and use early stopping on a validation split (Kingma and Ba, 2017). We fit a pooled model once and then run LODO refits (one dataset held out at a time).

5 Results: A Query-Feature Risk Landscape for Hallucination

Overview and hypotheses. We evaluate how human-confusing linguistic phenomena relate to LLM hallucination risk across datasets and task formats. Guided by the features in §B and the ordinal model in §3.3, we test:

- **H1 (Ambiguity/complexity → higher risk):** underspecification, anaphora, negation, and clause-level complexity increase risk.
- **H2 (Grounding → lower risk):** explicit intention and answerability reduce risk.
- **H3 (Domain effects):** domain-specificity has mixed association, moderated by model familiarity with the domain.

5.1 Feature and Dataset Effects

Figure 4 summarizes proportional-odds estimates for two specifications: $\mathbf{S}_{\beta}$ (feature-only) and $\mathbf{S}_{\beta,\gamma,\alpha}$ (scenario/dataset-adjusted). On features, *Answerability* shows the largest protective effect (negative β), while *Lack of Specificity*, *Negation*, and *Anaphora* are positively associated with risk, consistent with **H1 & H2**. Structure-related indicators (*Clause Complexity*, *Polysemous Words*, *Pragmatic Features*) also increase risk but with smaller magnitudes. On contexts, fixed effects mirror scenario difficulty: abstractive configurations are riskier on average (SQUAD (ABSTR.), HOTPOTQA), multiple-choice safer (SCIQ, ARC-EASY), and extractive in between. The signs and relative magnitudes of feature coefficients are stable with leave-one-dataset-out fits, indicating they are not artifacts of a single dataset or scenario mix.

5.2 Distributional Effects

To move beyond point estimates, we compare ECDFs of predicted $P(\text{Risky})$ for *Present* vs. *Absent* items per feature (Figure 2). The top separations by KS confirm the ordinal results: *Lack of Specificity*, *Excessive Details*, *Clause Complexity*, and *Query–Scenario Mismatch* shift mass toward higher risk (positive Δmedian), while *Answerability* and *Intention Grounding* shift mass lower (negative Δmedian).

5.3 Task Format Moderates Absolute Risk But Not Direction

Baseline differences by dataset/scenario (Figure 4, Figure 3) are substantial:risk rises sharply with length for **Abstractive**, remains low/flat for **Extractive**, and is intermediate for **Multiple-Choice**. Nevertheless, the *direction* of feature effects are stable across bins. Largest gaps appear for shorter, open-ended prompts, where ambiguity features notably raise empirical RISKY rates and grounding features reduce them. Risk–length profiles (Figure 10) further clarify that open-ended, longer prompts in **Abstractive** settings amplify risk,

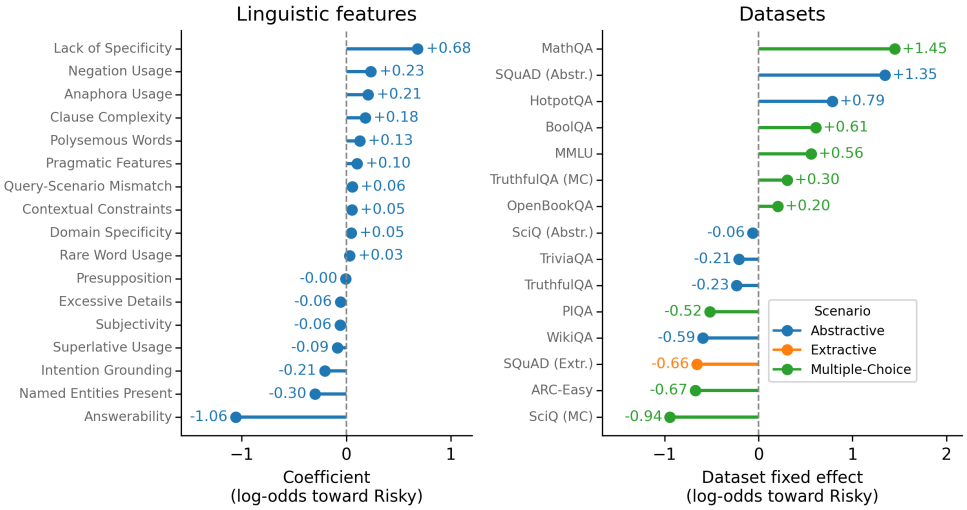

Figure 4: **Feature coefficients β (left) and dataset/scenario fixed effects α, γ (right) from the ordinal logit model. Positive values increase log-odds of *Risky*.** *Answerability* is strongly protective; *Lack of Specificity*, *Negation*, and *Anaphora* increase risk.

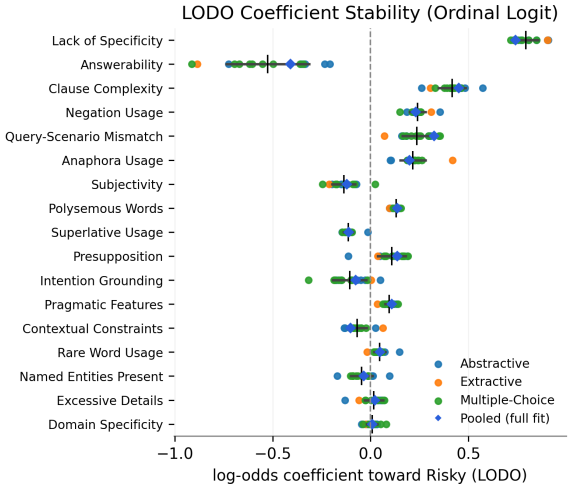

Figure 5: **LODO coefficient stability (ordinal logit).** Each point is a feature coefficient estimated when one dataset is held out (color = held-out dataset's scenario), with short horizontal bars showing the mean and ±1 s.d. across LODO runs. The blue diamond is the pooled (full-fit) coefficient. Signs and magnitudes are stable: *Lack of Specificity*, *Clause Complexity*, *Query–Scenario Mismatch* remain risk-increasing, while *Answerability* remains strongly protective.

whereas **Extractive** settings remain comparatively flat across context lengths.

5.4 Propensity overlap & uplifts

We estimate per-feature propensities $\hat{\pi}_f = \Pr(T_f{=}1 \mid Z_f)$ and plot Present/Absent KDEs to assess common support (App. Fig. 13). Where overlap is adequate, we compute uplifts in $\Pr(\textsc{Risky})$ via IPW and stratified matching (App. Table 6).

- **Well-supported (uplifts reported):** *Lack of Specificity*, *Clause Complexity* (top-two).
- **Degenerate overlap (associational only):** *Answerability*, *Intention Grounding* (top-two).

When queries are otherwise comparable, tightening specificity and simplifying clause structure offers the clearest, overlap-supported path to reduce $\Pr(\textsc{Risky})$; strongly protective signals like *Answerability* and *Intention Grounding* remain robust correlates but cannot be treated as causal toggles due to limited overlap.

5.5 Robustness Across Datasets

Leave-one-dataset-out refits (Figure 5) preserve the signs and relative ranks of the dominant features: *Answerability* remains protective; *Lack of Specificity*, *Clause Complexity*, and *Query–Scenario Mismatch* remain risk-increasing—indicating conclusions are not driven by any single dataset. Calibration within Present/Absent strata (App.Fig.12) is near-diagonal with small ECE, supporting the use of probability shifts as meaningful rather than artifacts of miscalibration. Correlation structure among features (Figure 6) clusters ambiguity markers together and grounding markers together, aligning with the observed risk directions.

5.6 Linguistic Trends (Figure 1)

Higher-Risk Queries Are Marked by Ambiguity and Complexity. Features such as *lack of specificity*, *anaphora usage*, *polysemy*, *pragmatic features*, and *clause complexity* show increased prevalence when moving from *Safe* to *Risky* queries, suggesting that higher ambiguity & syntactic depth

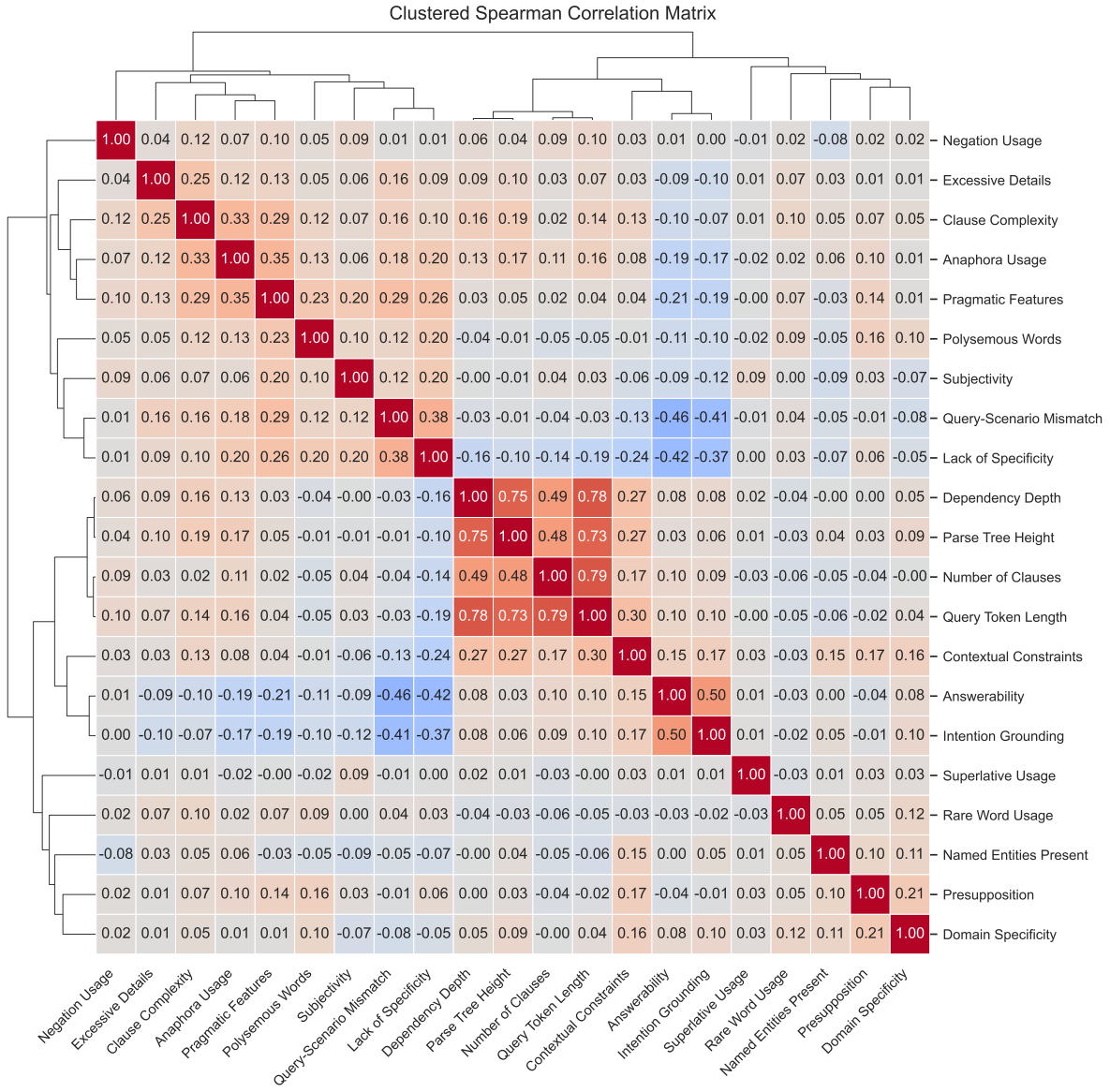

Figure 6: **Clustered Spearman correlation matrix using complete linkage & correlation distance.** The color scale ranges from red ($\rho = 1$, strong positive correlation) to blue ($\rho = -1$, strong negative correlation). Dendrograms group features with similar correlation patterns and share similar linguistic functions.

are more frequently associated with hallucination.

Presupposition Is Common Across Categories. Interestingly, *presupposition* occurs frequently even in *Safe* queries, suggesting that its presence alone does not imply elevated risk. However, its co-occurrence with other risk-associated features, such as structural complexity or misaligned context, may contribute to increased hallucination rates.

Core Anchors of "Safe" Queries: Intention Grounding and Answerability. Queries that explicitly convey user intent (*intention grounding*) and are demonstrably answerable from available context (*answerability*) are highly concentrated in the *Safe* category (over 90%). This pattern aligns with the hypothesis that semantic clarity and contextual grounding are predictive of lower hallucination propensity.

5.7 Correlation Clusters (Figure 6)

Syntactic Complexity: *Query Token Length*, *Dependency Depth*, *Parse Tree Height*, and *Number of Clauses* cluster tightly (with correlations up to $\rho = 0.79$). Notably, these features exhibit significant **inverse associations** with hallucination, i.e., richer contextual cues coincide with lower risk.

Semantic Grounding: *Intention Grounding* and *Answerability* correlate strongly ($\rho = 0.60$) and are moderately associated with *Contextual Constraints*. This cluster is linked to lower hallucination, consis-

tent with the hypothesis that semantically grounded queries tend to yield more accurate responses.

Ambiguity: *Lack of Specificity*, *Query-Scenario Mismatch*, *Polysemous Words*, and *Pragmatic Features* show moderate intercorrelations ($\rho = 0.38$ between *Lack of Specificity* and *Query-Scenario Mismatch*). This group appears frequently in queries with higher hallucination propensity, indicating shared ambiguity-related characteristics.

Lexical and Stylistic Features: Attributes such as *Negation Usage*, *Excessive Details*, *Subjectivity*, and *Superlative Usage* exhibit weak correlations overall. However, these features may interact with others to influence model behavior, though their individual contributions appear limited.

Domain-Oriented Group: *Domain Specificity*, *Named Entities Present*, and *Presupposition* form a loose cluster ($\rho = 0.21$ for *Named Entities Present* and *Domain Specificity*). This suggest that domain-driven queries may entail presuppositional assumptions, which could correlate with hallucination risk when the model lacks sufficient domain familiarity.

5.8 Regression-Based Associations with Risk

Table 1 integrates our ordinal logistic regression estimates with nonparametric correlation metrics with respect to the observed hallucination rates. A **positive** coefficient indicates a feature that is positively associated with hallucination propensity, whereas a **negative** coefficient signifies an inverse association.

High-Impact, Risk-Increasing:

- *Lack of Specificity* presents the highest positive coefficient (0.868) and an odds ratio (OR) of 2.382, suggesting that queries which omit concrete details or precise aims are more likely to be associated with higher-risk outputs.
- *Clause Complexity* (0.568, OR=1.764) is also strongly associated with hallucination, consistent with the observation that syntactically intricate prompts co-occur with elevated error rates, consistent with its ECDF right-shifts.

Protective Features:

- *Answerability* exhibits the largest negative coefficient (-1.106, OR = 0.331), suggesting that queries with clear, retrievable answers tend to have lower hallucination scores.
- *Intention Grounding* (-0.168) is also negatively associated, indicating that queries with explicit intent are less likely to exhibit hallucination. Both align with strong left-shifts in $P(\text{Risky})$ ECDFs.
- Syntactic features (*Query Token Length* (-0.212), *Dependency Depth* (-0.128), and *Number of Clauses* (-0.262)) are inversely correlated with risk, potentially reflecting that greater syntactic structure can provide helpful context.

Moderately Associated Features:

- *Negation Usage* (0.311) and *Anaphora Usage* (0.214) are positively associated with hallucination risk, possibly due to the interpretive ambiguity they introduce. However, they are both weaker than ambiguity/structure features.
- *Polysemous Words* (0.096) broaden interpretive pathways, causing LLMs to fill gaps with hallucinated details and erroneous responses.

Mixed or Context-Moderated:

- *Named Entities Present* is not statistically significant ($p = 0.205$), no clear association between entity presence and hallucination propensity.
- *Domain Specificity* has a near-zero coefficient (0.003, ρ = -0.013), suggesting highly variable associations, possibly dependent on the model's familiarity with the domain in question.

5.9 Findings with respect to hypotheses

- **H1:** *Lack of Specificity*, *Clause Complexity*, *Negation*, and *Anaphora* show **strong positive associations with** hallucination risk and upward ECDF shifts.
- **H2:** *Answerability*, *Intention Grounding* exhibit substantial negative coefficients and downward ECDF shifts. This suggests that well-defined queries provide a **protective effect** against hallucinations.
- **H3:** *Domain Specificity* has **mixed, variable associations**; effects appear moderated by dataset or model familiarity rather than uniformly positive or negative.

5.10 Practical Applications: Risk Triage and Low-effort Rewrites

Triage: At inference time, systems can (i) detect features, (ii) compute predicted $P(\text{Risky})$ under $S_{\beta,\gamma,\alpha}$, and (iii) route high-risk queries to either a clarifying step or a retrieval/tool-grounded path.

Low-effort rewrites: Our results yield three low-effort rules, directly tied to the highest-leverage features, generalize across tasks:

(1) add disambiguating *constraints* (time, place, entity) to raise specificity;

(2) always state *intent* explicitly (e.g., "summarize / compare / extract / verify");
(3) always resolve *polysemy* up front (e.g., Java the language vs. the island).

Our length-conditioned profiles indicate these edits are especially important for short, open-ended prompts, where risk gaps between Present/Absent are largest. These steps are potentially automatable and align with the strongest negative coefficients and ECDF separations.

6 Conclusion

Taken together, our results suggest that a substantial portion of "hallucination risk" is attributable to *how much the query commits the model to a determinate reading*. Queries that declare intent and make answerability explicit constrain the hypothesis space the model must explore; underspecified or structurally intricate queries expand that space and invite speculative completion. The correlation clusters are consistent with this view: grounding features co-cluster and associate with lower risk; ambiguity markers co-cluster and associate with higher risk; syntactic complexity interacts with these axes, sometimes compounding ambiguity (nested clauses, unresolved anaphora), sometimes adding helpful scaffolding when paired with explicit constraints. These findings highlight the potential for automated query filtering and rewriting strategies to enhance model reliability by flagging risk-associated linguistic markers *directly*.

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JPMorgan") and is not a product of the Research Department of JPMorgan. JPMorgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

Limitations

Our findings should be interpreted with the following limitations in mind. First, the study is primarily observational. Although we use overlap diagnostics (propensity/positivity) and ablations to qualify comparisons, these provide, at best, quasi-causal evidence. Several features (e.g., *Answerability*) are inherently semantic and not cleanly manipulable without changing meaning, so we treat their coefficients as empirical associations corroborated by multiple diagnostics rather than as causal effects. Furthermore, our experiments are limited to English-language queries and one class of LLMs. We do not account for multimodal inputs or evolving model behavior across versions. Additionally, feature extraction relies on existing NLP toolkits and LLM predictions, which may introduce parsing errors in noisy queries. Additionally, we treat the linguistic features as independent variables and do not model higher-order interactions. Future work could explore whether specific feature combinations jointly contribute to increased hallucination risk. Importantly, the feature correlations should not be interpreted as evidence of causality. Due to the opacity of neural representations and the challenge of tracing internal mechanisms, we frame our findings as empirical associations rather than causal claims. Finally, while our reward formulation is rigorously tuned using Pareto-optimal ROC–AUC analysis, it relies partially on an LLM-based judge, which may itself introduce systematic biases.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Preprint*, arXiv:2307.16877.

Michael Aerni, Javier Rando, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramèr. 2024. Measuring non-adversarial reproduction of training data in large language models. *Preprint*, arXiv:2411.10242.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024. Why does the effective context length of llms fall short? *Preprint*, arXiv:2410.18745.

Max Bachmann. 2024. rapidfuzz/rapidfuzz: Release 3.8.1.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. *Preprint*, arXiv:2211.07830.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Nicole Cho and William Watson. 2025. Multiq&a: An analysis in measuring robustness via automated crowdsourcing of question perturbations and answers. *Preprint*, arXiv:2502.03711.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 web track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, volume 500-278 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Scott A. Crossley and Stephen Skalicky. 2017. Making sense of polysemy relations in first and second language speakers of english. *International Journal of Bilingualism*, 23(2):400–416. (Original work published 2019).

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Preprint*, arXiv:2401.01301.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. *Preprint*, arXiv:2302.12246.

Janosch Haber and Massimo Poesio. 2024. Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

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 (ICLR)*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. `https://spacy.io/`. Accessed 2025-10-05.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, arXiv:1705.03551.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Muhammad Ali Khalidi. 2023. Domain specificity. In *Cognitive Ontology: Taxonomic Practices in the Mind-Brain Sciences*, pages 100–122. Cambridge University Press.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sanggoo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007, Miami, Florida, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Clam: Selective clarification for ambiguous questions with generative language models. *Preprint*, arXiv:2212.07769.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Factuality enhanced language models for open-ended text generation. *Preprint*, arXiv:2206.04624.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

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

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023a. We‘re afraid language models aren‘t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Brian MacWhinney, Elizabeth Bates, and Reinhold Kliegl. 1984. Cue validity and sentence interpretation in english, german, and italian. *Journal of Verbal Learning and Verbal Behavior*, 23(2):127–150.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

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

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Klara Marton, Richard G. Schwartz, Lajos Farkas, and Victoria Katsnelson. 2006. Effect of sentence length and complexity on working memory performance in hungarian children with specific language impairment (sli): A cross-linguistic comparison. *International Journal of Language & Communication Disorders*, 41(6):653–673.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Dan Milmo. 2023. Two US Lawyers Fined for Submitting Fake Court Citations from ChatGPT. The Guardian. Accessed: 2025-03-21.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *Preprint*, arXiv:2311.17035.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *The 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, Saarbrucken, Germany.

Yasuhiro Ozuru, Stephen Briner, Christopher A. Kurby, and Danielle S. McNamara. 2013. Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, 67(3):215–227.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Friedemann Pulvermüller and Yury Shtyrov. 2006. Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology*, 79(1):49–71.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *Preprint*, arXiv:2307.16789.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jerrold M. Sadock and Arnold M. Zwicky. 1985. Speech acts distinctions in syntax. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, pages 155–196. Cambridge University Press, Cambridge.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Silke Scheible. 2008. Annotating superlatives. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Preprint*, arXiv:1904.06707.

Ethel Schuster. 1988. Anaphoric reference to events and actions: A representation and its advantages. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

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, Bangkok, Thailand. Association for Computational Linguistics.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and coreference resolution: A review. *Preprint*, arXiv:1805.11824.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *Preprint*, arXiv:2306.08189.

Rob A. Van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *Preprint*, arXiv:2307.03987.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

William Watson, Nicole Cho, and Nishan Srishankar. 2025. Is there no such thing as a bad question? h4r: Hallucibot for ratiocination, rewriting, ranking, and routing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25470–25478.

William Watson, Nicole Cho, Nishan Srishankar, Zhen Zeng, Lucas Cecchi, Daniel Scott, Suchetha Siddagangappa, Rachneet Kaur, Tucker Balch, and Manuela Veloso. 2024. Law: Legal agentic workflows for custody and fund services contracts. *Preprint*, arXiv:2412.11063.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *Preprint*, arXiv:1707.06209.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhen Zeng, William Watson, Nicole Cho, Saba Rahimi, Shayleen Reynolds, Tucker Balch, and Manuela Veloso. 2024. Flowmind: Automatic workflow generation with llms. *Preprint*, arXiv:2404.13050.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2021. Sparse attention with linear units. *Preprint*, arXiv:2104.07012.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *Preprint*, arXiv:2405.12063.

A Distribution of Hallucination Across Query Type

Hallucination distributions vary across query scenarios:

- **Extractive:** Hallucinations are infrequent, likely due to the presence of explicit supporting context; most queries are classified as *Safe*.
- **Multiple Choice:** The presence of distractor options corresponds with a higher proportion of *Borderline* cases.
- **Abstractive:** Lacking external context, abstractive queries are most frequently associated with hallucinations, with a large share labeled *Risky*.

B Linguistic Features

B.1 Structural Features

Query and Context Length: Sentence length has long been studied as a core factor affecting working memory performance in children (Marton et al., 2006). An interesting finding posits that syntactic complexity has a far more negative impact on a child's comprehension than sentence length (Marton et al., 2006). We bring this perspective to our study and strive to understand whether LLMs are impacted by query length. Prior studies have dived into whether LLMs can actually comprehend windows that reach their nominal capacity (An et al., 2024). In contrast, instead of stress-testing the LLM by reaching the model's context window, we aim to measure the correlation between the total length of the *query* and *context* and hallucination propensity.

Anaphoric References: Anaphora refers to words (e.g., *he, she, it, this, that, these, those*) referencing previously mentioned entities, states, or actions (Schuster, 1988). For instance, in *"I like ice-cream. Do you think* ***it*** *is my favorite dessert?"*, "it" is an anaphor pointing back to "ice-cream." Anaphoric references and their effective representations for human understanding have long confounded linguists (Mann and Thompson, 1988). Traditional NLP research has focused on coreference resolution, linking pronominal or nominal mentions to antecedents (Sukthanker et al., 2018), rather than NER. We investigate whether the presence of anaphora itself is associated with LLM errors.

Clause Complexity: Syntactic complexity is known to hinder understanding (MacWhinney et al., 1984; Marton et al., 2006). We define clause complexity as the presence of multiple subordinate clauses, which introduce syntactic dependencies.

Query Type	Train	Val	Test	Total
Extractive	80,049	5,843	–	85,892
Multiple Choice	45,997	14,127	21,573	81,697
Abstractive	176,446	24,521	1,281	202,248
Overall	302,492	44,491	22,854	369,837

Table 2: Number of queries across *Extractive*, *Multiple Choice*, and *Abstractive* categories, split by train, validation (Val), and test sets. Note that we make no distinction between these splits in our analysis.

	Safe		Borderline		Risky	
Query Type	**Count**	**%**	**Count**	**%**	**Count**	**%**
Extractive	58,834	69.0	19,618	23.0	6,773	8.0
Multiple Choice	38,869	47.0	24,711	29.9	19,064	23.1
Abstractive	67,078	33.4	44,244	22.0	89,429	44.5

Table 3: Observed Risk counts and row-normalized percentages across query types. Each risk group shows the count and percentage of predictions labeled as *Safe*, *Borderline*, or *Risky*.

We study whether clause complexity is a feature that induces hallucinatory behavior. We count subordinate clauses using spaCy's dependency parser (Honnibal et al., 2020) and LLM-based predictions.

Dependency Tree Depth: This metric measures how many layers of syntactic dependencies a query contains. Deeper dependency trees often involve more complex resolution chains and long-term memory, as studied in cognitive science (Lewis et al., 2006). We compute dependency depth and **Parse Tree Height**, through spaCy's dependency parser (Honnibal et al., 2020) to understand whether it influences misunderstanding by LLMs.

B.2 Scenario-Based Features

Query Type: Cognitive science has long studied the different abilities required to answer open-ended (abstractive questions) compared to multiple-choice questions. Ozuru et al. (2013) studied how the efficacy in responding to open-ended questions was associated with the caliber of self-explanatory elaborations, whereas the accuracy in answering multiple-choice questions was linked to the extent of pre-existing knowledge pertinent to the text. These outcomes imply that open-ended and multiple-choice question formats assess distinct dimensions of comprehension mechanisms. While Watson et al. (2025) has studied the effects of different scenarios on LLMs, we delve deeper into whether it is associated with hallucinatory outputs.

Mismatch: Beyond query type, we also assess

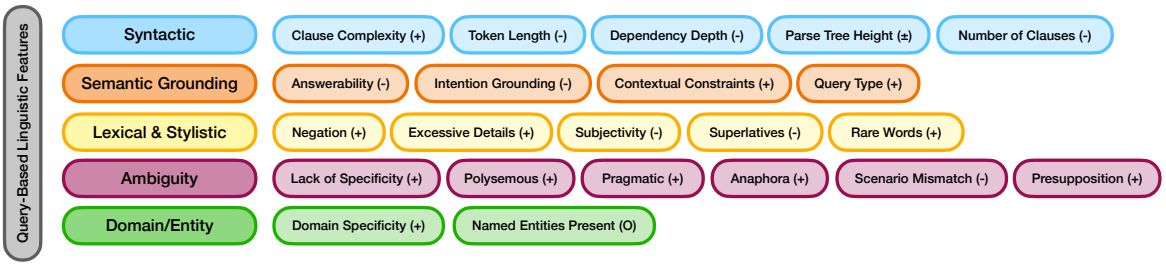

Figure 7: Illustration of our query-based linguistic features, grouped by correlation categories. The (+) or (-) signs indicate whether each feature is positively or negatively associated with hallucination risk, while "±" and "O" denote mixed or non-significant effects, respectively. Note: several **syntactic** features (e.g., token length, number of clauses) show **negative associations**, indicating richer structure can co-occur with *lower* risk.

whether a query is aligned with the scenario in which it is posed. For instance, prompts implying an *extractive* setup (*"refer to recent news"*) paired with an *abstractive* scenario that lacks such information. This feature is inspired by Mismatch Negativity (MMN) in cognitive science, where unexpected stimuli elicit neural responses (Pulvermüller and Shtyrov, 2006). Similarly, we examine if LLMs are impacted by contextual mismatches.

Presupposition: These are implicit assumptions that a query treats as true. For example, *"Who is the musician that developed neural networks?"* presupposes that such a musician exists (Levinson, 1983). We follow Van der Sandt (1992) in identifying presuppositional triggers such as interrogative words (*"Who is the football player with white hair?"*), possessive forms (*"Shelley likes her dog."*), and counterfactuals (*"I would be happy if I had money."*). Presupposition is known to hinder linguistic clarity; we study its correlation with hallucination in LLMs.

Pragmatics: Addresses context and discourse driven meanings that are not strictly encoded lexically or syntactically (Levinson, 1983; Sadock and Zwicky, 1985). For instance, *"Can you pass me the salt?"* is less about physical ability and more about willingness. Sravanthi et al. (2024) has released a benchmarks on tasks that involve understanding pragmatics - we extend this research in-depth to understand whether pragmatics impacts downstream LLM behavior.

B.3 Lexical Features

Word Rarity: Prior work from 2019 indicates that LLMs often struggle with rare vocabulary (Schick and Schütze, 2019); as LLMs have advanced rapidly in the past few years, our motivation is to understand whether word rarity is still a risk factor for LLM understanding.

Negation Usage: Mis-primed queries including *not*, *never*, and *no* have been shown to confuse LLMs more than humans (Kassner and Schütze, 2020; Truong et al., 2023). We include negation to understand its correlation for the latest LLMs.

Superlatives: Following Scheible (2008), superlative expressions (*biggest*, *fastest*, *best*) indicate comparisons within a set of options that may be ambiguous or not always apparent. The interpretation of superlative adjectives has long been a study in linguistics - we therefore select it as one of our features in this study.

Polysemy: Polysemy, or lexical ambiguity, is where words have multiple related meanings (Haber and Poesio, 2024). For instance, the word *"mouth"* can refer either to a bodily feature or the mouth of a river. Polysemy presents a significant challenge for English as a Second Language (ESL) learners, as it necessitates advanced cognitive processing to discern context-dependent semantic nuances and apply appropriate interpretations within varied linguistic frameworks (Crossley and Skalicky, 2017).

B.4 Stylistic Complexity

Answerability: Sarcastic or rhetorical questions pose a greater challenge for comprehension compared to straightforward, answerable queries, as they require the interpreter to discern underlying intent, contextual cues, and implicit meanings that deviate from literal interpretations, often necessitating a nuanced understanding of social and linguistic subtleties (Oraby et al., 2017). For example, the query *"Based on recent news, are investors expressing concern for Stock A?"* is composed with greater clarity than *"So, do you think Stock A is going to plummet?"*. Operationally, we prompt an LLM to mark a query as *answerable* if the query (i) has a single or small set of verifiable answers within the provided context/dataset, (ii) is not rhetorical/sarcastic, (iii) does not require external, time-varying

facts unless supplied.

Excessive Details: We examine whether queries overloaded with details influence hallucination probability. While chain-of-thought prompting (Sahoo et al., 2024; Diao et al., 2024) leverages detailed reasoning, it remains uncertain whether excessive details may instead overwhelm the model and trigger hallucinations.

Subjectivity: Traditional linguistics have studied the different formulations of fact-based and subjective opinions. Therefore, we strive to understand, whether a subjective opinion formulation for an LLM engenders more hallucinatory behavior.

Lack of Specificity: Queries that are broadly phrased and lack concrete details are inherently ambiguous and open to multiple interpretations (Brown et al., 2020; Kim et al., 2024; Liu et al., 2023a). Operationally, we prompt an LLM to mark *present* if ≥ 1 of: (i) missing disambiguating constraints (time/place/entity), (ii) multiple plausible interpretations without tie-breakers, (iii) underspecified task (e.g., *"tell me about X"* without scope). Not specific queries can include *"Tell me about Tesla."*, where multiple interpretations are valid (company, car, tech, stock). A specific query contains contextual clues to identify the scope and entity discussed: *"Summarize 2024 Q4 Tesla earnings call highlights in* ≤ 5 *bullets."*

B.5 Semantic Grounding

Intention Grounding: A query is well-grounded in intention if its purpose is immediately clear without requiring additional context (Clarke et al., 2009). For example, *"What are the tax implications of investing in municipal bonds in the U.S.?"* in contrast to *"What happens if I invest?"*.

Contextual Constraints: Precise constraints such as specific timeframes, locations, or conditions can guide language comprehension through optimized memory storage (Marton et al., 2006). We evaluate if contextually constrained queries are less prone to hallucination.

Named Entity Presence: People, organizations, and places (verifiable entities) may ground LLMs in external factual information (Lee et al., 2023). We take this study deeper and understand whether the presence of named entities has any measurable impact, if any, on downstream hallucination.

Domain Specificity: Domain specificity, a concept utilized across various research programs in cognitive science, refers to cognitive abilities that are constrained in specific manners. Certain cognitive abilities are confined to a particular domain, while others extend beyond it. The difficulty lies in defining the boundaries of a domain for a given capacity, particularly because knowledge areas are not inherently segmented into distinct compartments (Khalidi, 2023). Therefore, we measure whether domain-specific terminology can influence hallucination risk, depending on the model's expertise. Studies in finance (Zeng et al., 2024) and law (Watson et al., 2024) show that LLMs perform better with domain-specific tools.

C Reward-Weight Simplex Analysis:

We swept (w'_0, w'_1, w'_2) over a triangular grid ($w'_0 + w'_1 + w'_2 = 1$) and computed ROC–AUC on the 100 item human-labeled validation set. AUC degrades when relying on BLEU alone, and increases when dominated by the judge; the $0.6/0.3/0.1$ convex mix is on the Pareto plateau. The Pareto frontier (Appendix 8) for $\hat{h}(q_i)$ reveals the following:

- **LLM-Judge Robustness (w_0):** The ROC–AUC surface is nearly invariant when w_0 varies by ± 0.2: AUC shifts by <0.5%, indicating our formulation tolerates large w_0 weight swings.
- **Fuzzy-Match Sensitivity (w_1):** Small increases in w_1 rapidly exit the Pareto region, showing that the fuzzy-match term must be tuned carefully to avoid degrading overall accuracy.
- **BLEU-Only Pitfall (w_2):** As w_2 increases, AUC steadily declines, bottoming out at $w_2 = 1$, where the metric overemphasizes surface overlap at the expense of semantic correctness.
- **Pareto-Optimal Region:** We select (0.6, 0.3, 0.1) as our final weights, which lie deep in the high-AUC plateau, confirming it is a Pareto-optimal trade-off among semantic, fuzzy, and lexical signals.

D Feature Calculation Methodology

We computed our linguistic features using a combination of techniques:

- **LLM Structured Output Model:** For binary features, we employed an LLM structured output model (`gpt-4o-2024-08-06`) that leverages a Pydantic schema. This schema includes, for every feature dimension, a chain-of-thought slack variable, enabling the model to consider all relevant variables before predicting the final boolean values for each feature. Our in-context examples and definitions are itemized in Table 7.
- **spaCy Parsers:** Syntactic features such as the

Dataset	Scenario	Domain	License	Count	Citation
SQuADv2	E, A	Wikipedia	CC BY-SA 4.0	86K	Rajpurkar et al. (2016, 2018)
TruthfulQA	M, A	General Knowledge	Apache-2.0	807	Lin et al. (2022)
SciQ	M, A	Science	CC BY-NC 3.0	13K	Welbl et al. (2017)
MMLU	M	Various	MIT	15K	Hendrycks et al. (2021)
PIQA	M	Physical Commonsense	AFL-3.0	17K	Bisk et al. (2020)
BoolQ	M	Yes/No Questions	CC BY-SA 3.0	13K	Clark et al. (2019); Wang et al. (2019)
OpenBookQA	M	Science Reasoning	Apache-2.0	6K	Mihaylov et al. (2018)
MathQA	M	Mathematics	Apache-2.0	8K	Amini et al. (2019)
ARC-Easy	M	Science	CC BY-SA 4.0	5K	Clark et al. (2018)
ARC-Challenge	M	Science	CC BY-SA 4.0	2.6K	Clark et al. (2018)
WikiQA	A	Wikipedia QA	Other	1.5K	Yang et al. (2015)
HotpotQA	A	Multi-hop Reasoning	CC BY-SA 4.0	72K	Yang et al. (2018)
TriviaQA	A	Trivia	Apache-2.0	88K	Joshi et al. (2017)

Table 4: Overview of datasets used in our study, including domain, license, number of examples, and associated scenario types. These datasets span a diverse range of question types, knowledge areas, and reasoning skills, supporting robust evaluation across domains. Scenario types tested: E = Extractive, M = Multiple Choice, A = Abstractive.

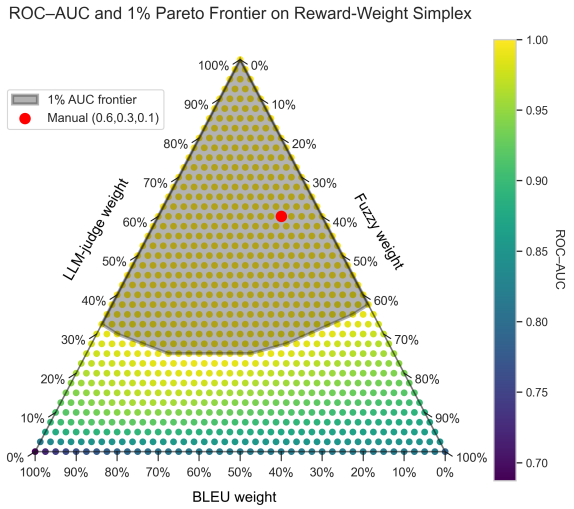

Figure 8: ROC–AUC Landscape over Reward-Weight Simplex. Each point represents a convex combination of weights (α, β, γ) over the LLM-judge, Fuzzy, and BLEU metrics. Color indicates ROC–AUC measured on a held-out validation set; the shaded region denotes the top 1% frontier. Our selected weights $(0.6, 0.3, 0.1)$ are marked in red.

number of clauses, dependency depth, and parse tree height were computed using spaCy's parsers, which provided robust dependency and constituency parsing capabilities (Honnibal et al., 2020).

- **OpenAI's tiktoken Library:** Token lengths were determined using OpenAI's `tiktoken` library,[1] with encoding `o200k_base`, ensuring consistency with the tokenization process used during simulation. Note that the observed risk is derived from responses generated with `gpt-4o-2024-08-06`.

[1] https://github.com/openai/tiktoken

E Ordinal Logistic Regression Details

We use an ordinal logistic regression model (`OrderedModel` from `statsmodels` (Seabold and Perktold, 2010)) to estimate the effect of linguistic features on hallucination risk. The response variable is ordinal with three levels: *Safe*, *Borderline*, and *Risky*. Predictor variables include the binary linguistic features described in Section 3. Trained using the BFGS optimization method. Table 1 reports the estimated coefficients, where positive values indicate a higher likelihood of hallucination risk.

Analysis of Results. For each binary feature f, we plot ECDFs of model-predicted $P(\text{Risky})$ for $f{=}0$ vs $f{=}1$, report the Kolmogorov–Smirnov distance (KS) and $\Delta\text{median} = \text{median}(P(\text{Risky}) \mid f{=}1) - \text{median}(P(\text{Risky}) \mid f{=}0)$, and shade the region of dominance. For length analyses we partition queries into $Q{=}30$ equal-mass bins by token length. Within each bin we compute the empirical rate of the *Risky* label for items with a feature *Present* vs *Absent*. We plot the bin means against the bin centers, with binomial 95% confidence bands. For ECDFs we compare the distributions of model-predicted $P(\text{Risky})$ under Present vs Absent and report the Kolmogorov–Smirnov distance and Δmedian. The distributions exhibit the expected ordering (e.g., *Risky* items have higher $P(\text{Risky})$ mass).

Calibration. We bucket predicted $P(\text{Risky})$ into 10 equal-mass bins and plot observed frequency vs mean predicted probability, separately for Present/Absent per feature. Expected Calibration Error (ECE) is the weighted average of absolute devia-

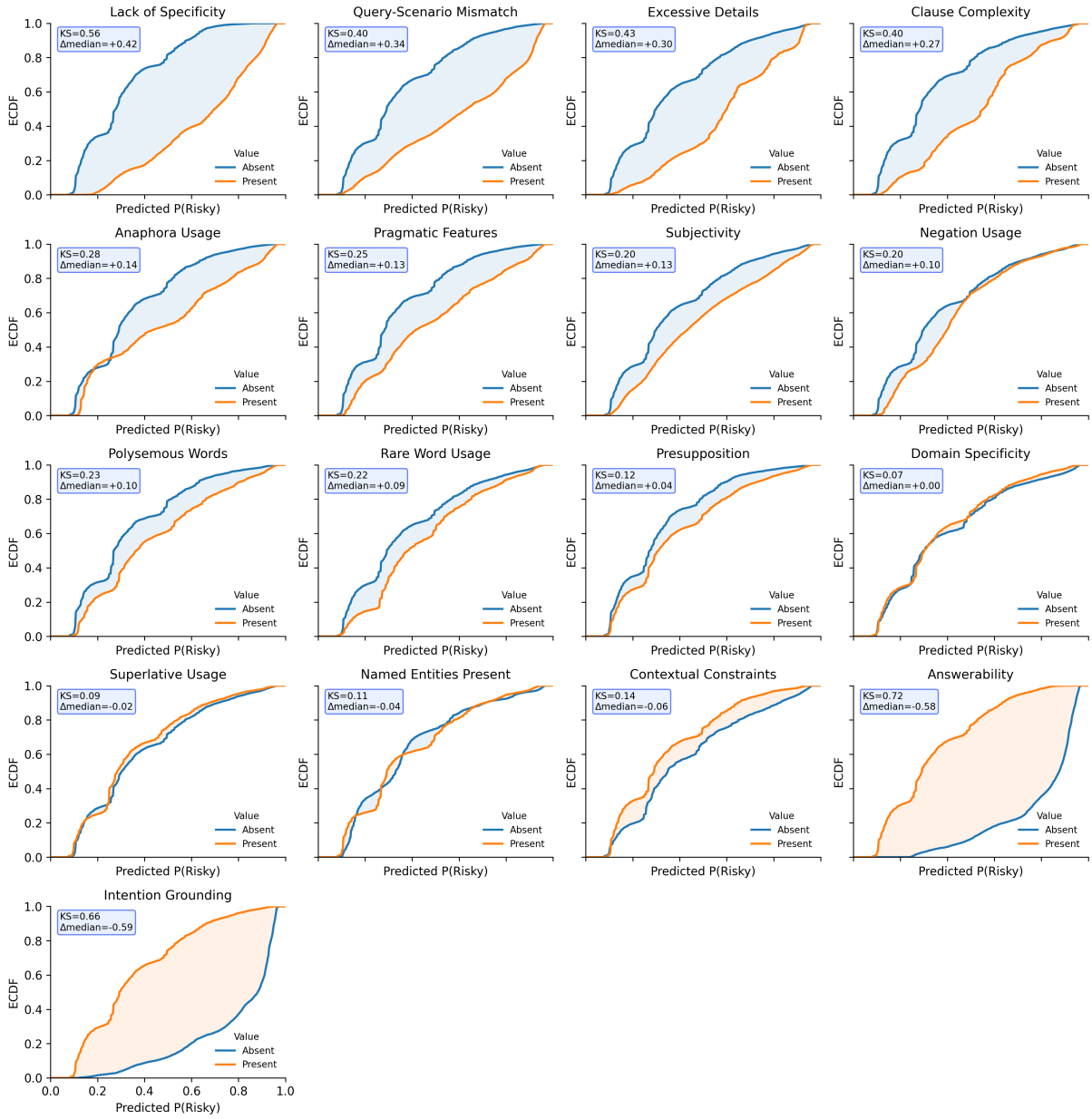

Figure 9: **All features: ECDFs of predicted $P(\textbf{Risky})$ by feature presence.** Same rendering as Figure 2.

tions between observed and predicted bin rates.

F Propensity and IPW Uplift Computation

Setup. For each binary linguistic feature f (treatment $T_{fi} \in \{0,1\}$ for item i), let Z_{fi} stack all *other* feature indicators $x_{-f,i}$ together with scenario and dataset indicators (γ, α). The outcome used for uplift is a scalar $O_i \in [0,1]$ (either the observed risky label $\mathbf{1}\{y_i{=}\textsc{Risky}\}$ or the model-implied $P_i(\textsc{Risky})$).

Propensity model. We estimate the feature-specific propensity

$$\pi_f(z) \;=\; \Pr(T_f{=}1 \mid Z_f{=}z)$$

with a separate logistic regression per f:

$$\hat{\pi}_{fi} \;=\; \sigma\big(\phi_{0f} + Z_{fi}^{\top}\phi_f\big),$$

holding the global covariate set fixed but excluding f to avoid leakage. To prevent extreme weights, we bound $\hat{\pi}_{fi} \in [10^{-3}, 1{-}10^{-3}]$.

Overlap (positivity) diagnostic. We quantify common support via the *overlap share*

$$\text{overlap}_f \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big\{\hat{\pi}_{fi} \in [\alpha, 1{-}\alpha]\big\}$$

where $\alpha{=}0.05$, and visualize $\hat{\pi}_{fi}$ for *Present* vs. *Absent* with KDEs (Fig. 13). We report matched/In-

Feature	KS	Δmedian	n_{abs}	n_{pres}	Direction
Answerability	0.72	−0.58	25,280	343,340	risk ↓
Intention Grounding	0.66	−0.59	13,576	355,044	risk ↓
Lack of Specificity	0.56	0.42	302,781	65,839	risk ↑
Excessive Details	0.43	0.30	361,479	7,141	risk ↑
Clause Complexity	0.40	0.27	307,849	60,771	risk ↑
Query–Scenario Mismatch	0.40	0.34	333,939	34,681	risk ↑
Anaphora Usage	0.28	0.14	287,833	80,787	risk ↑
Pragmatic Features	0.25	0.13	270,893	97,727	risk ↑
Polysemous Words	0.23	0.10	226,974	141,646	risk ↑
Rare Word Usage	0.22	0.09	329,378	39,242	risk ↑
Negation Usage	0.20	0.10	352,690	15,930	risk ↑
Subjectivity	0.20	0.13	356,043	12,577	risk ↑
Contextual Constraints	0.14	−0.06	126,145	242,475	risk ↓
Presupposition	0.12	0.04	47,411	321,209	risk ↑
Named Entities Present	0.11	−0.04	108,331	260,289	risk ↓
Superlative Usage	0.09	−0.02	338,308	30,312	risk ↓
Domain Specificity	0.07	0.00	71,623	296,997	risk↑

Table 5: **Feature ranking by ECDF separation.** KS and Δmedian computed on predicted $P(\text{Risky})$.

Feature	$n_{\textbf{Present}}$	$n_{\textbf{Absent}}$	**Overlap**	**ATE (IPW)**	**ATE (Strat.)**
Named Entities Present	260,289	108,331	1.000	-0.000	-0.001
Polysemous Words	141,646	226,974	1.000	+0.014	+0.016
Pragmatic Features	97,727	270,893	0.993	+0.007	+0.006
Contextual Constraints	242,475	126,145	0.983	+0.000	+0.003
Domain Specificity	296,997	71,623	0.979	+0.001	+0.006
Clause Complexity	60,771	307,849	0.969	**+0.103**	**+0.083**
Rare Word Usage	39,242	329,378	0.937	+0.015	+0.009
Anaphora Usage	80,787	287,833	0.918	**+0.059**	**+0.071**
Superlative Usage	30,312	338,308	0.890	**-0.032**	**-0.025**
Presupposition	321,209	47,411	0.838	+0.016	+0.007
Lack of Specificity	65,839	302,781	0.808	**+0.212**	**+0.199**
Query–Scenario Mismatch	34,681	333,939	0.488	+0.039	+0.012
Answerability	343,340	25,280	0.338	—	—
Negation Usage	15,930	352,690	0.338	—	—
Subjectivity	12,577	356,043	0.266	—	—
Intention Grounding	355,044	13,576	0.225	—	—
Excessive Details	7,141	361,479	0.126	—	—

Table 6: **Propensity overlap and overlap-conditioned uplifts by feature.** Overlap is the common-support share in $\pi_f(z)$ (Present vs. Absent). ATEs are percentage-point changes in $\Pr(\textsc{Risky})$ under IPW and propensity-stratified matching, reported only where overlap is adequate; **bold** marks salient non-zero effects. "—" indicates poor overlap (no uplift reported).

verse Probability Weighting (IPW) uplifts only where overlap is substantial (see App. Table 6).

IPW uplift. The inverse-probability-weighted (IPW) ATE for feature f on O is

$$\hat{\tau}_f^{\text{IPW}} = \frac{\sum_i \frac{T_{fi}}{\hat{\pi}_{fi}} O_i}{\sum_i \frac{T_{fi}}{\hat{\pi}_{fi}}} - \frac{\sum_i \frac{1-T_{fi}}{1-\hat{\pi}_{fi}} O_i}{\sum_i \frac{1-T_{fi}}{1-\hat{\pi}_{fi}}}.$$

IPW is interpretable as a quasi-causal contrast under unconfoundedness and positivity; where overlap is weak, we treat estimates as associational.

Matched (stratified) contrast. As a complementary, low-variance estimator, we stratify by $\hat{\pi}_{fi}$ quantiles into $K{=}10$ bins b and compute

$$\hat{\tau}_f^{\text{match}} = \sum_{b=1}^{K} \omega_b \big(\bar{O}_{1b} - \bar{O}_{0b} \big)$$

where $\omega_b \propto n_b$ and $\bar{O}_{tb}$ is the within-bin mean outcome for $T{=}t$ and n_b is the bin size. We report both $\hat{\tau}_f^{\text{IPW}}$ and $\hat{\tau}_f^{\text{match}}$ when overlap is adequate (App. Table 6).

Interpretation. These contrasts estimate the change in risky outcome associated with *feature presence*, conditional on the other query features and dataset/scenario mix. Practically, we trust uplift magnitudes only for features with strong overlap; for near-degenerate features (e.g., *Answerabil-*

ity, *Intention Grounding*), we report coefficients and ECDF gaps as correlational signals.

G Prompt Templates

Universal Feature Template

```
You are an expert linguist. Given a user query,
    decide whether it exhibits the FEATURE below
     using the operational rubric.
Return STRUCTURED OUTPUT with fields:
- label: true|false
- rationale: <=2 sentences (short, evidence-
    based)

FEATURE: {{FEATURE_NAME}}

OPERATIONAL RUBRIC:
- {{RUBRIC_BULLET_1}}
- {{RUBRIC_BULLET_2}}
- {{RUBRIC_BULLET_3}}

EXAMPLES (5-shot; mix of positive and negative):
[E1] FEATURE=Negation Usage; INPUT: Why didn't
    the test run?
OUTPUT: label=true; rationale="Contains explicit
     negation (didn't) affecting the main
    predicate."

[E2] FEATURE=Negation Usage; INPUT: Why did the
    test run?
OUTPUT: label=false; rationale="No negation
    markers present."

[E3] FEATURE=Lack of Specificity; INPUT: Tell me
     about Tesla.
OUTPUT: label=true; rationale="Multiple
    plausible scopes (company, vehicles, stock)
    with no constraints."

[E4] FEATURE=Lack of Specificity; INPUT:
    Summarize Tesla's 2024 Q4 earnings call in 5
     bullets.
OUTPUT: label=false; rationale="Time, scope, and
     format are clearly specified."

[E5] FEATURE=Named Entities Present; INPUT: Did
    the CDC issue RSV guidance in 2024?
OUTPUT: label=true; rationale="Contains named
    entity (CDC) and dated reference (2024)."

Now classify the following query.

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="<two short
    sentences>"
```

Anaphora Usage

```
You are an expert linguist. Given a user query,
    decide whether it exhibits the FEATURE below
     using the operational rubric.
Return STRUCTURED OUTPUT with fields:
- label: true|false
- rationale: <=2 sentences

FEATURE: Anaphora Usage

OPERATIONAL RUBRIC:
- Contains pronominal/definite references (it/
    this/that/they/he/she/these/those/that one)
    with an antecedent not locally introduced.
- Correct interpretation depends on prior
    discourse or missing antecedent.
- If read in isolation, resolution is unclear or
     ambiguous.

EXAMPLES (5-shot):
[E1] INPUT: Is he the same person who founded
    the company?  → STRUCTURED OUTPUT: label=
    true; rationale="'he' lacks an antecedent;
    resolution depends on prior context."
[E2] INPUT: How does this compare to that paper
    from last year? → label=true; rationale="'
    this' and 'that paper' require discourse to
    resolve."
[E3] INPUT: It was delayed again--when will it
    ship? → label=true; rationale="'It' is
    anaphoric with no antecedent in the query."
[E4] INPUT: Who founded Apple? → label=false;
    rationale="No anaphoric expressions; fully
    self-contained."
[E5] INPUT: Define photosynthesis. → label=
    false; rationale="No pronouns or anaphoric
    references."

Now classify the following query.

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="<two short
    sentences>"
```

Clause Complexity

```
You are an expert linguist. Decide whether the
    query exhibits the FEATURE below using the
    rubric.
Return STRUCTURED OUTPUT with fields {label,
    rationale (<=2 sentences)}.

FEATURE: Clause Complexity

OPERATIONAL RUBRIC:
- Contains multiple subordinate/relative/
    conditional clauses.
- Uses subordinators or relativizers (because,
    although, which/that, if/when/while, even
    though, so that).
- Meaning would materially change if reduced to
    a single clause.

EXAMPLES (5-shot):
[E1] INPUT: If the trial succeeds, which
    regulators will, according to the memo that
    leaked, approve it first? → label=true;
    rationale="Multiple embedded/conditional
    clauses."
[E2] INPUT: Summarize the study that was
    published last week, which compared three
    models. → label=true; rationale="Relative
    clauses 'that was published' and 'which
    compared'."
[E3] INPUT: Although sales fell, margins
```

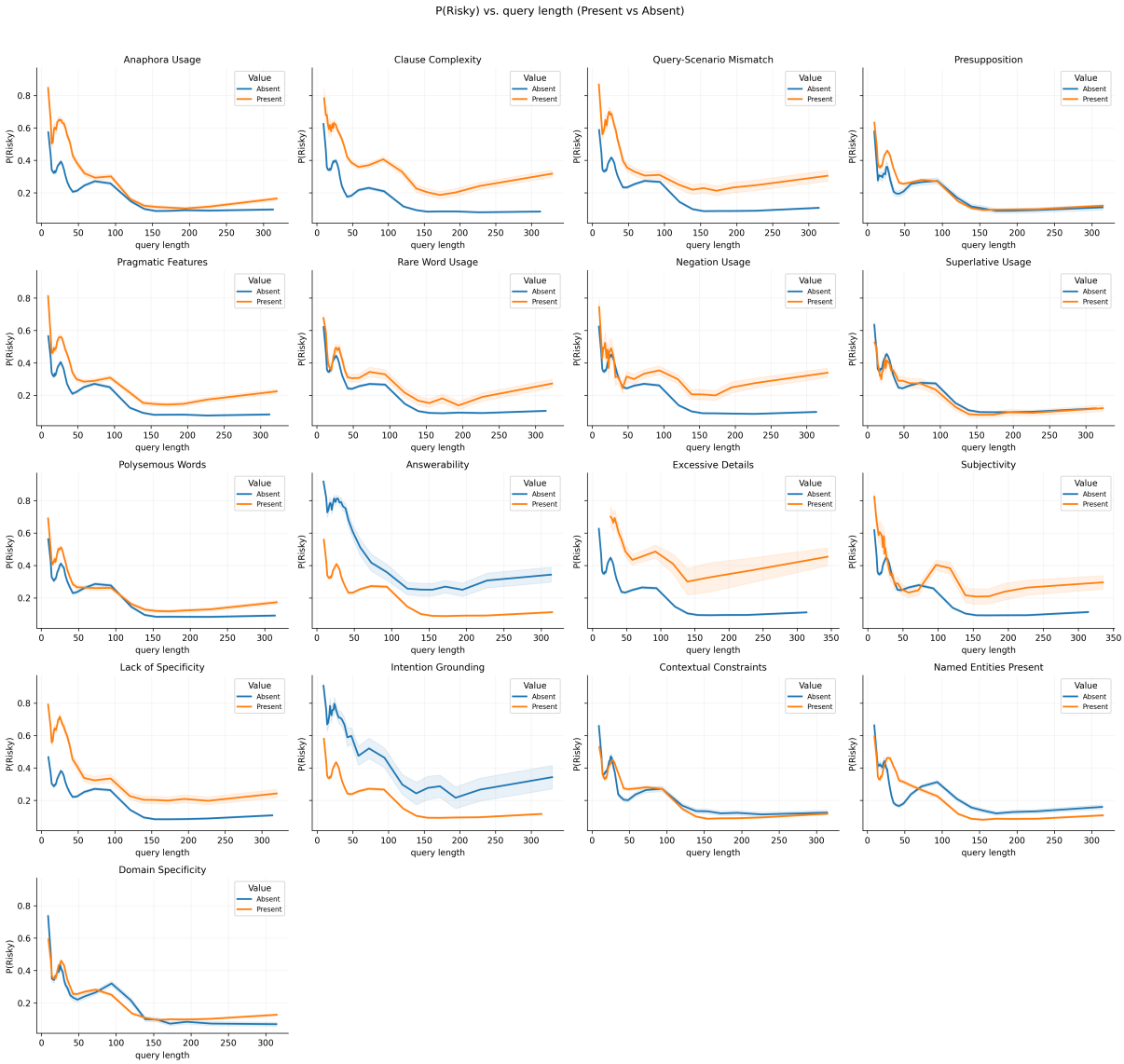

Figure 10: **Risk vs. query length (Present vs Absent).** For each feature, we bin queries into length quantiles and plot the empirical probability of the *Risky* label within each bin for *Present* vs *Absent*. Shaded bands are binomial CIs. Separation is largest at short lengths: features such as *Lack of Specificity* and *Excessive Details* increase risk, whereas *Answerability* and *Intention Grounding* reduce risk across lengths.

```
    improved. → label=true; rationale="
    Subordinate concessive clause."
[E4] INPUT: Who wrote The Road? → label=false;
    rationale="Single simple clause."
[E5] INPUT: Define GDP. → label=false;
    rationale="No subordination or embedding."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Query–Scenario Mismatch

```
You are an expert linguist. Decide whether the
    query is mismatched with the declared
    scenario.
Return STRUCTURED OUTPUT with fields {label,
    rationale (<=2 sentences)}.

FEATURE: Query--Scenario Mismatch

OPERATIONAL RUBRIC:
- Requested operation conflicts with SCENARIO (
    Extractive / Abstractive / Multiple-Choice).
- Expected answer format is incompatible with
    SCENARIO resources (e.g., asks for "exact
    span" but no passage; asks to "pick an
    option" but no options).
- The query presupposes inputs (choices/passage)
    absent in the scenario.

EXAMPLES (5-shot):
[E1] SCENARIO=Abstractive; INPUT: Extract the
    exact span containing the date. → label=
    true; rationale="Extraction request in
    Abstractive setting."
[E2] SCENARIO=Multiple-Choice; INPUT: Provide a
```

```
    free-form summary of the article. → label=
    true; rationale="Open summary in MC setting
    ."
[E3] SCENARIO=Extractive; INPUT: Choose the
    correct option (A--D). → label=true;
    rationale="MC instruction in Extractive
    scenario."
[E4] SCENARIO=Abstractive; INPUT: Summarize the
    passage in three bullets. → label=false;
    rationale="Matches Abstractive scenario."
[E5] SCENARIO=Multiple-Choice; INPUT: Select the
     best answer from the options. → label=
    false; rationale="Matches MC scenario."

INPUT:
SCENARIO: {{scenario}}
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Presupposition

```
You are an expert linguist. Decide whether the
    query embeds a nontrivial presupposition.
Return STRUCTURED OUTPUT with {label, rationale
    <=2 sentences}.

FEATURE: Presupposition

OPERATIONAL RUBRIC:
- Assumes some fact is true (existence/
    uniqueness/factivity) without evidence in
    the query.
- Removing the presupposition changes truth
    conditions (e.g., "When did X stop..."
    presupposes X used to...).
- The assumed fact may be false or unverifiable
    given typical inputs.

EXAMPLES (5-shot):
[E1] INPUT: When did the CEO admit the fraud? →
     label=true; rationale="Presupposes there
    was a fraud and an admission."
[E2] INPUT: Who is the king of France now? →
    label=true; rationale="Presupposes France
    has a king."
[E3] INPUT: Why did the model fail again? →
    label=true; rationale="Presupposes failure
    occurred previously."
[E4] INPUT: Who wrote Pride and Prejudice? →
    label=false; rationale="No hidden assumption
     beyond existence of the book."
[E5] INPUT: Define inflation. → label=false;
    rationale="No presupposed event/state."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Pragmatic Features

```
You are an expert linguist. Decide whether the
    query relies on pragmatics (implicature,
    deixis, indirect speech acts).
Return STRUCTURED OUTPUT with {label, rationale
    <=2 sentences}.

FEATURE: Pragmatic Features

OPERATIONAL RUBRIC:
- Literal form diverges from intended act (e.g.,
     "Can you pass the salt?" = request).
- Meaning depends on deixis ("here", "now", "
    this time") or shared situational context.
- Interpretation requires implicature/sarcasm/
    politeness beyond literal semantics.

EXAMPLES (5-shot):
[E1] INPUT: Could you maybe tone that down a bit
    ? → label=true; rationale="Indirect request
    , politeness strategy."
[E2] INPUT: It's cold in here. → label=true;
    rationale="Likely a request to close window/
    adjust temp (implicature)."
[E3] INPUT: Is that really how we want to do
    this? → label=true; rationale="Rhetorical/
    indirect suggestion."
[E4] INPUT: What is the capital of Japan? →
    label=false; rationale="Literal Q&A."
[E5] INPUT: Define entropy in thermodynamics. →
     label=false; rationale="No pragmatic
    inference required."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Rare Word Usage

```
You are an expert linguist. Decide whether the
    query uses rare/low-frequency or highly
    technical terms.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Rare Word Usage

OPERATIONAL RUBRIC:
- Contains niche jargon or low-frequency lexical
     items relative to general English.
- Common synonyms exist that would be much more
    frequent.
- A typical non-expert would flag the term as
    uncommon.

EXAMPLES (5-shot):
[E1] INPUT: Explain the pathophysiology of
    rhabdomyolysis. → label=true; rationale="'
    rhabdomyolysis' is rare, technical."
[E2] INPUT: Define syzygy in orbital mechanics.
    → label=true; rationale="'syzygy' is rare."
[E3] INPUT: What does heteroscedasticity mean?
    → label=true; rationale="Technical
    statistical term."
[E4] INPUT: What is a star? → label=false;
    rationale="Common vocabulary."
[E5] INPUT: Who was the first president of the
    US? → label=false; rationale="No rare words
    ."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Negation Usage

```
You are an expert linguist. Decide whether the
    query contains semantic negation.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Negation Usage

OPERATIONAL RUBRIC:
- Uses explicit negation tokens (not, no, never,
    without, hardly, scarcely).
- Negation scope changes the truth of the main
    predicate.
- Negative polarity is central to the request.

EXAMPLES (5-shot):
[E1] INPUT: Which vaccines are not mRNA-based?
    → label=true; rationale="Explicit negation
    'not' restricting set."
[E2] INPUT: Why didn't the test run? → label=
    true; rationale="Negated auxiliary 'didn't
    '."
[E3] INPUT: Summarize the paper without
    mentioning formulas. → label=true;
    rationale="'without' introduces negation
    constraint."
[E4] INPUT: Who wrote Hamlet? → label=false;
    rationale="No negation."
[E5] INPUT: Define polymerase. → label=false;
    rationale="No negation."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Superlative Usage

```
You are an expert linguist. Decide whether the
    query uses superlatives.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Superlative Usage

OPERATIONAL RUBRIC:
- Morphological/lexical superlatives (biggest,
    smallest, "the most/least", "of all").
- Implies an ordering over a set with an extreme
    endpoint.
- Expects a unique argmax/argmin or tie-breaking
    criterion.

EXAMPLES (5-shot):
[E1] INPUT: What is the fastest marine mammal?
    → label=true; rationale="Superlative '
    fastest'."
[E2] INPUT: Which city has the most museums? →
    label=true; rationale="'the most' indicates
    superlative count."
[E3] INPUT: What is the smallest prime number
    greater than 50? → label=true; rationale="'
    smallest' within a constrained set."
[E4] INPUT: Name a city with many museums. →
    label=false; rationale="Comparative/
    quantified, not superlative."
[E5] INPUT: Define prime number. → label=false;
    rationale="No superlative."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Polysemous Words

```
You are an expert linguist. Decide whether a key
    content word is polysemous and under-
    specified here.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Polysemous Words

OPERATIONAL RUBRIC:
- A salient word has multiple distinct senses (
    bank, cell, Java, Mercury).
- Local context does not disambiguate the
    intended sense.
- Different senses would change the answer.

EXAMPLES (5-shot):
[E1] INPUT: How do I open a new account at the
    bank? → label=false; rationale="Context
    favors financial institution."
[E2] INPUT: What is the weather like in Java? →
    label=true; rationale="Could be island or
    language; under-specified."
[E3] INPUT: Describe the function of a cell. →
    label=true; rationale="Could be biological
    cell or prison cell."
[E4] INPUT: Mercury's orbital period is what? →
    label=true; rationale="Planet vs. element;
    ambiguous."
[E5] INPUT: Who wrote The Hobbit? → label=false
    ; rationale="No polysemous ambiguity."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Answerability

```
You are an expert linguist. Decide whether the
    query is answerable on the basis of provided
    /commonly-known information (not speculation
    ).
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Answerability

OPERATIONAL RUBRIC:
- Has a verifiable answer given supplied context
    or widely-known facts.
- Not opinion-based, rhetorical, or forecasting
    without data.
- Does not require time-varying external info
    unless included.

EXAMPLES (5-shot):
[E1] INPUT: Who wrote The Road? → label=true;
    rationale="Single verifiable fact (Cormac
    McCarthy)."
[E2] INPUT: What is 17 × 19? → label=true;
    rationale="Deterministic computation."
```

```
[E3] INPUT: Will Stock X crash next month? → label=false; rationale="Speculative forecasting."
[E4] INPUT: Should I move to New York? → label=false; rationale="Subjective; no criteria."
[E5] INPUT: Is there life on Europa? → label=false; rationale="Unknown; not currently verifiable."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Excessive Details

```
You are an expert linguist. Decide whether the query includes extraneous details not needed to answer it.
Return STRUCTURED OUTPUT {label, rationale<=2 sentences}.

FEATURE: Excessive Details

OPERATIONAL RUBRIC:
- Contains descriptive asides that do not constrain the answer.
- Removing them would not change the target operation or output.
- Details distract or broaden scope without adding specificity.

EXAMPLES (5-shot):
[E1] INPUT: In my blue notebook from last summer's trip to Italy, can you define mitosis? → label=true; rationale="Notebook/trip details irrelevant to defining mitosis."
[E2] INPUT: Please, given my favorite mug and desk plant, what is 12 × 8? → label=true; rationale="Superfluous objects unrelated to arithmetic."
[E3] INPUT: When did WWI begin? → label=false; rationale="No extra details."
[E4] INPUT: Summarize this article in 3 bullets. → label=false; rationale="No extraneous info."
[E5] INPUT: What is the boiling point of water at sea level? → label=false; rationale="All details are relevant."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Subjectivity

```
You are an expert linguist. Decide whether the query requests a subjective judgment or preference.
Return STRUCTURED OUTPUT {label, rationale<=2 sentences}.

FEATURE: Subjectivity

OPERATIONAL RUBRIC:
- Invites personal taste/value judgment (best, beautiful, should, worth) without criteria.
- No objective rubric is provided to adjudicate correctness.
- Output depends on preferences rather than evidence.

EXAMPLES (5-shot):
[E1] INPUT: Which smartphone is the best right now? → label=true; rationale="'best' without criteria is subjective."
[E2] INPUT: Should I learn Rust or Go? → label=true; rationale="Advisory preference question."
[E3] INPUT: Is modern art good? → label=true; rationale="Value judgment."
[E4] INPUT: What's the battery capacity of iPhone 13? → label=false; rationale="Objective spec."
[E5] INPUT: Define convolution. → label=false; rationale="Objective definition."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Lack of Specificity

```
You are an expert linguist. Decide whether the query is under-specified.
Return STRUCTURED OUTPUT {label, rationale<=2 sentences}.

FEATURE: Lack of Specificity

OPERATIONAL RUBRIC:
- Missing disambiguating constraints (time/place/entity/scope).
- Multiple plausible interpretations; no tie-breaker.
- Task intent or output format is underspecified.


EXAMPLES (5-shot):
[E1] INPUT: Tell me about Tesla. → label=true; rationale="Company vs. cars vs. stock; scope unclear."
[E2] INPUT: Compare the models. → label=true; rationale="Which models? No domain or criteria."
[E3] INPUT: What happened yesterday? → label=true; rationale="No topic or domain given."
[E4] INPUT: Summarize Tesla's 2024 Q4 earnings call in 5 bullets. → label=false; rationale="Time, domain, and format specified."
[E5] INPUT: Extract the date of publication from the abstract. → label=false; rationale="Clear operation and target."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Intention Grounding

```
You are an expert linguist. Decide whether the user's intended operation is explicit.
Return STRUCTURED OUTPUT {label, rationale<=2 sentences}.
```

```
FEATURE: Intention Grounding

OPERATIONAL RUBRIC:
- Verb makes the operation clear (summarize,
    compare, extract, classify, translate).
- Expected output form is inferable (bullets,
    short answer, definition).
- Operation applies to supplied or implied
    content.

EXAMPLES (5-shot):
[E1] INPUT: Summarize the article in three
    bullets. → label=true; rationale="Clear
    directive and format."
[E2] INPUT: Extract the chemical formula from
    the passage. → label=true; rationale="
    Unambiguous extraction task."
[E3] INPUT: Compare Model A and Model B on
    latency and cost. → label=true; rationale="
    Operation and criteria stated."
[E4] INPUT: Java? → label=false; rationale="No
    operation specified."
[E5] INPUT: Tell me about space. → label=false;
     rationale="Vague goal, no operation."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Contextual Constraints

```
You are an expert linguist. Decide whether the
    query includes explicit constraints that
    narrow scope.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Contextual Constraints

OPERATIONAL RUBRIC:
- Names time, location, population, or
    conditions that meaningfully narrow the
    answer.
- Constraints are integral to fulfilling the
    request.
- Removing constraints would broaden or change
    the target.

EXAMPLES (5-shot):
[E1] INPUT: List three causes of inflation in
    the US during 2022. → label=true; rationale
    ="Time and location constraints."
[E2] INPUT: Summarize EU AI Act obligations for
    SMEs only. → label=true; rationale="
    Jurisdiction and population constraints."
[E3] INPUT: Give NYC subway delays after 10pm.
    → label=true; rationale="Location and time
    constraints."
[E4] INPUT: Define inflation. → label=false;
    rationale="No constraints."
[E5] INPUT: Summarize the article. → label=
    false; rationale="No narrowing conditions."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Named Entities Present

```
You are an expert linguist. Decide whether the
    query includes named entities (proper names)
    .
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Named Entities Present

OPERATIONAL RUBRIC:
- Contains proper names (persons, orgs, places,
    products, works, dates).
- Entities are pivotal to resolving the query.
- Generic categories alone (city, company) do
    not count as named entities.

EXAMPLES (5-shot):
[E1] INPUT: Did Sundar Pichai announce Gemini in
     2023? → label=true; rationale="Person and
    product names; year."
[E2] INPUT: What did the CDC advise about RSV in
     2024? → label=true; rationale="Org and
    year."
[E3] INPUT: When did World War I begin? → label
    =true; rationale="Named historical event."
[E4] INPUT: Who wrote that book? → label=false;
     rationale="No explicit names given."
[E5] INPUT: Define a balanced tree. → label=
    false; rationale="No proper names."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Domain Specificity

```
You are an expert linguist. Decide whether the
    query is specialized to a technical/
    professional domain.
Return STRUCTURED OUTPUT {label, rationale<=2
    sentences}.

FEATURE: Domain Specificity

OPERATIONAL RUBRIC:
- Requires discipline-specific knowledge/
    terminology (law, medicine, finance, ML,
    biology, etc.).
- A layperson would likely consult an expert/
    reference.
- Uses domain conventions or statutory/technical
     terms.

EXAMPLES (5-shot):
[E1] INPUT: Interpret Section 230 immunity in
    the context of platform moderation. → label
    =true; rationale="Legal doctrine."
[E2] INPUT: Differentiate Type I vs. Type II
    error and discuss α control. → label=true;
     rationale="Statistical terminology."
[E3] INPUT: Explain CRISPR-Cas9 off-target
    effects. → label=true; rationale="Molecular
     biology."
[E4] INPUT: What is a triangle? → label=false;
    rationale="Elementary concept."
```

```
[E5] INPUT: Who was Ada Lovelace? → label=false
     ; rationale="General knowledge biography."

INPUT:
{{query}}

STRUCTURED OUTPUT:
label=<true|false>; rationale="..."
```

Paraphrase Generation ($K \leq 6$)

```
You are a careful rewriter. Generate up to 6
     paraphrases that preserve the original
     meaning exactly while avoiding any added or
     lost constraints.
Return STRUCTURED OUTPUT with fields:
- paraphrases: list of strings (length 1...6)

RULES:
- Preserve all factual content, entities,
     temporal/locational constraints, and
     requested operation.
- Do not simplify away constraints or add new
     ones.
- Keep reading level and tone similar; avoid
     stylistic extremes.
- Each paraphrase must be self-contained and
     faithful.

EXAMPLES (5-shot):
[E1]
INPUT: Summarize the article in three bullets.
OUTPUT:
paraphrases=[
  "Provide a three-bullet summary of the article
     .",
  "Summarize the article using exactly three
     bullet points.",
  "Give a three-point bulleted summary of the
     article."
]

[E2]
INPUT: Extract the publication date from the
     abstract.
OUTPUT:
paraphrases=[
  "Identify the publication date contained in
     the abstract.",
  "From the abstract, extract the date of
     publication."
]

[E3]
INPUT: Compare Model A and Model B on latency
     and cost.
OUTPUT:
paraphrases=[
  "Compare Model A with Model B in terms of
     latency and cost.",
  "Evaluate Model A versus Model B focusing on
     latency and cost."
]

[E4]
INPUT: List three causes of inflation in the US
     during 2022.
OUTPUT:
paraphrases=[
  "List three causes of U.S. inflation in 2022.",
  "Provide three factors that drove inflation in
     the United States during 2022."
]

[E5]
INPUT: Translate the sentence into Spanish
     without adding explanations.
OUTPUT:
paraphrases=[
  "Translate the sentence to Spanish with no
     additional explanation.",
  "Render the sentence in Spanish only; do not
     include any commentary."
]

INPUT:
{{query}}

STRUCTURED OUTPUT:
paraphrases=["...", "..."]
```

Hallucination LLM-Judge

```
You are an impartial grader. Decide whether the
     model's answer is correct for the given
     query and gold reference.
Return STRUCTURED OUTPUT with fields:
- correct: true|false
- rationale: <=2 sentences (concise, cite the
     decisive mismatch or match)

CRITERIA:
- Accept semantic equivalence to the gold even
     if phrased differently.
- Penalize invented facts, contradictions, or
     unsupported specifics.
- If gold is a set, accept any member as correct.

- If the gold is not applicable or missing, mark
     correct=false and explain.

EXAMPLES (5-shot):
[E1]
INPUT:
query: Who wrote "The Road"?
model_answer: Cormac McCarthy.
gold: Cormac McCarthy
OUTPUT:
correct=true; rationale="Exact match to the
     reference author."

[E2]
INPUT:
query: What is the boiling point of water at sea
     level?
model_answer: 90°C.
gold: 100°C
OUTPUT:
correct=false; rationale="Numerical value
     contradicts the reference (90 ≠ 100)."

[E3]
INPUT:
query: Name one prime number greater than 10.
model_answer: 13.
gold: {11, 13, 17, 19, ...}
OUTPUT:
correct=true; rationale="Answer (13) is a valid
     member of the acceptable set."
```

```
[E4]
INPUT:
query: Define photosynthesis.
model_answer: It is the process by which plants
    convert light into chemical energy,
    producing glucose and oxygen from carbon
    dioxide and water.
gold: Process converting light energy into
    chemical energy, producing glucose and
    oxygen from CO2 and water.
OUTPUT:
correct=true; rationale="Semantically equivalent
     definition."

[E5]
INPUT:
query: Who is the current king of France?
model_answer: Louis XX.
gold: No current king of France.
OUTPUT:
correct=false; rationale="Asserts a non-existent
     monarch; contradicts the reference."

Now grade the following example.

INPUT:
query: {{query}}
model_answer: {{answer}}
gold: {{gold}}

STRUCTURED OUTPUT:
correct=<true|false>; rationale="<two short
    sentences>"
```

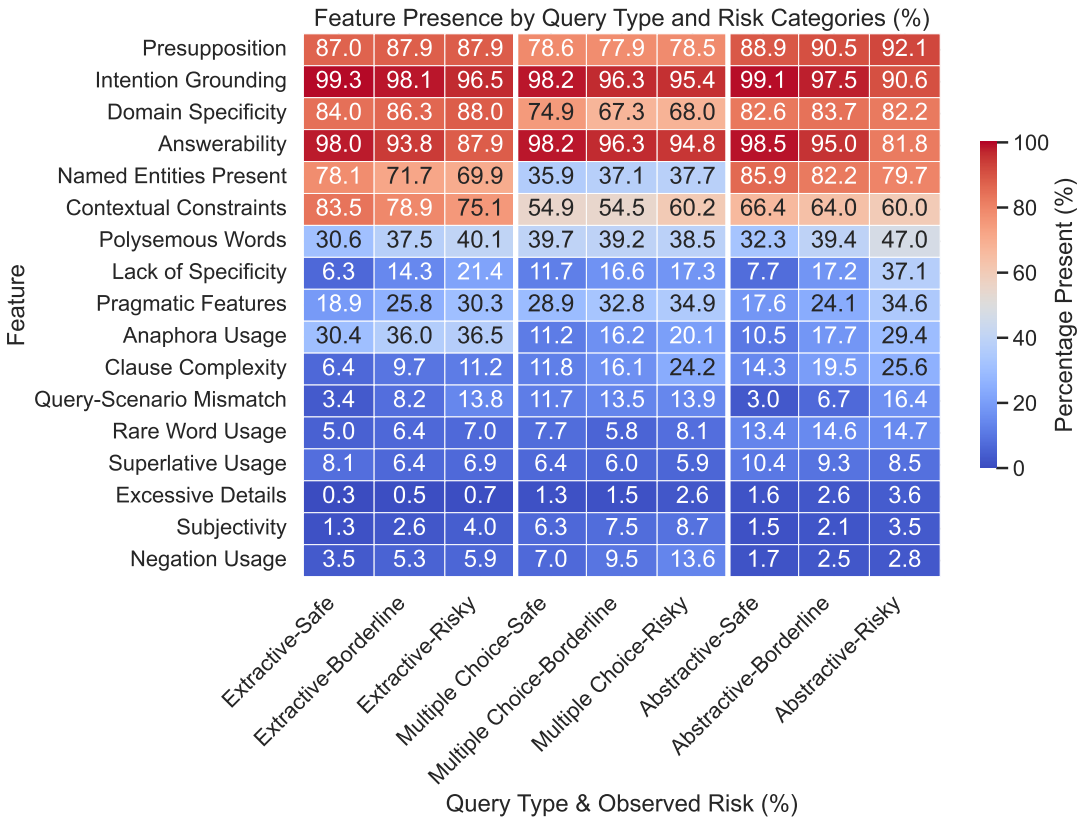

Figure 11: Heatmap of the percentage of queries exhibiting each binary linguistic feature, grouped by query type (extractive, multiple choice, abstractive) and categorized by observed risk level (*Safe*, *Borderline*, *Risky*). Warmer colors (reds) indicate higher prevalence of a feature, while cooler colors (blues) indicate lower prevalence. Several features (*lack of specificity*, *clause complexity*, *polysemous words*) increase most prominently from *Safe* to *Risky*, showing a clear monotonic rise in prevalence across risk categories. In contrast, *answerability* and *intention grounding* decrease steadily, and certain features (*domain specificity* and *contextual constraints*) display opposite trends across different query types.

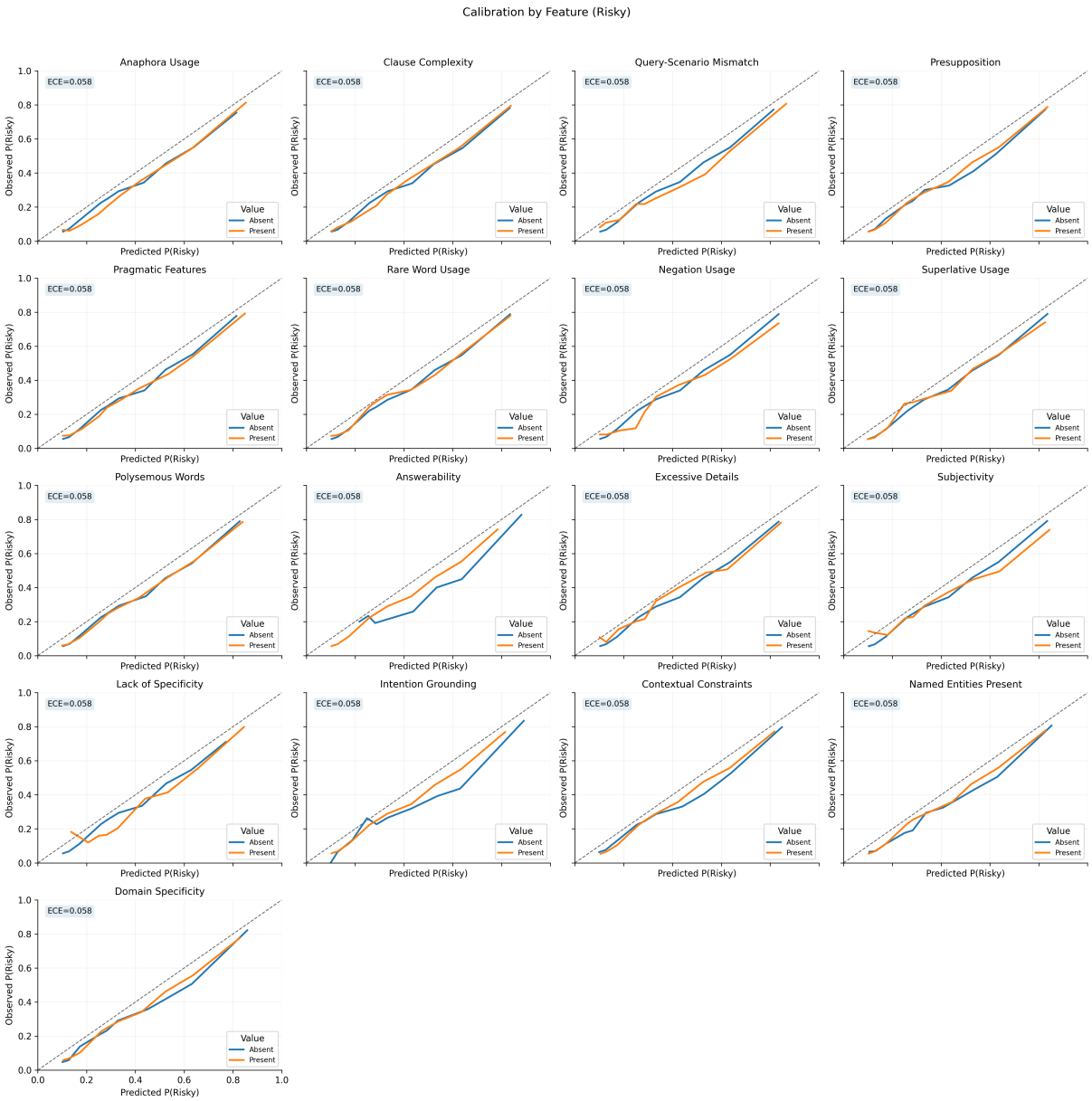

Figure 12: **Reliability by feature.** Binned reliability curves for predicted $P(\text{Risky})$ when a feature is *Present* vs *Absent*. Dashed line is perfect calibration. An overall ECE is reported per panel (10 equal-mass bins). Calibration across strata. We show reliability curves stratified by feature presence. The model is reasonably calibrated across strata (ECE $\approx$ 0.05–0.06). Importantly, the *direction* of miscalibration does not reverse between Present/Absent strata for the dominant features (e.g., Answerability, Lack of Specificity), supporting that the feature effects observed in the ECDFs translate to well-behaved risk scores rather than artifacts of calibration.

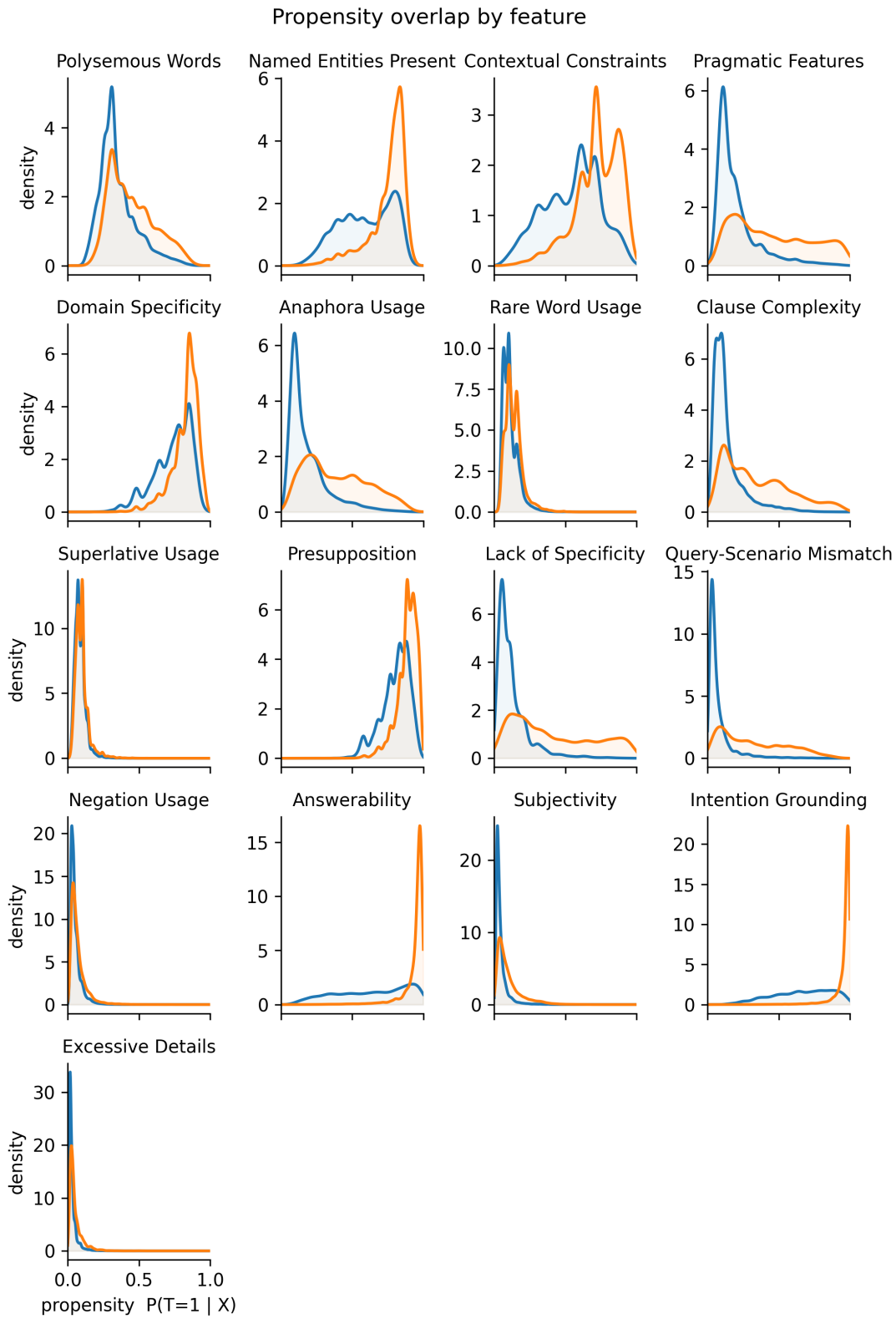

Figure 13: **Propensity overlap by feature.** For each feature f, we fit a logistic model $\pi_f(z) = \Pr(T_f{=}1 \mid Z_f{=}z)$ over $Z_f = (x_{-f}, \alpha, \gamma)$ and plot KDEs of $\hat{\pi}_f$ for *Present* ($T_f{=}1$) vs. *Absent* ($T_f{=}0$). Substantial overlap indicates adequate support for balancing or matching; near-degenerate propensities (mass near 0 or 1) warn that causal comparisons will be fragile. Several features (e.g., *Answerability*, *Intention Grounding*) show limited overlap, which we treat with weighting and sensitivity analyses.

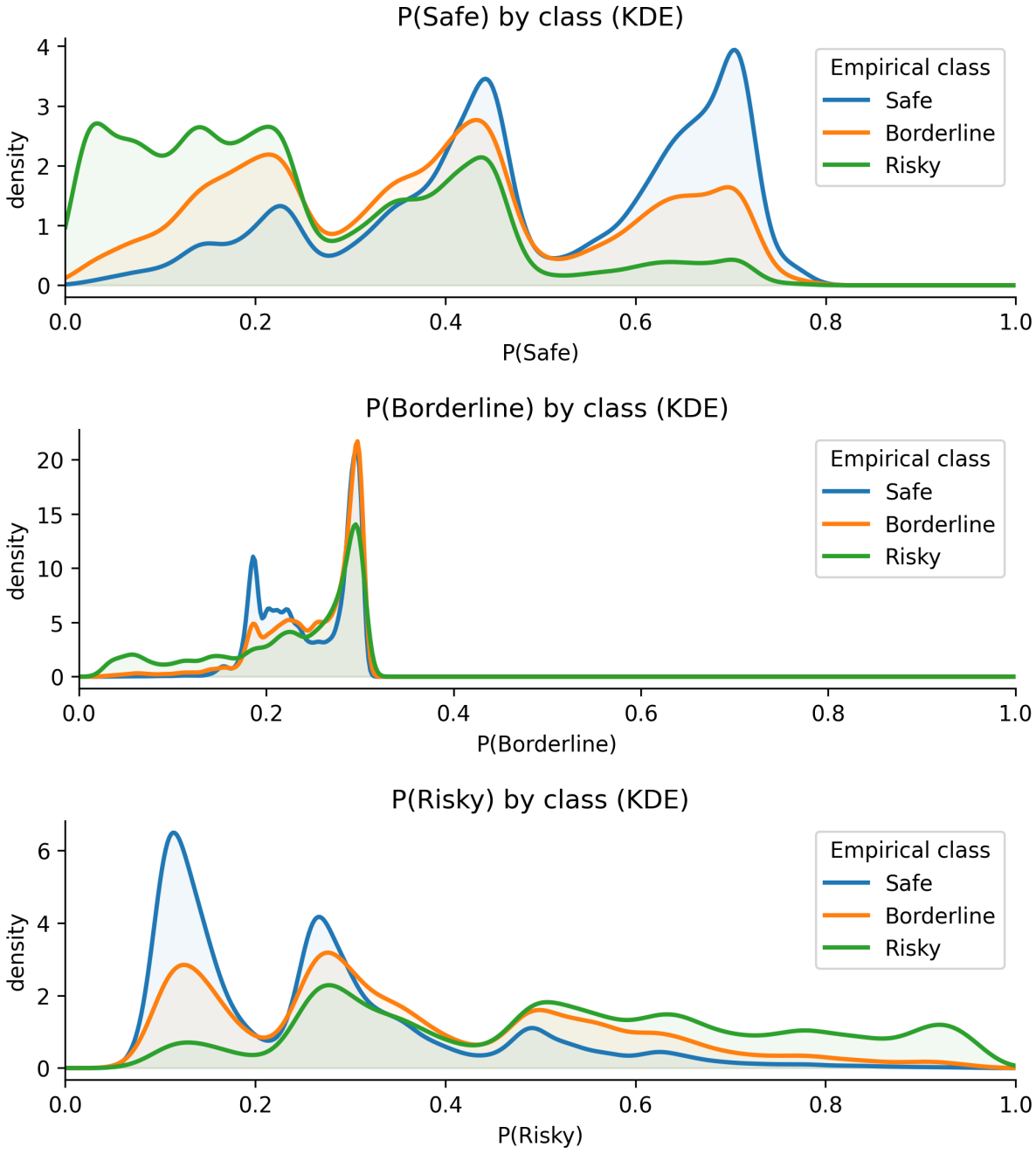

Figure 14: **Per-class probability KDEs.** KDEs of model-predicted $P(\text{Safe})$, $P(\text{Borderline})$, and $P(\text{Risky})$ grouped by empirical class labels.

	Feature	Definition	Example
Structural	Query and Context Length	Total number of tokens in the query (and context, when applicable)	"How does reinforcement learning work?" (6 tokens)
	Anaphoric Reference	Presence of pronouns or references requiring external context.	"What about that one?" (Unclear reference)
	Clause Complexity	Measures the presence of multiple subordinate clauses	"While I was walking home, I saw a cat that looked just like my friend's."
	Dependency Tree Depth	Depth of the query's syntactic dependency tree.	"Describe the structure of a sentence that contains multiple levels of embedding." (Depth: 7)
	Parse Tree Height	Height of the parse tree, providing a secondary measure of syntactic complexity.	"Analyze a sentence with nested relative clauses." (Height: 4)
Scenario-Based	Query Type	Extractive, Multiple Choice, or Abstractive.	"Summarize the latest economic report." (Requires retrieval)
	Query Scenario Mismatch	Mismatch between the query's intended output and its actual structure.	"List all prime numbers" (Infeasible output expectation)
	Presupposition	Unstated assumptions embedded in the query.	"Who is the musician that developed neural networks?" (Assumes such a musician exists)
	Pragmatics	Captures context-dependent meanings beyond literal interpretation.	"Can you pass the salt?" (A request, not a literal ability)
Lexical	Word Rarity	Use of rare or niche terminology.	"What are the ramifications of quantum decoherence?" (Uses low-frequency terms)
	Negation Usage	Presence of negation words (*not*, *never*).	"Is it not possible to do this?"
	Superlatives	Detection of superlative expressions (*biggest*, *fastest*).	"What is the fastest algorithm?"
	Polysemy	Presence of ambiguous words with multiple related meanings.	"Explain how a bank operates." (Ambiguity: financial institution vs. riverbank)
Stylistic	Answerability	Assesses whether the query has a verifiable answer.	"What is the exact number of galaxies?" (Unanswerable)
	Excessive Details	Evaluates whether a query is overloaded with information, potentially distracting the model.	"Can you explain how convolutional neural networks work, including all mathematical formulas?"
	Subjectivity	Detects the degree of opinion or personal bias in the query.	"What is the best programming language?"
	Lack of Specificity	Assesses the breadth or vagueness of a query.	"Tell me about history." (Too broad)
Semantic	Intention Grounding	Evaluates how clearly the query's purpose is expressed.	"How does reinforcement learning optimize control in robotics?" (Clear intent)
	Contextual Constraints	Identifies explicit constraints (time, location, conditions) provided in the query.	"What was the inflation rate in the US in 2023?"
	Named Entity Presence	Checks for the inclusion of verifiable named entities.	"Who founded OpenAI?"
	Domain Specificity	Determines whether the query belongs to a specialized domain (e.g., finance, law).	"What are the legal implications of the GDPR ruling?"

Table 7: Summary of our feature categories, definitions, and examples (See Section 3)

Feature	Presence	Question	Chain of Thought
Anaphora Usage	✓	Who was the guitarist for the English Rock band who Terry Kirkbride performed live in the studio with?	The question contains an anaphoric reference ('the English Rock band') without clear contextual information.
	✗	Isotopes are named for their number of protons plus what?	The question does not contain anaphoric references; it is a straightforward scientific inquiry.
Clause Complexity	✓	During evolution, something happened to increase the size of what organ in humans, relative to that of the chimpanzee?	The query has multiple clauses, increasing its complexity.
	✗	What do some animals do to adjust to hot temperatures?	The question is simple, consisting of a single clause.
Query-Scenario Mismatch	✓	What type of forested areas can be found on the highest terrace?	The query asks about 'forested areas,' but without a specific location or context, creating a mismatch.
	✗	What date in 2009 saw the heaviest UK snowfall since 1991?	The question has a direct and valid scenario, asking for a factual historical date.
Presupposition	✓	Central America's Panama seceded from which country in 1903?	The question presupposes that Panama seceded from a specific country in 1903.
	✗	What is the scientific name of the true creature featured in "Creature from the Black Lagoon"?	The question does not assume any prior knowledge; it is a straightforward request for a name.
Pragmatic Features	✓	Where did this pattern come from?	The meaning of 'this pattern' relies on prior discourse, making pragmatics necessary.
	✗	What is the name of plant-like protists?	The question does not rely on pragmatics; it seeks a factual term.
Rare Word Usage	✓	Where in the human body can you find the Trapezium bone?	The term 'Trapezium' is a less commonly known anatomical term.
	✗	What is an organism at the top of the food chain called?	The phrase 'apex predator' is well known and lacks rare words.
Negation Usage	✓	Which is not an inherited trait in humans?	The presence of 'not' reverses the expectation of the query.
	✗	Along with Walt Disney, who created Oswald the Lucky Rabbit?	The question is affirmative without negation.
Superlative Usage	✓	What is the first stage of cellular respiration?	The word 'first' introduces a superlative comparison.
	✗	Which river forms a natural border between Argentina and Uruguay?	No superlative forms are present in the query.
Named Entities Present	✓	What borough are the neighborhood of Chelsea and the office building, 10 Hudson Yards, both a part of?	Named entities include 'Chelsea' and '10 Hudson Yards.'
	✗	Some plants can detect increased levels of what when reflected from leaves of encroaching neighbors?	No specific named entities are present in the query.

Table 8: Representative queries illustrating the presence and absence of selected linguistic features, with accompanying chain-of-thought explanations (Part 1).

Feature	Presence	Question	Chain of Thought
Polysemous Words	✓	Who supervised the sting operation that implicated Evelyn Dawn Knight?	The word 'supervised' could have different meanings but in this context refers to oversight.
	✗	Which string instrument often played the basso continuo parts?	The terms 'string instrument' and 'basso continuo' are not polysemous in this context.
Subjectivity	✓	What is a criticism of other streaming services?	The query invites subjective responses based on personal opinions.
	✗	What is the second book in the Harry Potter series?	The question is factual and does not involve subjectivity.
Answerability	✓	How long was Warsaw occupied by Germany?	The question can be answered based on explicit historical data.
	✗	Beyoncé would take a break from music in which year?	The event may not have a definitive, verifiable answer.
Excessive Details	✓	SkyWest Airlines is a North American airline owned by SkyWest, Inc. and headquartered in which city in Utah, U.S., it flies as SkyWest Airlines in a partnership with Alaska Airlines?	The question includes excessive details about partnerships that are unnecessary for identifying the headquarters.
	✗	What is giving birth to dogs called?	The question is concise and does not contain excessive information.
Domain Specificity	✓	What is the term for a series of biochemical reactions by which an organism converts a given reactant to a specific end product?	The question is highly specific to biochemistry.
	✗	Fado is a type of folk music found in which country?	The question is not highly specialized; it relates to general cultural knowledge.
Lack of Specificity	✓	What division is the Canadian Army Doctrine of?	The query lacks clarity in defining what is meant by 'division.'
	✗	Winchester was the capital of which Anglo Saxon kingdom?	The question is specific in its historical context.
Intention Grounding	✓	Which of the two mines, Discovery Mine or Big Dan Mine, produced more gold?	The question is well-grounded in intent by seeking a clear comparison.
	✗	What are the two blocks of Catalan?	The intention is unclear due to the ambiguity of 'blocks.'
Contextual Constraints	✓	Which is the least densely populated county of England?	The question is constrained to a specific geographical location.
	✗	Who was the lyricist partner of Richard Rogers prior to Oscar Hammerstein?	No explicit constraints limit the question's scope.

Table 9: Representative queries illustrating the presence and absence of selected linguistic features, with accompanying chain-of-thought explanations (Part 2).