

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 MEASURING INTENT COMPREHENSION IN LLMs: A VARIANCE DECOMPOSITION FRAMEWORK

Anonymous authors

Paper under double-blind review

ABSTRACT

People judge interactions with large language models (LLMs) as successful when outputs match what they want, not what they type. Yet LLMs are trained to predict the next token solely from text input, not underlying intent. Because written language is an imperfect proxy for intent, and correlations between phrasing and desired outcomes can break down in training data, models that rely too heavily on surface cues may respond inconsistently to semantically equivalent prompts. This makes it essential to evaluate whether LLMs can reliably infer user intent—especially in high-stakes settings where robustness and generalization are critical. We introduce a formal framework for assessing intent comprehension in LLMs: whether a model demonstrates robust understanding of user intent by producing consistent outputs across semantically equivalent prompts while differentiating between prompts with distinct intents. Our evaluation approach is based on a variance decomposition of model responses into three components: variability due to user intent, user articulation, and model uncertainty. Models that understand what users want, and are not overly sensitive to textual cues, should attribute most output variance to intent differences, rather than articulation style. Applying this framework across diverse domains, we find that, within the five LLaMA and Gemma models we evaluate, larger models typically assign a greater share of variance to intent, indicating stronger comprehension of intent, although gains are uneven and often modest with increasing model size. These results motivate moving beyond accuracy-only benchmarks toward semantic diagnostics that directly assess whether models understand what users intend.

1 INTRODUCTION

Human communication involves a fundamental socio-cognitive process: we form an intent we want to articulate, then we choose words to express it, hoping the recipient will sympathetically decipher our intended meaning (Smith (1759)). Consider a frustrated traveler at an airport asking "Is there any way to get to Terminal B faster?" They might equally say "I need to catch my flight—what's the quickest route to B?" or "Terminal B is so far—any shortcuts?" They could even ask indirectly, "How do I get to the gate for flight 718?" Each phrasing differs dramatically, yet all express the same underlying intent: finding the fastest path to their destination. Much communication, therefore, centers on extracting underlying intent from observable signals, with success depending on the recipient's ability to see past surface variation to grasp the purpose beneath.

This challenge of intent extraction becomes particularly critical for large language models (LLMs), where all interaction occurs through text, making the model's ability to understand user intent the first step in effective human-AI interaction. Despite the centrality of intent understanding, most model evaluation approaches focus on whether models can perform specific tasks. These evaluation frameworks assume users express their intentions clearly, but this assumption often fails in practice. Users frequently struggle to articulate their needs and may revise their requests multiple times until they express their intent clearly. This implies that for models to be reliable from a user perspective, they must not simply respond to literal text, but rather infer the underlying intent from limited textual cues, disregard unimportant surface variations, and respond to the user's true purpose.

In this paper, we propose a framework for measuring how well models capture users' underlying intentions. We treat intent as a latent variable that underlies every prompt: while the same purpose

054 can be expressed in many ways, a model that understands intent should produce consistent response
 055 distributions across surface variations, yet shift appropriately when purpose changes. We define intent
 056 comprehension as the property that responses remain invariant to phrasing when purpose is fixed, but
 057 vary systematically when purpose changes.

058 To operationalize this idea, we present a diagnostic method that decomposes variation of model
 059 outputs into three components: Intent Sensitivity, the share of variation due to changes in purpose
 060 or intent; Articulation Sensitivity, the share due to phrasing variations; and Model Uncertainty,
 061 the residual variation stemming from the model’s inherent uncertainty. The variation in a model
 062 that understands user intent should exhibit high Intent Sensitivity and low Articulation Sensitivity,
 063 indicating consistent meaning extraction that disregards superficial linguistic cues.

064 Because intent is unobserved and generating equivalent prompts that differ only in form is non-trivial,
 065 we construct semantically equivalent prompts through cross-lingual translation, inspired by universal
 066 semantic structures across languages (Youn et al. (2016)). Starting from a base prompt, we translate
 067 it through a sequence of typologically diverse languages and back to English, inducing natural
 068 variation in phrasing while preserving intent. Translation serves this purpose well because its core
 069 objective aligns with our needs: altering surface form while maintaining meaning. Additionally,
 070 LLMs demonstrate strong automated translation capabilities, making this approach both theoretically
 071 motivated and practically feasible.

072 Using this pipeline, we evaluate five language models from the LLaMA and Gemma families of
 073 varying sizes across tasks in health, logistics, finance, travel, and social planning. Our findings
 074 show that within each family, the higher-parameter variants tend to attribute a larger share of output
 075 variability to changes in user intent (Intent Sensitivity), indicating stronger alignment with user
 076 goals and more coherent internal representations. However, this improvement is not uniform: the
 077 larger models do not consistently outperform smaller models across all domains, and its robustness
 078 advantages are often modest. Larger models also demonstrate greater sensitivity to prompt articulation,
 079 reflecting a trade-off between semantic generalization and responsiveness to surface cues.
 080 Furthermore, model robustness varies significantly across domains. These findings demonstrate our
 081 framework’s diagnostic value and highlight the need for more targeted evaluations of models’ ability
 082 to respond to user intent rather than varied textual cues.

083 **Intent in Computational Models.** The concept of intent has emerged as a fundamental construct
 084 in artificial intelligence to bridge observable behaviors and underlying cognitive states. In natural
 085 language processing, intent traditionally represents the underlying purpose or goal behind user
 086 utterances (Qin et al. (2021); Zhang and Wang (2022)), evolving from early slot-filling paradigms
 087 to sophisticated neural architectures (Louvan and Magnini (2020)). This aligns with philosophical
 088 foundations in Brentano’s intentionality theory—the directedness of mental states toward objects
 089 (Jacob (2019))—and Gricean pragmatics, which emphasizes the role of communicative intentions
 090 in meaning (Grice (1957; 1989)). Recent work has formalized these intuitions through a Bayesian
 091 calculus, most notably Baker, Tenenbaum, and Saxe’s inverse planning framework (Baker et al.
 092 (2007; 2009)), which models intent recognition as Bayesian inference over an agent’s goals given
 093 observations of their actions under an assumption of approximate rationality.

094 **Contributions** This paper makes three key contributions. First, we use these foundations to propose
 095 a formal definition of intent comprehension, specifically tailored to language models, grounded in the
 096 notion of response invariance to surface variations. Second, we introduce a variance decomposition
 097 method that quantifies the relative contributions of intent, phrasing, and uncertainty to model outputs,
 098 providing interpretable measures of model behavior. Third, we apply this method to real-world tasks
 099 across multiple domains, demonstrating how our framework reveals meaningful differences in the the
 100 ability of models to understand users intent.

101 2 CONCEPTUAL FRAMEWORK

102 In this section, we provide the conceptual foundation for our proposed measure in section 3. Let
 103 $T \in \mathcal{T}$ denote users’ intent. This intent or purpose represents the communicative goal underlying a
 104 user’s request to the model. Users articulate their intent through prompts p_i , which may vary widely
 105 in surface form even when seeking to convey the same underlying objective. For example, the intent
 106 of knowing the capital of France may be expressed through prompts such as “What’s the capital of

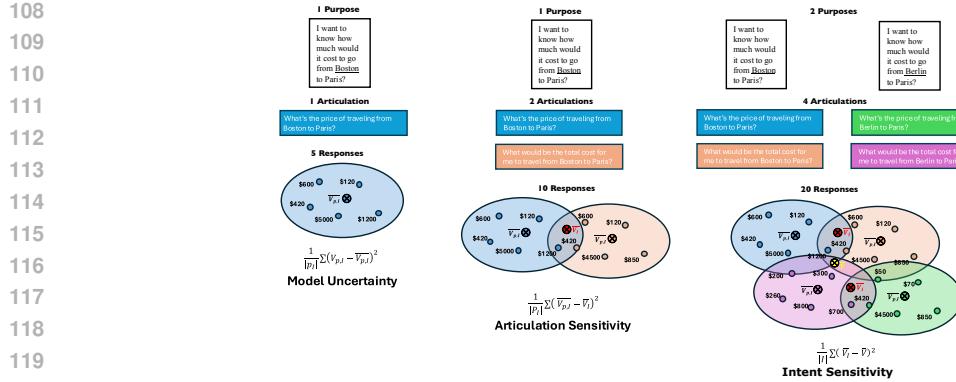


Figure 1: Conceptual Illustration of our decomposition Measures

France?” or “Which city is France’s capital?”. Prompts should be thought of as the full textual inputs provided to the model that captures the implied intent of the user. Finally, let $a \in \mathcal{A}$ be the model response.

We say that a language model understands the user’s intent if it produces identical response distributions for requests or prompts that express the same underlying intent, and distinct response distributions for prompts that express different intents. We formalize this as follows:

Definition 2.1 (Intent Comprehension) *Let $\tau : \mathcal{P} \rightarrow \mathcal{T}$ be the ground-truth mapping that assigns each prompt an intent label. A language model that generates a response distribution π over \mathcal{A} , is said to understand the user’s intent if, for every pair of prompts $p_i, p_j \in \mathcal{P}$, the following properties hold:*

1. **Consistency:** If $\tau(p_i) = \tau(p_j)$, then

$$\pi(\cdot | p_i) = \pi(\cdot | p_j).$$

2. **Sensitivity:** If $\tau(p_i) \neq \tau(p_j)$, then

$$\pi(\cdot | p_i) \neq \pi(\cdot | p_j).$$

Our definition requires a model that understands user intent to distinguish between surface-level variations in articulation and substantive changes in user intent or purpose. In particular, a model that varies its response distribution in reaction to inconsequential articulation changes is likely overfitting to superficial correlations in the training data rather than genuinely understanding the user’s underlying purpose.

Remark 1 Our definition of Intent Comprehension is directly tied to the broader notion of a world model. In robotics and model-based control, agents infer a latent state s_t and learn dynamics/observation maps (f, g) so that behavior depends on s_t rather than raw measurements. Performance is typically judged by correctness-centric criteria such as one-step prediction error, trajectory likelihood, or task return under a known oracle. In our text-only setting, the prompt stream serves as observations, and the payoff-relevant latent state is the user’s intent. As in robotics, a language model with a usable world model should (i) behave invariantly across paraphrases that preserve this state and (ii) shift its output distribution when the state (intent) changes. Crucially, unlike standard evaluations of world models—which assess both invariance across equivalent states and whether outcomes are correct relative to an external oracle—IC is consistency-centric: it tests whether the LLM can identify and respond to the latent state rather than the surface cues. IC therefore focuses on the first property of a world model: recognizing its latent state. See more in Appendix C.1.

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

Remark 2 In our framework, intent captures what the user wants the model to say—specifically, the expected distribution over exact responses. Two prompts have the same intent if they reliably elicit the same output distribution from the model, even if they differ in wording or domain. For example, “Will a fair coin land heads or tails? Answer 1 for heads, 0 for tails” and “I drew a number from [-1, 1]. Is it positive or negative? Answer 1 for positive, 0 for negative” differ in content but share the same intent: both aim to produce a 50/50 distribution over identical outputs—‘1’ and ‘0’.

Our criteria for whether a model has intent comprehension may be over-exacting. Precisely defining intent is inherently difficult due to its latency and potential ambiguity. Evaluating model responses adds another layer of complexity. This often requires estimation of the distribution of full-text responses, which are inherently high-dimensional (e.g., a detailed description of a driving itinerary from Boston to Miami.) Focusing on the distribution of responses, an evaluator may view distinct answers as carrying the same underlying meaning. For example, in response to the prompt “How much is $1 + 2$ ”, a model might answer “three” or “3.” Although slight variations in the prompt may influence the likelihood of generating “three” versus “3,” we should still regard the model as consistent.

To overcome these challenges, we propose a more relaxed notion of intent comprehension: a *sufficient* intent comprehension. This concept softens the standard definition by introducing an evaluator who judges prompts and responses. Specifically, we assume the existence of two functions: one that determines whether two prompts express the same intent, and another that assigns values to different responses. We say that a language model understands an the user intent if, whenever two prompts are judged to share the same intent, the distributions over evaluator-assigned values of responses remain identical. In this sense, we say that it’s sufficient to say that a model understands the user’s latent intent if the user cannot distinguish between model responses to the same intent.

Formally, let $\tilde{\tau} : \mathcal{P} \rightarrow \mathcal{T}$ map from prompts to intents (i.e., articulations to purposes), defined according to some evaluation criterion. Let $V : \mathcal{A} \rightarrow \mathbb{R}$ be a function that maps responses to values. Finally, let $\pi_V(\cdot | p)$ denote the distribution over response values induced by model responses to prompt p , as evaluated by V . The relaxed definition is as follows:

Definition 2.2 (Sufficient Intent Comprehension with respect to an evaluator $\tilde{\tau}$ and V) Let $\tilde{\tau} : \mathcal{P} \rightarrow \mathcal{T}$ be a fixed intent–evaluator that assigns an intent label for each prompt, and let $\pi_V(\cdot | p)$ be the value distribution induced by the model, in response to prompt p . A language model possesses a sufficient intent comprehension with respect to $\tilde{\tau}$ and V if for every pair of prompts $p_i, p_j \in \mathcal{P}$:

1. **(Sufficient Consistency)** If $\tilde{\tau}(p_i) = \tilde{\tau}(p_j)$, then

$$\pi_V(\cdot | p_i) = \pi_V(\cdot | p_j).$$

2. **(Sufficient Sensitivity)** If $\tilde{\tau}(p_i) \neq \tilde{\tau}(p_j)$, then

$$\pi_V(\cdot | p_i) \neq \pi_V(\cdot | p_j).$$

Compared to Definition 2.1, this relaxed notion makes two changes: the intent map $\tilde{\tau}$ is evaluator–dependent, and consistency is enforced only on the pushforward distributions of evaluated responses $V(a)$, rather than on the full response distributions over \mathcal{A} .

Remark 3 Both Intent Comprehension and sufficient Intent Comprehension criteria focus on measuring consistency, rather than assessing whether a language model’s responses are factually correct. This is because our primary goal is to evaluate how well the model understands user intent—not whether it produces the correct answer. This approach differs from traditional benchmarks, which typically rely on a binary notion of correctness, classifying answers as strictly right or wrong. To understand why consistency matters, consider a language model trained only on economic data from 2020. If asked about economic conditions in 2023, it may consistently produce outdated yet internally coherent answers across differently phrased questions, reflecting the information on which it was trained. While these answers are incorrect in light of present-day facts, their internal consistency suggests that the model is responding to the user’s underlying intent, and not simply their textual variation.

Our definition of a (sufficient) intent comprehension is motivated by LLM training data, which consists of human-generated text, replete with spurious correlations between writing style and user intent. Small stylistic changes—like tone or word choice—often co-occur with shifts in latent attributes of the writer, such as personality, mood, or communicative norms, even when the underlying intent and objective remains constant. This makes prompt phrasing a confounded signal, blending true intent with incidental articulation. As a result, a model may change its responses not because users change their intent, but because it has learned correlations between superficial cues and different responses. Evaluating whether an LLM understands the latent user intent, thus mirrors the causal inference problem: Can it separate variation due to intent from variation due to articulation? In a hypothetical world where prompt phrasing perfectly reveals intent and carries no noise, models would trivially satisfy our definition. But in reality, disentangling intent from correlation is non-trivial—making consistency a meaningful test of deeper understanding.

3 MEASURING (SUFFICIENT) INTENT COMPREHENSION

In this section, we present a simple method, motivated by our conceptual framework, for evaluating whether an LLM understands the user’s latent intent. We assume access to a sample of prompts along with corresponding model responses. Some prompts are designed to share the same underlying intent, while others are not. In the experimental section, we provide a detailed description of how we construct such a sample.

Our evaluation approach is based on a variance decomposition of model responses. Specifically, we break down response variance into three components: *Intent Sensitivity* (IS), which captures the model’s responsiveness to changes in the intent or purpose of the input prompt; *Articulation Sensitivity* (AS), which reflects the model’s sensitivity to variations in how users express the same intent; and *Model Uncertainty* (MU), which accounts for the inherent ambiguity or variability in the model’s responses.

To formalize these components, we consider a domain D consisting of related purposes or intents $I \in D$. Each intent can be expressed through multiple prompts $p_I \in \tau(I)$, and we denote the model’s value of response to a prompt p as v . To enable comparability across tasks, domains, and models, we define the standardized response as $\tilde{v} = \frac{v}{\text{std}(v|D)}$. Furthermore, let R_I^2 be the standard coefficient of determination, the proportion of total variance explained by differences in intent, and $R_{p_I}^2$ be the proportion of variance explained by differences in prompt phrasing, holding intent I fixed. With this notation, we define our three core measures, sketched in Figure 1, as:

$$\begin{aligned} IS &= \text{Var}_I(\mathbb{E}[\tilde{v}|I]) = R_I^2 && \text{(Intent Sensitivity),} \\ AS &= \mathbb{E}_I [\text{Var}_{p|I}(\mathbb{E}[\tilde{v} | p_I])] = \mathbb{E}_I [R_{p_I}^2] && \text{(Articulation Sensitivity),} \\ MU &= \mathbb{E}_I [\mathbb{E}_{p|I}(\text{Var}(\tilde{v} | p_I))] = \mathbb{E}_I [1 - R_{p_I}^2] && \text{(Model Uncertainty).} \end{aligned}$$

The first term, IS, is the variance of the conditional expectation of responses given intent. It reflects the sufficient Intent Comprehension property in 2.2, quantifying how much the model’s mean response shifts with changes in the user’s underlying purpose. A model appropriately sensitive to intent should produce distinctly different responses for different tasks. Therefore, we expect IS to be high in models that understand the user’s latent intent or purpose. The second term, AS, corresponds to the consistency property. It quantifies how much, on average, the model’s mean response varies with changes in phrasing or wording of the prompt, holding the user’s purpose fixed. A model that understands the user’s latent intent, rather than being swayed by syntactic or stylistic differences, should exhibit low sensitivity to surface-level changes in articulation.

The third term, MU, captures the residual variance that remains after conditioning on both intent and phrasing. This component conflates at least three sources: (i) sampling stochasticity (e.g., due to temperature or nucleus sampling); (ii) epistemic uncertainty, where the model lacks a confident internal representation and spreads probability mass over multiple plausible answers; and (iii) aleatoric uncertainty, which is inherent to the task itself (e.g., when the prompt is “Tell me a random joke”). Whether a high MU is desirable depends on the context. For deterministic tasks with a clear correct answer (e.g., “What is 1+1?”), high MU reflects unwanted uncertainty. In contrast, for inherently open-ended or subjective tasks—such as those we examine in the experimental section—some degree of response variability is expected and even appropriate.

270 These three measures are all non-negative and sum to one: $IS + AS + MU = 1$. This identity
 271 ensures that each term represents the relative contribution of a distinct source of variation in model
 272 responses. Together, they provide a principled way to assess the robustness and interpretability of a
 273 model’s behavior.

274 An alternative perspective on this decomposition is in terms of R^2 : how much of the variation in
 275 model responses is predictable from different parts of the request. From this viewpoint, the normative
 276 target is clear: *there should be essentially no predictive signal in the paraphrase once intent is*
 277 *fixed*, and nearly all systematic variation should come from changes in intent. While we use an
 278 R^2 -style “explained variation” framing to connect directly to the equations, the same idea can be
 279 operationalized with other performance metrics better matched to the output type—for example
 280 AUC for binary decisions, accuracy/F1 for discrete labels, or log loss/cross-entropy for probabilistic
 281 outputs. We emphasize the variance decomposition because it is model-agnostic, easy to implement,
 282 and intuitive to interpret, but other approach to operationalize the intent comprehension measures are
 283 possible.

284 We also define two summary statistics that condense the decomposition into intent-centric diagnostics.
 285 The first is the Meaningful Variability Share (MVS),

$$287 \quad MVS = \frac{IS}{IS + AS},$$

$$288$$

$$289$$

290 which serves as a signal-to-noise ratio over the *explainable* portion of variability: among the variance
 291 attributable to request features—either genuine changes in intent (IS) or superficial changes in
 292 articulation (AS)—MVS measures how much is signal rather than surface noise. A high MVS
 293 indicates that most of the model’s explainable variation reflects meaningful distinctions between user
 294 intents, whereas a low MVS suggests the model is overly influenced by superficial differences in
 295 wording, indicating a lack of intent understanding. Another interpretation is that, when decoding
 296 is effectively deterministic (temperature near zero) so that residual randomness is negligible, MVS
 297 approximates the two-way split between intent and articulation, because variability then stems only
 298 from intent and phrasing.

299 To align directly with our definition of sufficient intent comprehension—distributions should move
 300 with intent (sensitivity) and remain invariant to non-intent articulation (consistency)—we combine
 301 MVS with the overall share of total variance due to intent into a single index,

$$302 \quad ICI = \frac{2 MVS \cdot IS}{MVS + IS} \in [0, 1],$$

$$303$$

$$304$$

305 which we call the *Intent Comprehension Index (ICI)*. ICI is the harmonic mean of “intent purity” and
 306 “intent coverage.” The MVS component captures the sufficient consistency by penalizing articulation-
 307 driven variability: if non-intent phrasing moves the distribution, AS rises, MVS falls, and so does
 308 ICI. The IS component captures the sufficient sensitivity clause by rewarding models whose outputs
 309 shift when intent changes: if the distribution barely moves with intent, IS is small and ICI again falls.
 310 Therefore, the ICI summarizes the definition of intent comprehension.

311 This construction mirrors the classic precision–recall analogy in information retrieval. The *MVS*
 312 term plays the role of precision because, among the variance that is systematically explained by
 313 features of the request (intent or articulation), it measures the fraction that is aligned with intent rather
 314 than surface form. In this analogy, IS is the analogue of “true positives,” while AS acts like “false
 315 positives”: variance that is explained by the request, but driven by articulation rather than intent. The
 316 recall-like quantity, $IS/(IS + AS + MU) = IS$, instead asks what fraction of *all* observed variability
 317 in responses is driven by intent; the remaining components, $AS + MU$, are then analogous to “false
 318 negatives,” capturing variation that does *not* move with changes in intent—either because the model
 319 is reacting to non-intent phrasing (AS) or because its behavior is intrinsically noisy (MU). Read this
 320 way, ICI is an F1-type score over units of variance: it is large exactly when intent-driven variation is
 321 both pure (high MVS) and covers a large share of total variability (high IS).

322 **Intuitive interpretation.** Our decomposition asks a simple question: when the model’s answers
 323 change, why do they change? Intent Sensitivity captures variation driven by genuine differences in
 what users are trying to do. Articulation Sensitivity captures variation driven by superficial differences

324 in how users phrase the same request. Model Uncertainty is the residual: randomness from sampling,
 325 uncertainty in the model’s beliefs, or genuine task ambiguity. The three components sum to one,
 326 forming a “pie chart” of the model’s behavior. A model that truly understands intent will place most
 327 of the mass on IS, little on AS, and only as much MU as the task naturally requires. The ICI is high
 328 exactly when responses are primarily driven by intent rather than phrasing or noise.
 329

330 **Remark 4** *In practice, when estimating the components of the decomposition, we deliberately*
 331 *restrict attention to domains where requests share a common structure but differ in intent. This*
 332 *controlled setup reflects a conservative and more stringent approach to evaluating whether the*
 333 *model understands the user’s intent. By fixing the overall structure of the query and varying*
 334 *only a key element that changes the underlying intent (such as modifying the income level in*
 335 *a tax-related prompt while keeping the rest of the wording unchanged), we eliminate spurious*
 336 *cues and force the model to rely on its understanding. If the model systematically adapts its*
 337 *output to such minimal yet meaningful changes, it suggests the presence of a coherent internal*
 338 *representation capable of capturing user intent and purpose.*

339 Our variance decomposition is one concrete operationalization of sufficient intent comprehension,
 340 but it is not the only way to quantify whether a model is responding to *intents* rather than to surface
 341 form. Conceptually, the core behavioral question we ask is whether prompts that share the same
 342 intent but differ in articulation (paraphrases, typos, different levels of detail) induce similar response
 343 distributions, while prompts that differ in intent induce systematically different distributions. The
 344 IS–AS–MU decomposition captures this by attributing variation either to changes in intent (IS), to
 345 changes in articulation conditional on intent (AS), or to residual model uncertainty (MU). Other
 346 scoring rules or distances over response distributions could in principle be plugged into the same logic;
 347 the key desideratum is that they separate variation due to intent from variation due to articulation and
 348 noise.

349 In the Appendix, we show how our decomposition framework extends to settings where the model
 350 produces free-form text (Appendix ??) or discrete, categorical outputs (Appendix ??). The key
 351 challenge in the discrete case is that variance is not naturally defined—any numeric encoding
 352 of categories is arbitrary. We address this by replacing the variance of the outcome v with its
 353 Shannon entropy $H(v)$. Using the chain rule for entropy, we derive an analogous decomposition
 354 $H(v) = IS + AS + MU$. Here, IS becomes the mutual information $I(I; v)$ between intent and output;
 355 AS becomes a conditional mutual information term that captures the additional information about v
 356 conveyed by the articulation p_I beyond intent; and MU becomes the remaining conditional entropy
 357 $H(v | I, p_I)$, representing residual unpredictability. Normalizing by $H(v)$ again yields unit-sum
 358 shares that play the same role as in our variance-based decomposition. Appendix ?? provides an
 359 illustrative application in which the model is tasked with ranking and approving loans and returns
 360 only categorical decisions, demonstrating how the entropy-based decomposition operates in practice.

362 4 EXPERIMENTS

364 In this section, we provide a brief description of how we construct our experiments; a detailed
 365 description is presented in Section F of the Appendix. We begin by designing an automatic question
 366 generator. Our questions focus on open-ended, guesstimation-style tasks. These are chosen because
 367 they naturally elicit a wide distribution of plausible responses, which is essential for disentangling
 368 intent sensitivity from articulation sensitivity and model uncertainty. We utilize an LLM (GPT-4.1) to
 369 construct 24 questions across five domains—transportation, personal finance, health and nutrition,
 370 logistics, and social planning—via a two-stage pipeline: first, we generate unit-specified templates
 371 with placeholders, and then we populate them with realistic values to define specific intents. For each
 372 task, we generate 12 distinct intents. This procedure yields a diverse and structured set of quantitative
 373 estimation tasks that reflect naturalistic usage while remaining well controlled.

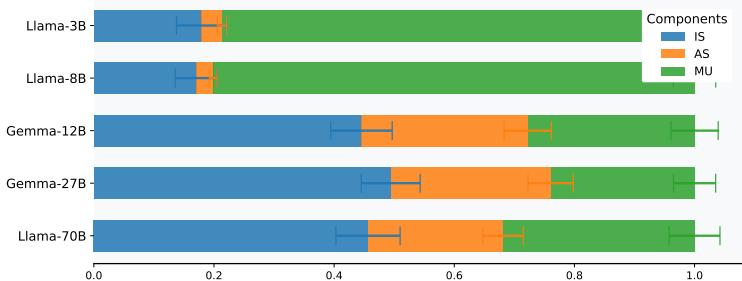
374 To evaluate articulation sensitivity, we need to generate sets of prompts that convey the same intent.
 375 This is non-trivial, as creating prompts with exactly equivalent definitions is difficult even for humans.
 376 To address this, we generate equivalent prompts through cross-lingual translation chains, leveraging
 377 structural differences across languages to induce lexical and syntactic diversity while preserving
 meaning. GPT 4.1-based checks filter paraphrases to ensure intent equivalence, and we then select a

378 diverse final set using an embedding-based approach. Each paraphrase is paired with varying input
 379 values and repeatedly posed to different LLMs. Specifically, for each of the 24 tasks, we generate 10
 380 paraphrases. We then submit these prompt variations to five different language models—LLaMA
 381 3.2 3B, LLaMA 3.1 8B, LLaMA 3.3 70B (Aaron Grattafiori et al. (2024)), and Google’s Gemma 3
 382 12B-IT, and 27B-IT (Gemma Team: Aishwarya Kamath et al. (2025a))—requesting 25 responses
 383 for each prompt–intent pair. Unless otherwise noted, we fix the sampling temperature at 1 for all
 384 models. This places each model in a single, untuned “natural” decoding regime, so that differences in
 385 IS/AS/MU reflect the models themselves rather than model-specific temperature tuning; Appendix G
 386 explores other temperature setting. In total, we obtain 1,772,492 model responses, averaging 354,499
 387 per model¹. Finally, we perform the variance decomposition described in Section 3. To mitigate
 388 finite-sample issues, we apply ANOVA for the variance decomposition, which we discuss further in
 389 Section D of the Appendix. Human evaluation, described in Appendix B, verifies that our paraphrases
 390 largely preserve intent and that our numeric extraction is reliable, with high inter-annotator agreement
 391 for both checks.

392 4.1 RESULTS

393 We now turn to explore our results. Figure 2 presents our key findings across models. The figure
 394 shows the average Intent Sensitivity (IS), Articulation Sensitivity (AS), and Model Uncertainty (MU)
 395 across all tasks and across models. The figure reveals that the two smallest models, LLaMA 8B
 396 and LLaMA 3B, exhibit very high MU across tasks, accompanied by both low IS and low AS. This
 397 is accompanied by higher variance in general, as shown in Figure 7. In contrast, the three larger
 398 models—LLaMA 70B, Gemma 27B, and Gemma 12B—show comparable results across these three
 399 components, with the largest share attributed to IS. LLaMA 70B shows slightly higher MU. Within
 400 our evaluation, this indicates that models with more parameters in these families are not uniformly
 401 superior, and that high IS can be achieved by a 12B-parameter Gemma model despite its smaller size
 402 relative to LLaMA 70B. As the two families differ in training practices and training data this indicate
 403 that high IS can be achieved within smaller models.

404 Figure 3² presents our
 405 results for the Intent Com-
 406 prehension Index (ICI),
 407 Meaningful Variability
 408 Share (MVS), and IS
 409 across models. The blue
 410 line illustrates that, in
 411 this five-model set, the
 412 higher-parameter LLaMA
 413 and Gemma variants tend to
 414 yield higher IS, indicating
 415 that changes in user inten-
 416 tion are more effectively
 417 reflected in adjustments to
 418 the response distribution.
 419 Specifically, the smaller



420 Figure 2: The Average Variance Decomposition Across Models, shaded
 421 area is 95% CI

422 LLaMA models account for approximately 20% of the response variation due to changes in intention,
 423 while the larger models account for around 50%. Because these models differ in architecture and
 424 training regimes as well as parameter count, we treat this as a descriptive association rather than a
 425 causal scaling law.

426 The second component, MVS, shown by the orange line, demonstrates an inverse pattern: the two
 427 smaller models exhibit higher MVS than their larger counterparts. This implies that smaller models
 428 respond relatively more to changes in intent than to changes in prompt articulation, aligning with the
 429 general lower responsiveness suggested by their Intent Sensitivity scores. Interestingly, larger models,
 430 in are evaluation, show greater sensitivity to articulation shifts. This suggests that some larger models
 431 are more sensitive to prompt text, therefore more responsive to intent, but at the cost of greater overall

¹We only keep track of valid responses, but sometimes the model fails to reach the desired set of responses within 125 attempts, and in such cases, we move on to the next section.

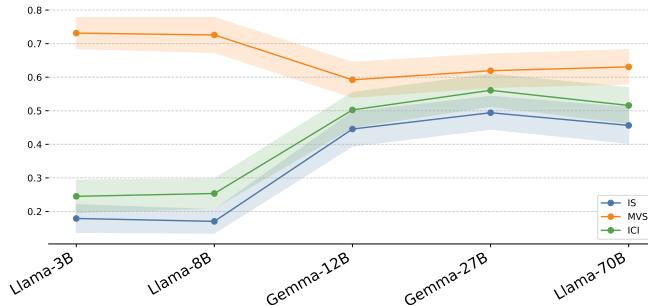
²Full distribution of the component for each tasks, by model, is in figure 8 in the Appendix

432 susceptibility to surface-level prompt variations, reflecting overfitting rather than an emerging deeper
 433 understanding in larger models.
 434

435 Finally, examining the Intent Comprehension Index, we observe higher values for larger models,
 436 but the improvement appears to plateau—the difference between the 70B model and the 12B model
 437 is modest. This suggests that intent comprehension is not necessarily an emergent property that
 438 scales linearly with model size, and that substantial improvements in understanding user intent can
 439 be achieved without requiring the largest available models.

440 In Figure 4, we examine heterogeneity in model performance across the five topical areas. The results indicate that larger models
 441 are not uniformly superior across domains; instead, their sensitivity varies by
 442 topic. For example, the largest model, LLaMA 70B, achieves the highest IS
 443 scores in Health and Nutrition and Transportation, whereas LLaMA 27B per-
 444 forms best in Logistics, Personal Finance, and Social
 445 Planning. The figure further highlights differences in the share of AS: questions in Social Planning
 446 are consistently more sensitive to writing style across models than those in Health and Nutrition,
 447 Transportation, or Personal Finance. Taken together, these findings suggest that model capability is
 448 domain-dependent. In this way, intent comprehension is not a uniform property of models but rather
 449 varies systematically across topical areas.

450 The differences we observe across models reflect not only scale but also model family. These family-
 451 level gaps are more plausibly attributed to training procedures than to data access: both LLaMA
 452 3 and Gemma 3 are trained on web-scale, multi-trillion-token mixtures (Meta AI (2024), Gemma
 453 Team: Aishwarya Kamath et al. (2025b)). Yet their training pipelines diverge sharply. LLaMA 3
 454 relies primarily on supervised instruction tuning and preference-based alignment, whereas Gemma
 455 3 incorporates knowledge distillation during pre-training and follows a structured four-stage post-
 456 training recipe: distillation from a larger instruct teacher, RLHF, RL from machine feedback for math,
 457 and RL from execution feedback for code. This contrast suggests that training design—rather than
 458 size or corpus alone—can substantially influence a model’s ability to infer and match a user’s latent
 459 intent, precisely the behavior our intent measure targets.



454 Figure 3: ICI, MVS and IS across models, shaded area is 95% CI

471 5 DISCUSSION AND LIMITATIONS

472
 473 In this project, we present a formal framework to assess whether LLMs understand the user’s
 474 latent intent. Instead of focusing solely on whether the model responds correctly to a question,
 475 we decompose the response variability of the model into user intent (Intent Sensitivity), irrelevant
 476 information (Articulation Sensitivity), and intrinsic randomness (Model Uncertainty). Applying
 477 this framework across models of varying sizes and across families and domains, we observe that
 478 larger models in our set of models generally exhibit a more robust understanding of users’ intent
 479—but an increased number of parameters does not guarantee better intent comprehension. Difference
 480 across llama and gemma model also indicate that different training approaches and data sources may
 481 contribute to the model sensitivity to the intent and articulation.

482 In some cases, we find that smaller models outperform larger ones, and robustness varies across
 483 different domains, underscoring the need for evaluation across diverse contexts. Our findings
 484 emphasize semantic consistency and generalization as key indicators of model quality, offering a
 485 scalable method to distinguish between true understanding and pattern matching.

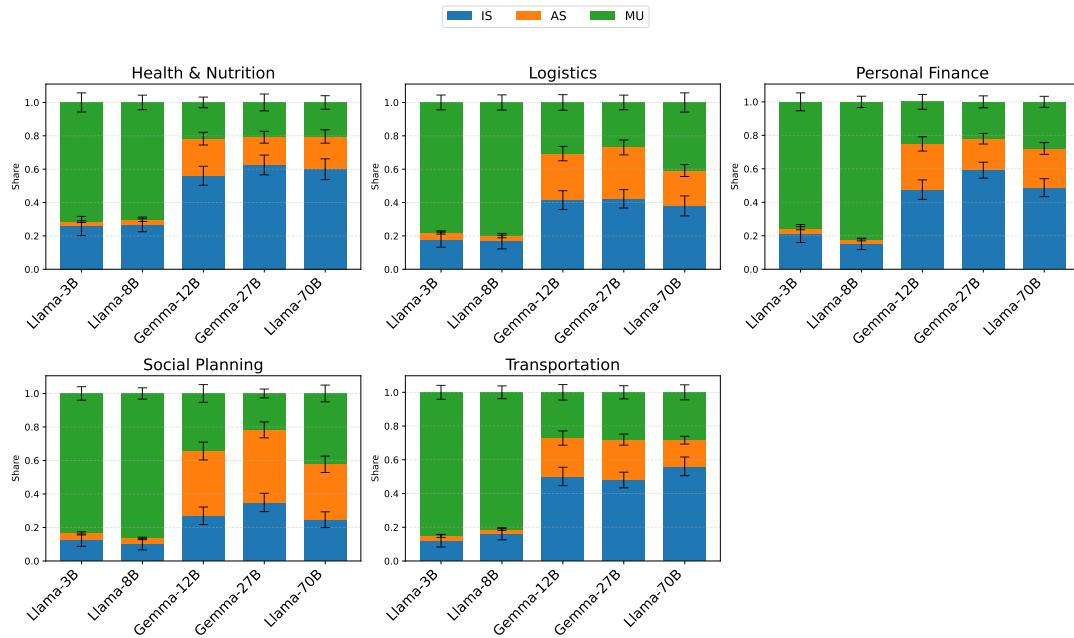


Figure 4: Variance Decomposition across the different topics, with black bars indicating the 95% confidence intervals from using bootstrap with 100 repetitions.

We emphasize that our notion of intent comprehension is not meant to replace standard accuracy metrics, but to complement them. Accuracy evaluates whether a model’s answers are correct, whereas our measure quantifies whether its responses track changes in a user’s underlying intent while remaining stable across paraphrases. In this sense, our framework helps distinguish genuine sensitivity to intent from brittle pattern matching on surface form. See Appendix E for a detailed discussion on incorporating correctness in our framework.

For practitioners, the IS–AS–MU decomposition offers a comprehensive risk assessment framework suitable for informed decision-making in LLM-enabled products. Accuracy tests alone tell a developer or user whether a model can answer a benchmark question; our metric tells them how the model will behave when real users inevitably rephrase, typo, or embed multiple requests in one prompt. Our metric also informs practitioners on potential fairness issues. Articulation sensitivity often correlates with dialect, accent, or education level; a model whose answers swing with phrasing differences can systematically disadvantage certain groups. Tracking AS gives model producers a concrete, quantitative way to demonstrate equity, complementing broader accuracy and bias checks.

Our measurement approach has several limitations. First, we only consider the first and second moments of response distributions. Although some decision-makers may prefer higher-order statistics or full distributional analyses, our method strikes a balance between complexity and practicality: mean and variance already capture key risks for many applications. Second, in settings where evaluators must assign qualitative categories—such as “great,” “ok,” “mid,” “bad,” or “worst”—subjective judgments can lead different annotators to produce different orderings. In such cases, results should be accompanied by sensitivity analyses across scoring rules or annotators to ensure robustness.

Our research highlights the importance of understanding the intent that drives users to utilize LLMs and other AI models. In this study, we show that LLMs often do not fully understand the user intent and respond to differences in superficial phrasing. Future research could more deeply and directly quantify user intent, processes through which intent emerges from preferences or through interaction with generative models, and the inherent limitations of fully expressing complex and original intent. Improved intent identification will enable to better measure models’ ability to capture and respond appropriately.

540 REFERENCES
541

542 Abhinav Jauhri Abhinav Pandey Abhishek Kadian Ahmad Al-Dahle Aiesha Letman Akhil Mathur
543 Alan Schelten Alex Vaughan Amy Yang Angela Fan Anirudh Goyal-Anthony Hartshorn Aobo Yang
544 Archi Mitra Archie Sravankumar Artem Korenev Arthur Hinsvark Arun Rao Aston Zhang Aurelien
545 Rodriguez Austen Gregerson Ava Spataru Baptiste Roziere Bethany Biron-Binh Tang Bobbie
546 Chern Charlotte Caucheteux Chaya Nayak Chloe Bi Chris Marra Chris McConnell Christian
547 Keller Christophe Touret Chunyang Wu Corinne Wong-Cristian Canton Ferrer Cyrus Nikolaidis
548 Damien Allonsius Daniel Song Danielle Pintz Danny Livshits Danny Wyatt David Esiobu Dhruv
549 Choudhary Dhruv Mahajan Diego Garcia-Olano Diego Perino Dieuwke Hupkes Egor Lakomkin
550 Ehab AlBadawy Elina Lobanova Emily Dinan Eric Michael Smith Filip Radenovic Francisco
551 Guzmán Frank Zhang Gabriel Synnaeve Gabrielle Lee Georgia Lewis Anderson Govind Thattai
552 Graeme Nail Gregoire Mialon Guan Pang Guillem Cucurell Hailey Nguyen Hannah Korevaar Hu
553 Xu Hugo Touvron Iliyan Zarov Imanol Arrieta Ibarra Isabel Kloumann Ishan Misra Ivan Evtimov
554 Jack Zhang Jade Copet Jaewon Lee Jan Geffert Jana Vranes Jason Park Jay Mahadeokar Jeet
555 Shah Jelmer van der Linde Jennifer Billock Jenny Hong Jenya Lee Jeremy Fu Jianfeng Chi Jianyu
556 Huang Jiawen Liu Jie Wang Jiecao Yu Joanna Bitton Joe Spisak Jongsoo Park Joseph Rocca Joshua
557 Johnstun Joshua Saxe Junteng Jia Aaron Grattafiori, Abhimanyu Dubey et al. The llama 3 herd of
558 models. *arXiv preprint*, 2024. Used for Meta Llama 3.1 8B.
559

560 Chris L. Baker, Joshua B. Tenenbaum, and Rebecca R. Saxe. Goal inference as inverse planning.
561 In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 779–784,
562 2007.

563 Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning.
564 *Cognition*, 113(3):329–349, 2009.

565 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is
566 to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in
567 neural information processing systems*, 29, 2016.

568 Melanie Brucks and Olivier Toubia. Prompt architecture induces methodological artifacts in large
569 language models. *PLoS one*, 20(4):e0319159, 2025.

570 Barry R Chiswick and Paul W Miller. Linguistic distance: A quantitative measure of the distance
571 between english and other languages. *Journal of multilingual and multicultural development*, 26
572 (1):1–11, 2005.

573 Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale.
574 In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,
575 pages 489–500, 2018.

576 Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural
577 machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational
578 Linguistics (Volume 2: Short Papers)*, pages 567–573, 2017.

579 Shreya Pathak Nino Vieillard Ramona Merhej-Sarah Perrin Tatiana Matejovicova Alexandre Ramé
580 Morgane Rivière Louis Rouillard Thomas Mesnard Geoffrey Cideron Jean-bastien Grill Sabela
581 Ramos Edouard Yvinec Michelle Casbon Etienne Pot Ivo Penchev Gaël Liu Francesco Visin
582 Kathleen Kenealy Lucas Beyer Xiaohai Zhai Anton Tsitsulin Robert Busa-Fekete Alex Feng
583 Noveen Sachdeva Benjamin Coleman Yi Gao Basil Mustafa Iain Barr Emilio Parisotto David Tian
584 Matan Eyal Colin Cherry Jan-Thorsten Peter Danila Sinopalnikov Surya Bhupatiraju Rishabh
585 Agarwal Mehran Kazemi Dan Malkin Ravin Kumar David Vilar Idan Brusilovsky Jiaming Luo
586 Andreas Steiner Abe Friesen Abhanshu Sharma Abheesht Sharma Adi Mayrav Gilady Adrian
587 Goedeckemeyer Alaa Saade Alex Feng Alexander Kolesnikov Alexei Bendebury Alvin Abdagic
588 Amit Vadi András György André Susano Pinto Anil Das Ankur Bapna Antoine Miech Antoine
589 Yang Antonia Paterson Ashish Shenoy Ayan Chakrabarti Bilal Piot Bo Wu Bobak Shahriari
590 Bryce Petrini Charlie Chen Charline Le Lan Christopher A. Choquette-Choo CJ Carey Cormac
591 Brick Daniel Deutsch Danielle Eisenbud Dee Cattle Derek Cheng Dimitris Paparas Divyashree
592 Shivakumar Sreepathihalli Doug Reid Dustin Tran Dustin Zelle Eric Noland Erwin Huijzenge
593 Eugene Kharitonov Frederick Liu Gagik Amirkhanyan Glenn Cameron Hadi Hashemi Hanna
Klimczak-Plucińska Harman Singh Harsh Mehta Harshal Tushar Lehri Hussein Hazimeh Ian

594 Ballantyne Idan Szpektor Ivan Nardini et al. Gemma Team: Aishwarya Kamath, Johan Ferret et al.
 595 Gemma 3 technical report. *arXiv preprint*, 2025a. Used for Gemma 3 12B.
 596

597 Shreya Pathak Nino Vieillard Ramona Merhej-Sarah Perrin Tatiana Matejovicova Alexandre Ramé
 598 Morgane Rivière Louis Rouillard Thomas Mesnard Geoffrey Cideron Jean-bastien Grill Sabela
 599 Ramos Edouard Yvinec Michelle Casbon Etienne Pot Ivo Penchev Gaël Liu Francesco Visin
 600 Kathleen Kenealy Lucas Beyer Xiaohai Zhai Anton Tsitsulin Robert Busa-Fekete Alex Feng
 601 Noveen Sachdeva Benjamin Coleman Yi Gao Basil Mustafa Iain Barr Emilio Parisotto David Tian
 602 Matan Eyal Colin Cherry Jan-Thorsten Peter Danila Sinopalnikov Surya Bhupatiraju Rishabh
 603 Agarwal Mehran Kazemi Dan Malkin Ravin Kumar David Vilar Idan Brusilovsky Jiaming Luo
 604 Andreas Steiner Abe Friesen Abhanshu Sharma Abheesht Sharma Adi Mayrav Gilady Adrian
 605 Goedeckemeyer Alaa Saade Alex Feng Alexander Kolesnikov Alexei Bendebury Alvin Abdagic
 606 Amit Vadi András György André Susano Pinto Anil Das Ankur Bapna Antoine Miech Antoine
 607 Yang Antonia Paterson Ashish Shenoy Ayan Chakrabarti Bilal Piot Bo Wu Bobak Shahriari
 608 Bryce Petrini Charlie Chen Charlaine Le Lan Christopher A. Choquette-Choo CJ Carey Cormac
 609 Brick Daniel Deutsch Danielle Eisenbud Dee Cattle Derek Cheng Dimitris Paparas Divyashree
 610 Shivakumar Sreepathihalli Doug Reid Dustin Tran Dustin Zelle Eric Noland Erwin Huizenga
 611 Eugene Kharitonov Frederick Liu Gagik Amirkhanyan Glenn Cameron Hadi Hashemi Hanna
 612 Klimczak-Plucińska Harman Singh Harsh Mehta Harshal Tushar Lehri Hussein Hazimeh Ian
 613 Ballantyne Idan Szpektor Ivan Nardini et al. Gemma Team: Aishwarya Kamath, Johan Ferret et al.
 614 Gemma 3 technical report. *arXiv preprint*, 2025b. Used for Gemma 3 27B.

615 H. Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957.
 616

617 H. Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1989.
 618

619 Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and
 620 Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint*
 621 *arXiv:2411.10915*, 2024.

622 David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in
 623 neural information processing systems*, 31, 2018a.

624 David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018b.
 625

626 Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning
 627 behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

628 Pierre Jacob. Intentionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.
 629 Stanford University, 2019.

630 Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level
 631 context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*, 2023.

632 Molly Lewis, Aoife Cahill, Nitin Madnani, and James Evans. Local similarity and global variability
 633 characterize the semantic space of human languages. *Proceedings of the National Academy of
 634 Sciences*, 120(51):e2300986120, 2023.

635 Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Watten-
 636 berg. Emergent world representations: Exploring a sequence model trained on a synthetic task.
 637 *Proceedings of the International Conference on Learning Representations*, 2023.

638 Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for
 639 mdps. *AI&M*, 1(2):3, 2006.

640 Samuel Louvan and Bernardo Magnini. Recent neural methods on slot filling and intent classification
 641 for task-oriented dialogue systems: A survey. *arXiv preprint arXiv:2011.00564*, 2020.

642

643 Meta AI. Introducing meta llama 3. <https://ai.meta.com/blog/meta-llama-3/>, 2024.
 644 Accessed: 2025-12-01.

645

646 Libo Qin, Qiguang Chen, Wanxiang Che, Yangming Li, Minheng Ni, Yue Li, Min Liu, Weiwei Deng,
 647 and Ting Liu. A survey on spoken language understanding: Recent advances and new frontiers.
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3032–3045, 2021.

648 Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity
 649 to spurious features in prompt design or: How i learned to start worrying about prompt formatting.
 650 *arXiv preprint arXiv:2310.11324*, 2023.

651

652 Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models
 653 with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for
 654 Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 2016.

655 Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and
 656 Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.

657

658 Adam Smith. *The Theory of Moral Sentiments*. Penguin Books, 1759. Republished 2010.

659

660 Yu Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *NeurIPS*, 2021.

661

662 Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for
 663 language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
 664 volume 36, pages 11385–11393, 2022.

665 Keyon Vafa, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating
 666 the world model implicit in a generative model. *Advances in Neural Information Processing
 667 Systems*, 37, 2024.

668

669 Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data
 670 augmentation for consistency training. In *Advances in Neural Information Processing Systems*,
 671 volume 33, pages 6256–6268, 2020.

672 Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xianchao Zhu, and Yue Zhang. Benchmarking
 673 gpt-4 against human translators: A comprehensive evaluation across languages, domains, and
 674 expertise levels. *arXiv preprint arXiv:2411.13775*, 2024a.

675

676 Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. Gpt-4 vs. human
 677 translators: A comprehensive evaluation of translation quality across languages, domains, and
 678 expertise levels. *arXiv preprint arXiv:2407.03658*, 2024b.

679

680 Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William
 681 Croft, and Tammooy Bhattacharya. On the universal structure of human lexical semantics. *Proceedings
 682 of the National Academy of Sciences*, 113(7):1766–1771, 2016.

683

684 Xiaodong Zhang and Houfeng Wang. A survey of joint intent detection and slot filling models in
 685 natural language understanding. *ACM Computing Surveys*, 55(8):1–38, 2022.

686

687 Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. Prosa:
 688 Assessing and understanding the prompt sensitivity of llms. *arXiv preprint arXiv:2410.12405*,
 689 2024.

690

691

692

693

694

695

696

697

698

699

700

701

702 703 Appendix 704

705 A WORKING WITH NON-NUMERIC RESPONSES 706

707 The main text focuses on settings where model outputs are numerical, for which the variance-based
708 decomposition provides a direct and transparent operationalization of our definition of (sufficient)
709 Intent Comprehension. In many applications, however, the desired outputs are discrete, categorical,
710 or fully free-form text. This section outlines how the same conceptual framework extends to these
711 cases by replacing raw responses with an appropriate representation extracted from the model’s
712 output. We first describe how to handle unrestricted natural-language responses through value-
713 extraction functions such as labelers, scoring models, and embedding-based representations. We then
714 introduce an analogue of our variance decomposition for discrete and categorical settings based on an
715 entropy-based decomposition of variability.

717 A.1 FREE-FORM RESPONSES 718

719 Our theoretical framework in Section 3 does not assume that model outputs are numeric or categori-
720 cal. When models produce unrestricted natural-language responses—explanations, multi-sentence
721 arguments, or other open-ended text—we require a mapping from raw text to an analyzable quantity.
722 This mapping is the value function V , after which the variance decomposition proceeds exactly as in
723 the numeric case.

724 The function $V(a)$ serves to extract the aspect of the response relevant to evaluation. Depending on
725 the application, V may be implemented through human annotation, a domain-specific scoring rule,
726 an evaluator model, or an embedding-based representation. If the analyst wishes to track a specific
727 semantic property (such as stance, polarity, or task success), $V(a)$ can return a scalar or discrete
728 label derived from the text. For more open-ended generation, $V(a)$ may be a continuous embedding
729 vector capturing semantic or stylistic content. In such cases one may analyze individual embedding
730 coordinates, low-dimensional projections, or task-specific directions, each of which functions as the
731 response variable whose variation is decomposed across intent and articulation.

732 The only requirement is that the representation induced by V meaningfully captures the dimension
733 along which intent comprehension is to be assessed. When this dimension is well defined, V may
734 be a deterministic extractor; when it is diffuse or multidimensional, embeddings offer a natural
735 surrogate. In all cases, once $V(a)$ is specified, the decomposition attributes variation in this induced
736 representation to changes in intent, changes in articulation, or residual randomness in exactly the
737 same manner as with scalar outputs.

738 A.2 DECOMPOSITION FOR DISCRETE OUTPUTS 739

740 In the main text, we use the variance decomposition for our measure
741

$$742 \text{Var}(v) = \underbrace{\text{Var}(\mathbb{E}[v | I])}_{IS} + \underbrace{\mathbb{E}_I[\text{Var}(\mathbb{E}[v | p_I] | I)]}_{AS} + \underbrace{\mathbb{E}_I[\mathbb{E}[\text{Var}(v | p_I) | I]]}_{MU},$$

745 which is appropriate when v is continuous (or ordinal), as $\text{Var}(\cdot)$ is then well-defined. For categorical
746 or otherwise non-ordinal outcomes, however, variance cannot be uniquely specified without imposing
747 an arbitrary numerical scale. For example, consider a travel agency that uses a chatbot to recommend
748 destinations based on user preferences. From the agency’s perspective, the recommendations corre-
749 spond to discrete destinations. To handle such cases, we propose an information-theoretic analogue,
750 derived directly by applying the chain rule of entropy twice.

751 For a discrete random variable X with distribution $P(X)$ we write $H(X) = -\sum_x P(X =$
752 $x) \log P(X = x)$ for its Shannon entropy. We denote the mutual information between random
753 variables X and Y as $I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$, where $H(Y | X)$
754 is the conditional entropy of Y given X , $H(X | Y = y) = -\sum_x P_{X|Y}(x | y) \log P_{X|Y}(x | y)$. Let
755 v be a discrete outcome variable, and denote again I as intent and p_I as the prompt constructed from
intent, as defined in the main text.

756 With these definitions, we obtain the decomposition
 757

$$758 H(v) = \underbrace{I(I; v)}_{IS} + \underbrace{E_i[I(p_I; v | I = i)]}_{AS} + \underbrace{E_i[H(v | p_I, I = i)]}_{MU}.$$

$$759$$

$$760$$

761 where IS (Intent Sensitivity) is the mutual information between the intent I and the model output
 762 quantifies how much output reveals about the intended meaning before the prompt is supplied.
 763 AS (Articulation Sensitivity) is now the conditional mutual information that measures the extra
 764 information carried by the prompt p_I , given the intent. Finally, MU (Model uncertainty) is the
 765 remaining conditional entropy that captures irreducible uncertainty in the model’s prediction once
 766 both intent and prompt are known.

767 We can further divide by the total entropy to get a unit-sum normalization that facilitates comparison
 768 across models:

$$770 1 = \frac{I(I; v)}{H(v)} + E_i \left[\frac{I(p_I; v | I = i)}{H(v)} \right] + E_i \left[\frac{H(v | p_I, I = i)}{H(v)} \right].$$

$$771$$

$$772$$

773 This expression mirrors the continuous-outcome variance decomposition in structure while remaining
 774 well-defined for any discrete output space.

775 A.2.1 EXAMPLE OF USAGE FOR THE DISCRETE CASE

$$776$$

777 We illustrate the discrete decomposition by applying it in a manner analogous to the variance
 778 decomposition approach we used in the main text. As an example, we focus on decision problems in
 779 consumer credit.

780 As in the main analysis, the experimental pipeline consists of synthetic intent generation, multilingual
 781 paraphrasing of task prompts, large-scale sampling of model outputs across all intent–paraphrase
 782 combinations, and an information-theoretic entropy decomposition applied separately for each model
 783 and task.

784 We evaluate two consumer-credit decision problems. The first task, loan approval, is a binary decision
 785 problem in which the model outputs $v \in \{0, 1\}$, where 1 indicates approval and 0 rejection. The
 786 second task, risk rating, is an ordinal classification problem in which the model assigns a risk score
 787 $v \in \{1, 2, 3, 4, 5\}$, with 1 denoting very low risk and 5 very high risk. For both tasks, the prompt
 788 specifies a detailed natural-language description of the bank’s decision policy, and the output label
 789 is parsed from the model’s response. Any output whose extracted label falls outside the valid set is
 790 discarded and the prompt is resampled.

791 **INTENT GENERATION** For each task, we generate a small synthetic set of *intents*—applicant profiles
 792 that serve as proxies for underlying world states. Intents are generated automatically using ChatGPT
 793 4.1. The prompt instructs the model to produce twelve diverse and realistic loan applicant profiles,
 794 each containing specified attributes (age, income, existing debt, employment characteristics, credit
 795 history, loan amount, and loan term). The model returns a JSON list, from which we extract twelve
 796 applicant profiles per task. These profiles are shared across all evaluation models and all downstream
 797 prompting steps.

798 **PARAPHRASE GENERATION** Paraphrased versions of each base prompt are constructed via back-
 799 translation, following the method described in the main text. Each task has a base English prompt
 800 template that describes the decision-making role, the policy criteria, the format of the applicant
 801 profile, and the required output label. A placeholder token `applicant_profile` is replaced with
 802 an intent-specific description. Translations and back-translations—performed again using ChatGPT
 803 4.1—yield candidate paraphrased templates. To ensure paraphrase fidelity, we apply the automatic
 804 equivalence check used in the main experiment.

805 **SAMPLING DESIGN** Given the sets of intents and paraphrased templates, we form a full grid of
 806 evaluation conditions over tasks, intents, paraphrases, and models. Each task combines twelve intents
 807 with ten paraphrases, yielding $12 \times 10 = 120$ prompt templates per task. For each template and each
 808 model, we repeatedly sample outputs by instantiating the template with the corresponding applicant

profile and submitting the resulting prompt. We evaluate the same five instruction-tuned models used in the main experiment. For each *task, intent, paraphrase, model* tuple, we target 25 valid samples, where validity requires that the parsed output lies within the task’s label set. Across both tasks, intents, and paraphrases, this yields 6,000 samples per model and 30,000 samples in total.

EMPIRICAL DECOMPOSITION To measure intent comprehension in the discrete case, we apply empirical plug-in estimators for the decomposition in Appendix ???. The observed data consist of $(v_n, i_n, p_n)_{n=1}^N$, where v_n is the discrete model output, i_n the intent index, and p_n the paraphrase index. For any (v, i, p) , define the empirical joint distribution $\hat{p}(v, i, p) = \frac{1}{N} \sum_{n=1}^N \mathbf{1}_{v_n = v, i_n = i, p_n = p}$, with the corresponding marginals obtained by summation and conditional distributions formed whenever denominators are nonzero. Using these empirical distributions, we compute plug-in estimators for the relevant entropies:

$$\hat{H}(v) = \sum_v \hat{p}(v) \log \hat{p}(v),$$

$$\hat{H}(v | I) = \sum_i \hat{p}(i) \left(* \sum_v \hat{p}(v | i) \log \hat{p}(v | i) \right),$$

$$\hat{H}(v | p_I, I) = \sum_{i,p} \hat{p}(i, p) \left(* \sum_v \hat{p}(v | i, p) \log \hat{p}(v | i, p) \right).$$

The empirical mutual informations are

$$\hat{I}(I; v) = \hat{H}(v) - \hat{H}(v | I), \quad \hat{I}(p_I; v | I) = \hat{H}(v | I) - \hat{H}(v | p_I, I).$$

Finally, we compute the empirical shares

$$\hat{IS} = \frac{\hat{I}(I; v)}{\hat{H}(v)}, \quad \hat{AS} = \frac{\hat{I}(p_I; v | I)}{\hat{H}(v)}, \quad \hat{MU} = \frac{\hat{H}(v | p_I, I)}{\hat{H}(v)},$$

with all shares set to zero whenever $\hat{H}(v) = 0$. We construct 95% confidence intervals using the percentile bootstrap with $B = 100$ replications.

Results Figures 5 and 5 report the information-theoretic decomposition for *loan approval* and *risk rating*, respectively. Across both tasks, we observe a clear separation between model families and scales. The smaller LLaMA models allocate a large share of total label entropy to model uncertainty and relatively little to intent sensitivity, indicating that the realized decision label remains hard to predict even after conditioning on the applicant profile (intent) and the specific paraphrase. In other words, much of their variation looks like “intra-condition wobble” rather than systematic differentiation across applicants. By contrast, Gemma-12B and Gemma-27B exhibit substantially higher IS, implying that most of the explainable variation in discrete decisions is driven by changes in applicant profiles rather than by superficial prompt differences; these models behave more like stable decision rules that react to the underlying state.

A notable nuance is that LLaMA-70B tends to sit between these extremes: it shows strong IS (it reacts to intent), but also a visibly non-trivial articulation sensitivity component. In this discrete setting, AS has a concrete interpretation: conditional on the same applicant profile, some paraphrases systematically tilt the predicted label distribution. For a high-stakes decision framing, this is exactly the failure mode our framework is designed to surface—outcomes can become partially a function of how the request is phrased rather than what the applicant profile implies.

Comparing tasks also hints why articulation effects can matter more in ordinal prediction than in binary decisions. The risk rating task introduces “adjacent-category ambiguity”: even when a model has the right qualitative assessment, mapping that assessment into one of five discrete bins (e.g., 3 vs. 4) can be more susceptible to framing. Thus, increases in AS for risk rating are especially informative: they suggest not just noise, but potentially a prompt-dependent calibration of the decision threshold on the ordinal scale.

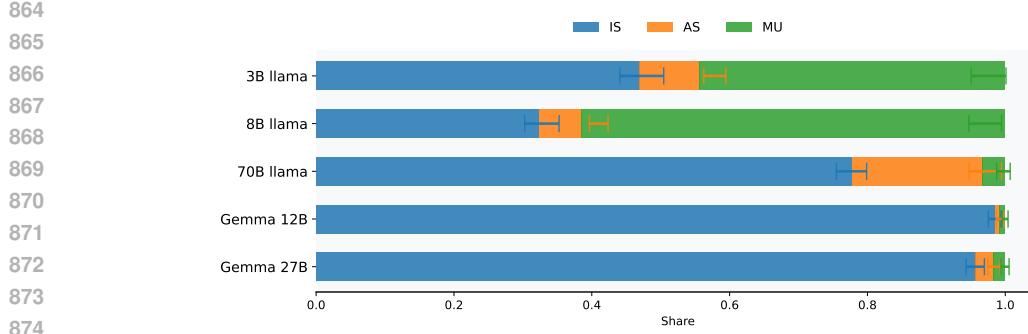


Figure 5: Loan Approval - Information Theoretic Decomposition of Intent Comprehension

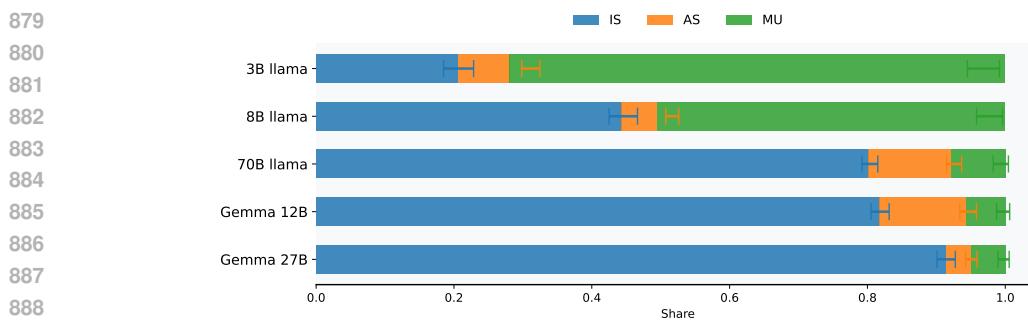


Figure 6: Risk Rating - Information Theoretic Decomposition of Intent Comprehension

Across both loan tasks, the decomposition pinpoints how intent comprehension fails when mapping an applicant profile to an outcome. High MU means the model has not formed a stable applicant→decision rule: even holding the candidate attributes and the exact articulation fixed, it vacillates, consistent with an internally ambiguous or weakly specified credit assessment. High AS, in contrast, directly reflects imperfect intent comprehension: when the candidate is held fixed, paraphrases that should be equivalent nonetheless shift the output, implying the model is not reliably reacting to the applicant profile but is letting unintended, surface-level factors (framing, emphasis, narrative style, wording) enter the decision rule. In other words, the model behaves less like it extracted “what is this applicant’s risk?” and more like it is pattern-matching on presentation artifacts.

Prompt Templates for the discrete case The following list the exact prompt templates used in the experiment. Placeholders enclosed in braces, such as `{applicant_profile}`, are instantiated programmatically at runtime.

BASE TASK PROMPTS For the `loan_approval` task, the base prompt template used when querying evaluation models is:

You are a conservative loan underwriter at a retail bank.
The bank’s policy:

Approve loans only if the borrower has a high likelihood of repaying on time.

Reject loans with substantial risk of default.

Use only the information below. Ignore any information not provided.

Applicant profile:

`{applicant_profile}`

Decision:

Should the bank approve this loan?

```

918
919 Answer with a single digit:
920 1 = approve
921 0 = reject
922 Do not output anything else.
923
924 For the risk_rating task, the base prompt template is:
925
926 You are a risk analyst at a bank.
927 Rate the credit risk of the applicant for the specified loan.
928
929 Applicant profile:
930 {applicant_profile}
931
932 Rating:
933 On a scale from 1 to 5, where
934 1 = very low risk,
935 2 = low risk,
936 3 = moderate risk,
937 4 = high risk,
938 5 = very high risk,
939
940 answer with a single integer between 1 and 5, and nothing else.
941
942 INTENT-GENERATION PROMPTS Intents are generated once per task using an Azure-hosted model
943 with the following system message:
944
945 You generate realistic bank loan applicant profiles for credit assessment.
946 Each profile should include:
947
948 Age
949 Annual income
950 Existing debt
951 Employment type and tenure
952 Credit history (late payments, defaults, etc.)
953 Loan amount and term
954
955 Keep each profile in 5-8 bullet points, simple and plausible.
956
957 and the following user message:
958
959 Generate 12 diverse applicant profiles that vary widely in income, debt-to-income ratio, and risk profile.
960
961 The JSON returned by this interaction is parsed, and the value of the "applicant_profile" field in each object is used as an intent.
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999

```

918

919 Answer with a single digit:

920 1 = approve

921 0 = reject

922 Do not output anything else.

923

924 For the risk_rating task, the base prompt template is:

925

926 You are a risk analyst at a bank.

927 Rate the credit risk of the applicant for the specified loan.

928

929 Applicant profile:

930 {applicant_profile}

931

932 Rating:

933 On a scale from 1 to 5, where

934 1 = very low risk,

935 2 = low risk,

936 3 = moderate risk,

937 4 = high risk,

938 5 = very high risk,

939

940 answer with a single integer between 1 and 5, and nothing else.

941

942 INTENT-GENERATION PROMPTS Intents are generated once per task using an Azure-hosted model

943 with the following system message:

944

945 You generate realistic bank loan applicant profiles for credit assessment.

946 Each profile should include:

947

948 Age

949 Annual income

950 Existing debt

951 Employment type and tenure

952 Credit history (late payments, defaults, etc.)

953 Loan amount and term

954

955 Keep each profile in 5-8 bullet points, simple and plausible.

956

957 and the following user message:

958

959 Generate 12 diverse applicant profiles that vary widely in income, debt-to-income ratio, and risk profile.

960

961 The JSON returned by this interaction is parsed, and the value of the "applicant_profile" field in each object is used as an intent.

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

Overall, the human evaluations provide strong evidence that both components of the pipeline behave reliably. In the intent-preservation task, both annotators judged the vast majority of paraphrases as faithfully preserving the original intent, with Annotator 1 marking 95% of items as intent-preserving and Annotator 2 marking 91% as such, where in 98.5% of cases, at least one of the annotator indicate the intent was preserved. The annotators agreed on 89% of cases, reflecting consistent judgments, and disagreements were concentrated in borderline paraphrases involving minor stylistic shifts or slight differences in emphasis. Because negative labels were extremely rare, Cohen’s κ is mechanically low ($\kappa \approx 0.16$) despite the high raw agreement; this effect is well-known in settings with extreme class imbalance and does not indicate substantive disagreement between annotators. Table ?? is the confusing matrix across the two annotators.

		Annotator 2	Different	Same
		Annotator 1		
		Different	3	7
		Same	15	175

Table 1: Confusion Matrix Between the Two Annotators

The evaluation of the numerical-value extractor shows similarly encouraging results. Annotator 1 labeled 94% of extractions as correct, and Annotator 2 labeled 91% as correct, and the two annotators agreed on 93% of items. Disagreements were limited and typically arose when the model provided a range (e.g., “around 5–10”) or its reasoning was cut by the token limit, leading annotators to differ slightly in whether the extractor should be credited for selecting a particular value. Cohen’s κ is approximately 0.50, interpreted as moderate agreement and consistent with the very high raw agreement. The table below shows the confusion matrix for both annotators.

		Annotator 2	Incorrect	Correct
		Annotator 1		
		Incorrect	8	4
		Correct	10	178

Table 2: Confusion Matrix Between the Two Annotators

Taken together, these results indicate that both the paraphrase-generation and numerical-extraction components of the pipeline perform reliably in practice. The high agreement rates across two independent annotators, supported by chance-corrected measures where appropriate, demonstrate that paraphrases almost always preserve intent and that numerical values are extracted correctly in the vast majority of model outputs.

C RELATED LITERATURE

LLMs Prompt Robustness Recent papers have investigated the sensitivity of large language models (LLMs) to prompting using a variety of frameworks and metrics. Zhuo et al. (2024) introduce the ProSA framework, which quantifies prompt sensitivity through a novel metric called PromptSensiScore (PSS). PSS measures the average variation in performance across semantically equivalent prompt variants at the instance level, offering insight into how much a model’s output changes with different prompt formulations. Their findings show that prompt sensitivity varies across tasks, models, and prompt types, with particularly high sensitivity observed in reasoning and creative tasks. Similarly, Sclar et al. (2023) examine the role of “spurious features” in prompt formatting—such as punctuation, spacing, and capitalization—and propose FORMATSPREAD, a metric that captures the spread in task accuracy across equivalent prompt formats. FORMATSPREAD is defined as the difference in performance between the best and worst format, and their results reveal that such superficial changes can lead to performance swings of up to 76 accuracy points—effects that are not mitigated by model size or few-shot learning. Brucks and Toubia (2025) take a different perspective, framing prompt sensitivity as a methodological artifact inspired by the concept of choice architecture.

1026 Using full-factorial experiments, they show that prompt order, labeling, framing, and justification
 1027 systematically bias model responses—and that even instructing the model to ignore these features
 1028 does not eliminate the effect. Their central claim is that no prompt is neutral. To address this, they
 1029 recommend aggregating responses across multiple prompts, akin to ensemble methods, to counteract
 1030 individual prompt biases.

1031 Our behavioral framing complements and extends this literature by offering a diagnostic framework
 1032 that quantifies prompt sensitivity in terms of structured variance. Unlike prior work that focuses on
 1033 scalar measures of how model responses change, we take seriously the idea that model outputs should
 1034 be stochastic—but emphasize that the source of this randomness should not arise from arbitrary
 1035 variations in prompt phrasing. As we know, no other measure of prompt sensitivity linked the
 1036 importance of robustness to the world model and focused on the variance aspect of responses.

1037
 1038 **Measuring Semantic Robustness and Fairness.** Our decomposition also offers practical diagnostic
 1039 value. High *articulation sensitivity* suggests that small variations in phrasing disproportionately affect
 1040 model outputs, raising concerns for *fairness* and *accessibility*. Prior work shows that models may
 1041 perform differently for users with dialectal, non-standard, or less “typical” phrasing (Bolukbasi et al.,
 1042 2016; Si et al., 2022; Tan and Celis, 2021; Guo et al., 2024). Our framework allows developers to
 1043 quantify this fragility and track improvements over time. In this respect, our method serves as a
 1044 form of semantic reliability auditing, applicable in safety-critical domains such as healthcare, finance,
 1045 and legal reasoning, where output variability under minor phrasing shifts can lead to unacceptable
 1046 inconsistency.

1047
 1048 **Translation Chains** Back-translation is a data augmentation technique in natural language processing,
 1049 particularly in neural machine translation (Sennrich et al. (2016)). Back-translation involves
 1050 translating monolingual target-language text into the source language using a reverse translation
 1051 model, and then using the resulting synthetic parallel data to train the forward model. Subsequent
 1052 work has extended the approach to include round-trip translation and multilingual back-translation,
 1053 where intermediate languages are introduced to further diversify the training data and improve gener-
 1054 alization (Fadaee et al., 2017; Edunov et al., 2018; Xie et al., 2020; Youn et al., 2016). These variants
 1055 exploit linguistic variation introduced during translation to create richer training distributions, which
 1056 have been beneficial not only for translation tasks but also for classification, question answering, and
 1057 style transfer. In our setup, we do not use translation chains as a data augmentation tool, but instead
 1058 think of them as a way to diversify language while preserving meaning.

1059 C.1 RELATION TO OTHER WORK ON MEASURING WORLD MODELS

1060 As discussed in Remark 1, our definition of Intent Comprehension, is highly related to world
 1061 models. Our notion of a world model draws inspiration from research in reinforcement learning
 1062 (RL) and model-based control, where a world model captures the dynamics of an environment and
 1063 supports planning. In these contexts, a world model is typically a learned transition function or latent
 1064 representation that abstracts the environment’s relevant state space [Ha and Schmidhuber (2018b);
 1065 Hafner et al. (2019); Ha and Schmidhuber (2018a)].

1066 Analogously, we treat user intent as a latent variable and evaluate whether a model can infer and
 1067 represent this variable consistently across diverse inputs. This is conceptually related to state
 1068 abstraction in RL Li et al. (2006), where different observations map to the same underlying state if
 1069 they yield equivalent value functions. In our case, different prompts map to the same intent if they
 1070 elicit the same output distribution (or value-weighted distribution).

1071 Recent work has examined whether generative models develop an internal world model in the context
 1072 of games. Toshniwal et al. (2022) and Li et al. (2023) helped establish games—such as chess
 1073 and Othello—as testbeds for evaluating the emergence of world models. A common approach in
 1074 this literature involves using probes to assess whether a model’s internal representations encode
 1075 latent game states. In contrast, our evaluation metrics are model-agnostic: rather than probing
 1076 representations or relying on an external notion of state feasibility, we focus on the internal consistency
 1077 of a model’s responses as an indicator of world model quality.

1078 Most closely aligned with our work is Vafa et al. (2024), who test an LLM’s world model by asking
 1079 whether it recovers the deterministic finite automaton (DFA) governing a sequence-generation

task. Leveraging the Myhill–Nerode theorem, they introduce two metrics: (i) sequence compression—do any two prefixes that land in the same DFA state admit the same set of valid continuations? and (ii) sequence distinction—do prefixes that reach different states generate different permissible continuations?

Both their framework and ours impose a common behavioral mandate: (1) inputs sharing a latent condition (DFA state or user intent) must elicit indistinguishable outputs, and (2) inputs differing in that condition must yield measurably different outputs. Crucially, both approaches assess world-model quality based solely on outputs, without inspecting internal parameters.

The divide lies in how “correctness” is anchored. Vafa et al. (2024) use a built-in, binary oracle—the known DFA—to label continuations as right or wrong at the syntax level. Our variance-decomposition framework instead delegates that judgment to an external evaluator V , which maps a $\langle \text{prompt}, \text{response} \rangle$ pair to latent intent and task value. V can imitate a hard 0/1 oracle, but it can also apply graded semantic scores. Importantly, our test goes further: when V deems two prompts equivalent, we require the entire distribution of (possibly stochastic) responses to match across those prompts. If V itself enforces strict correctness, this collapses to zero model uncertainty—any variability in outputs must stem from differing intents. Thus, while both methods rely on an evaluator, Keyon’s oracle is intrinsic and binary, whereas ours is external, tunable, and distribution-aware.

Finally, our notion of a IC demands distributional equality across equivalent prompts, whereas the DFA test assumes deterministic continuations. Extending Myhill–Nerode to stochastic automata would miss the point we target: open-ended tasks with no unique correct answer. In such settings, the likelihoods assigned to alternative continuations encode the model’s assumptions. A genuine world model, therefore, aligns those likelihoods whenever the underlying intent is the same, preserving semantic sufficiency even under uncertainty.

D ESTIMATING VARIANCE DECOMPOSITION VIA ANOVA

Our goal is to decompose variability in numeric responses into three components: IS, AS, and MU. A natural starting point is the variance decomposition

$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y | i, p]) + \mathbb{E}[\text{Var}(Y | i, p)],$$

where (i, p) indexes task and prompt. A naive sample-analogue implementation replaces the conditional means $\mathbb{E}[Y | i, p]$ with cell-wise sample means and the conditional variances with cell-wise sample variances, and then computes the variance of these estimated means across tasks and prompts. However, this plug-in approach can induce finite-sample bias when cells are sparse or unbalanced.

To see the problem, let $\mu_{ip} = \mathbb{E}[Y | i, p]$ and let $\hat{\mu}_{ip}$ denote the sample mean in cell (i, p) based on n_{ip} draws. Then

$$\text{Var}(\hat{\mu}_{ip}) = \text{Var}(\mu_{ip}) + \mathbb{E}[\text{Var}(\hat{\mu}_{ip} | \mu_{ip})] \approx \text{Var}(\mu_{ip}) + \frac{\sigma_R^2}{n_{ip}},$$

so the variance of *estimated* cell means mechanically includes sampling noise on the order of σ_R^2/n_{ip} . When many cells have small n_{ip} , this additional term inflates the apparent between-cell variation and therefore biases upward the estimated contributions of IS and AS. Unbalanced designs exacerbate the issue: cells with very small n_{ip} have highly variable $\hat{\mu}_{ip}$ but still enter equally in a simple variance-of-means calculation. In contrast, the within-cell component can be biased downward if the between-cell inflation is not correctly accounted for. In short, naive variance decomposition treats noisy cell means as if they were observed without error.

As the number of samples per cell grows, the sampling variance of $\hat{\mu}_{ip}$ shrinks and the naive decomposition approaches the population decomposition. In the limit $n_{ip} \rightarrow \infty$ for all (i, p) , the plug-in estimator is consistent. Practically, however, we work with finite and often heterogeneous n_{ip} . Precision improves with (i) the number of draws per task–prompt cell, and (ii) the diversity of paraphrases per task, which helps distinguish task-level and prompt-level variability rather than confounding them with noise. As a rough rule of thumb, naive cell-wise decompositions are reasonably stable only when most (i, p) cells have at least 20–30 responses and very few cells have fewer than about 5; below this range, between-cell variance estimates become highly sensitive to sampling noise, and partial pooling is strongly preferable.

To address these issues directly, we use a hierarchical (mixed-effects) specification that *borrows strength* across related cells via *partial pooling*: group means are shrunk toward a common mean in proportion to their sampling variance and group size. This stabilizes variance components, especially when some tasks or prompt variations have few observations.

Index tasks by $i = 1, \dots, I$, prompt variations (nested within tasks) by $p = 1, \dots, P_i$, and individual observations within a prompt by k . We model the standardized response y_{ipk} as

$$y_{ipk} = \mu + u_i + v_{ip} + \varepsilon_{ipk}, \quad u_i \sim \mathcal{N}(0, \sigma_I^2), \quad v_{ip} \sim \mathcal{N}(0, \sigma_A^2), \quad \varepsilon_{ipk} \sim \mathcal{N}(0, \sigma_R^2), \quad (1)$$

with v_{ip} nested in task i . Here μ is a global intercept, u_i captures intent-level variation, IS, v_{ip} captures prompt-level heterogeneity within intent, AS, and ε_{ipk} is idiosyncratic noise, MU). Partial pooling arises because the random effects (u_i, v_{ip}) are estimated jointly: noisy cells with small n_{ip} are shrunk more aggressively toward the global mean, preventing them from exerting disproportionate influence on the between-cell variance.

Let $\sigma_T^2 = \sigma_I^2 + \sigma_A^2 + \sigma_R^2$ denote the total variance in model responses. We report variance *shares*

$$\text{IS} = \frac{\sigma_I^2}{\sigma_T^2}, \quad \text{AS} = \frac{\sigma_A^2}{\sigma_T^2}, \quad \text{MU} = \frac{\sigma_R^2}{\sigma_T^2},$$

which sum to 1 and are invariant to scale. If we standardize y before fitting, these coincide with absolute contributions on the standardized scale. We estimate $(\mu, \sigma_I^2, \sigma_A^2, \sigma_R^2)$ by *Restricted Maximum Likelihood* (REML), which reduces the small-sample bias of variance component MLEs. Practically, we fit (1) via `lme4::lmer` (through `pymer4`). Finally, we report uncertainty for the *shares* using a bootstrap that respects the nesting: we resample at the task level, then within tasks at the prompt level, and then within prompts at the response level.

Throughout our estimation we report bootstrap standard errors for all figures, from 100 repetitions.

E INCORPORATING CORRECTNESS INTO THE FRAMEWORK

Our main experiments focus on numeric responses, where the valuation function V maps a model output a (possibly together with the prompt p) to a scalar value $Y = V(a, p)$ (e.g., a probability, price, or duration). When ground-truth labels are available, the same framework can enhance the standard analysis of *correctness*. This appendix explains how and also clarifies what is lost when we collapse multiple plausible answers into a single correctness label.

E.1 BINARY CORRECTNESS

Suppose we have a notion of correctness at the level of the evaluator. For a given prompt p and response a , let

$$V_{\text{corr}}(a, p) \in \{0, 1\}$$

indicate whether the response is correct (1) or incorrect (0) according to ground truth or human judgment. We can then apply the variance decomposition (or the decomposition suggested in appendix ?? for discrete values) to the random variable

$$Y_{\text{corr}} = V_{\text{corr}}(A, P),$$

where A is the model’s sampled answer and P ranges over prompts in our design.

All definitions carry over directly:

- IS_{corr} measures how much of the variation in *correctness* is explained by changes in intent, holding articulation fixed.
- AS_{corr} measures how much variation in *correctness* is explained by changes in articulation, holding intent fixed.
- MU_{corr} captures residual variation in *correctness* not explained by either.

In degenerate cases where a model is always correct (or always wrong) on a given task, Y_{corr} is constant and there is no variation to explain. This is appropriate: when correctness is deterministic,

1188 there is no remaining uncertainty about where errors come from and we can't distinguish between cases
 1189 where the model is simply matching patterns to cases where the model is reacting to the underlying user
 1190 intent. The decomposition is most informative in the realistic regime where correctness lies strictly
 1191 between 0 and 1, and we wish to understand whether failures are driven by intent, articulation, or
 1192 residual noise.

1193 In this sense, using correctness as the valuation V does not replace our original analysis on raw
 1194 numerical answers, but complements it. When we apply the decomposition to raw answers, we study
 1195 *how the model's responses change with intent and articulation*. When we apply it to correctness
 1196 indicators, we instead study *how the probability of being correct changes with intent and articulation*,
 1197 that is, whether errors are concentrated on certain intents, certain phrasings, or are largely unexplained
 1198 noise.

1199

1200

E.2 MULTIPLE ACCEPTABLE ANSWERS AND PARTIAL CREDIT

1201

1202 In many settings there is not a single uniquely correct answer, but instead a set of *acceptable* answers
 1203 $\mathcal{A}_{\text{acc}}(p)$ for a prompt p (e.g., a range of reasonable numerical estimates, or several semantically
 1204 equivalent textual completions). Our framework can accommodate this in several ways.

1205 A simple approach is to define a binary acceptability indicator

1206

$$V_{\text{acc}}(a, p) = \mathbf{1}\{a \in \mathcal{A}_{\text{acc}}(p)\},$$

1208

and apply the same decomposition to

1209

$$Y_{\text{acc}} = V_{\text{acc}}(A, P).$$

1211

1212 The resulting IS_{acc} , AS_{acc} , and MU_{acc} then describe how the *probability of producing an acceptable*
 1213 *answer* varies with intent and articulation.

1214

1215 When numeric magnitude also matters, we can combine acceptability and value. For example, in a
 1216 guesstimation task we might consider

1217

$$V_{\text{hyb}}(a, p) = V(a) \cdot \mathbf{1}\{a \in \mathcal{A}_{\text{acc}}(p)\},$$

1218

1219 where $V(a)$ is the relevant numeric value in the response, so that only acceptable answers contribute
 1220 their numeric value, while unacceptable answers are mapped to zero. More generally, one can use a
 1221 *graded scoring function*

1222

$$V_{\text{score}}(a, p) \in [0, 1],$$

1223

1224 which assigns partial credit based on distance to the correct value or semantic similarity to a reference
 1225 answer. In all these cases, our variance decomposition applies unchanged to the induced random
 1226 variable $Y = V(A, P)$.

1227

1228 In cases where there are multiple plausible answers, simply encoding correctness as a binary indicator
 1229 may miss important structure. Collapsing an entire set $\mathcal{A}_{\text{acc}}(p)$ into the value 1 and everything else
 1230 into 0 discards information about *how* the model's responses are distributed within the acceptable set
 1231 and how they deviate when they are not acceptable. For example, two models might have the same
 1232 overall correctness rate, but one produces a tight distribution around a canonical answer while the
 1233 other oscillates between several qualitatively different, yet still acceptable, solutions. A binary V_{acc}
 1234 does not distinguish these behaviors.

1235

1236 Our framework is flexible to this choice. If the evaluation goal is to separate *successes from failures*,
 1237 correctness-based encodings such as V_{corr} or V_{acc} are appropriate, and the decomposition reveals
 1238 whether variability in success is driven by intent, articulation, or model uncertainty. If, instead, the
 1239 goal is to understand how the model navigates a space of multiple plausible answers, one can retain a
 1240 richer V (e.g., a continuous score or a calibrated utility) and apply the decomposition to that quantity
 1241 rather than to a binary indicator.

1242

1243 Thus, accuracy-based evaluation and our intent-comprehension metrics are not competing notions.
 1244 Accuracy tells us *how often* the model is right; our framework, applied either to raw answers or to
 1245 correctness-encoded values, helps explain *why* it succeeds or fails, by attributing variation to intent,
 1246 articulation, and residual uncertainty, while leaving room to capture the structure of multiple plausible
 1247 answers when that structure is evaluatively relevant.

1242 **F EXPERIMENTAL DETAILS**

1244 In this section, we describe how we construct our experiments. We first describe how we construct
 1245 our set of tasks and intents, and then discuss how we construct an intent-equivariant set of prompts.
 1246

1247 **Constructing the set of tasks** To construct our evaluation metric, we focus on generating open-ended,
 1248 guesstimation-type questions, using a two-stage LLM workflow. We focus on these types of questions
 1249 as our IC metric for two reasons. First, our measure is defined over response distributions. To identify
 1250 whether variation is driven by intent (IS) rather than articulation (AS) or model uncertainty (MU),
 1251 we need tasks that produce genuine dispersion in plausible answers³. Open-ended questions, with
 1252 various sets of answers, also match real usage, where users ask for estimates, recommendations, and
 1253 contextual judgments; Together, this yields a naturalistic yet controlled setting in which IS, AS, and
 1254 MU are identifiable and informative.

1255 We focus on 5 areas of interest: Transportation, Personal Finance, health and nutrition, logistics,
 1256 and social planning. and use an automatic procedure to generate 24 questions for each topic. To
 1257 construct our dataset of guesstimation-style prompts, we implemented an automated pipeline that
 1258 leverages large language models guided by structured templates and semantic constraints. Each
 1259 generated question contains exactly one {placeholder} token, which is later substituted with concrete
 1260 values. The system prompt enforces that questions are self-contained, specify explicit reporting units
 1261 (e.g., “dollars per year,” “kWh per month”), and use the placeholder to denote a general semantic
 1262 role such as a material, process, or population subgroup. To ensure coverage across domains, we
 1263 draw from a library of question templates spanning categories such as counts, rates, intensities,
 1264 costs, and probabilities (see Appendix I.1). For each placeholder, a second prompt generates a set of
 1265 plausible replacements, or intents, which are short noun phrases consistent with the semantic role
 1266 and unit of the question. This two-step process—first generating unit-specified question structures
 1267 and then populating them with realistic values—produces a diverse set of open-ended, quantitative
 1268 guesstimation questions (see Appendix J for the exact wording of the prompts).

1269 **Creating a Set of Intent-Equivalent Prompts** Given the set of categories and tasks, our next
 1270 challenge is to generate a collection of prompts that express the *same intent* but differ in surface form
 1271 and semantics. This is a non-trivial challenge. Even for humans, writing multiple prompts that convey
 1272 exactly the same intent without introducing subtle shifts in meaning is difficult. Language models
 1273 are highly sensitive to linguistic cues, and small differences in phrasing can unintentionally signal
 1274 different goals or assumptions. This makes it essential to generate paraphrases that are semantically
 1275 faithful to the original prompt but lexically and syntactically distinct.

1276 To tackle this, we adopt a two-stage approach. Our primary method relies on *cross-lingual translation*,
 1277 a process inherently designed to preserve meaning while altering surface expression (e.g., Youn et al.
 1278 (2016)). Translation captures the core idea of transferring intent across languages while abstracting
 1279 away from specific phrasings. LLMs have demonstrated impressive performance in this task, often
 1280 on par with human translators in consistency and fluency (Karpinska and Iyyer (2023); Yan et al.
 1281 (2024a;b)). We leverage this ability by using GPT-4.1 to perform the translations: starting with an
 1282 original English prompt, we translate it sequentially through two randomly selected intermediate
 1283 languages and then back into English. The result is a paraphrase that retains the original intent but
 1284 exhibits different syntactic and semantic features due to translation-induced variation. Different
 1285 languages vary in how they structure, resulting in varying cross-translations as we vary the source
 1286 languages into and from which we translate the prompts (Lewis et al. (2023)).

1287 To maximize semantic divergence while preserving intent, we randomly sample intermediate lan-
 1288 guages from a diverse set.⁴ Each translation chain is required to include at least one of Chinese,
 1289 Japanese, or Arabic, given their significant structural and lexical distance from English (Chiswick
 1290 and Miller (2005); Lewis et al. (2023)), which helps introduce greater variation in the back-translated
 1291 output.

1292 To ensure the resulting prompts still reflect the original intent, we also use GPT-4.1 as an additional
 1293 check to verify that the intent remains unchanged. Given the original and translated prompt, the

1294 ³Single-answer tasks can make the IC estimator ill-conditioned if the model returns just a single response,
 1295 because there is no variance to decompose; the shares (and ratios like MVS, ICI) become numerically unstable
 1296 or undefined, and the standardized scale becomes ill-posed.

1297 ⁴Chinese, Japanese, Arabic, Korean, Portuguese, Spanish, German, Russian, Italian, French, Hindi

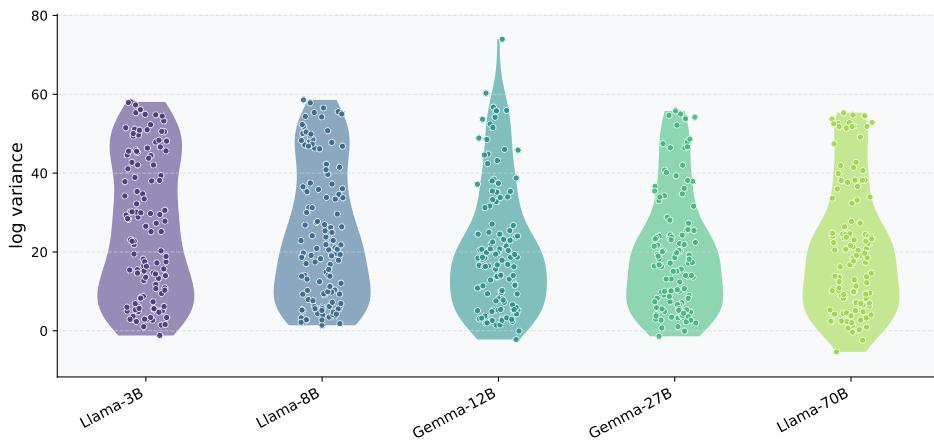
1296 model assesses whether they express the same underlying request. Only those pairs judged equivalent
 1297 are retained. To further enrich the diversity of the prompt set, we generate 500 paraphrases for
 1298 each original question, embed all prompts using sentence embeddings, and apply a greedy selection
 1299 algorithm to identify the 10 most diverse prompts—those with the highest mutual semantic distance.
 1300 This final step ensures that our prompt set not only shares intent but also spans a wide semantic space,
 1301 enabling robust evaluation of the model’s Articulation Sensitivity.

1302 After generating intent-equivalent prompts, we present them to the language model under evaluation.
 1303 Each prompt is paired with three input values that vary the prompt’s main intent — for example,
 1304 income level in a tax question. Then, for each of the 24 prompts, at each of the 5 topics, and for
 1305 each of the 12 values, we prompt the candidate LLM and elicit 25 model responses to capture the
 1306 within-prompt variation.

1307 As discussed in the definition of a sufficient IC, we need to determine how to evaluate the models’
 1308 responses. We focus on extracting a clear bottom-line value: if the model provides a range, we
 1309 take the average. Responses without a definitive answer are discarded, and we continue generating
 1310 responses until we obtain the desired 25 responses. All outputs are generated with a temperature of
 1311 1 to accurately reflect the model’s response distribution ⁵. To extract a usable numeric value from
 1312 each response, we employ GPT-4.1-mini as a post-processor that identifies and retrieves the relevant
 1313 quantity from the model’s output. In order to assure stability of the ANOVA estimation, for each
 1314 model–question pair, we winsorize the outcome by removing responses whose numeric estimate lies
 1315 more than 3 standard deviations from the mean response for that intent, across all paraphrase.

1316 Finally, in our analysis, we evaluate five models: Meta’s LLaMA 3.2 3B Instruct, 3.1 8B Instruct and
 1317 LLaMA 3.3 70B Instruct, as well as Google’s Gemma 3 12B-IT and 27B-IT. We use the OpenRouter
 1318 API to generate multiple responses from each model.

G ADDITIONAL FIGURES



1344 Figure 7: Log variance distribution across the 100 tasks, by model.
 1345
 1346
 1347
 1348

1349 ⁵in the Appendix H we show the effect of temperature on the results

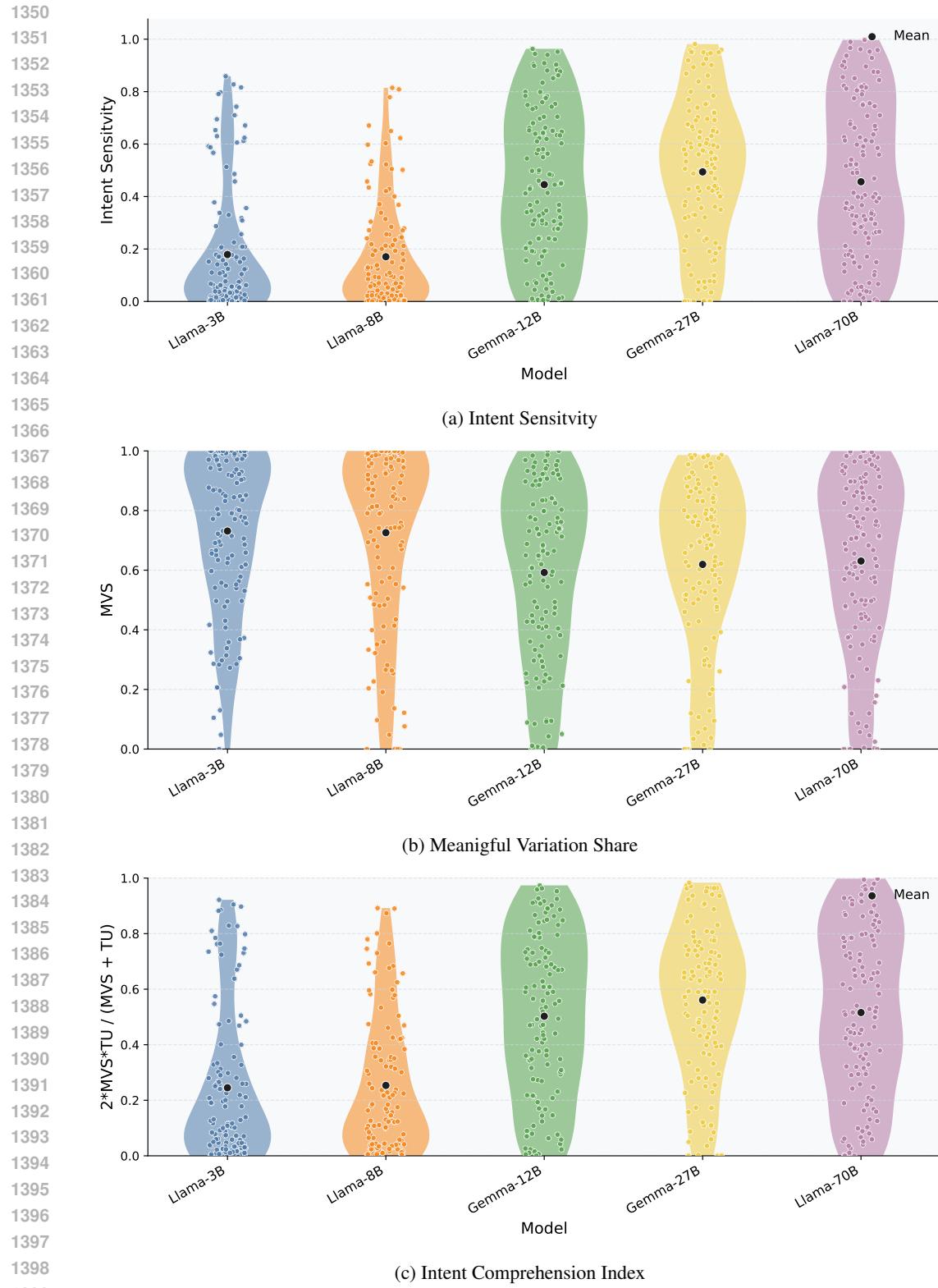


Figure 8: Distribution of components across tasks

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

H TEMPERATURE

The main results report model performance under the natural temperature of 1. Figure 2 illustrates this with a spider graph estimated on 5 questions for each of the 5 topics. To explore the effect of sampling temperature, Figures 9 and 11 compare the same tasks estimated with temperature 0.2 versus temperature 1. Reducing the temperature decreases model uncertainty across all models, with the effect particularly pronounced for the smaller models. The figure further suggests that, within this set, the higher-parameter LLaMA and Gemma models are more sensitive to surface text, whereas the smaller LLaMA models appear relatively more responsive to meaningful changes in intent. Figure 10 presents the corresponding results for temperature 0.5⁶. As expected, its behavior lies between the two other temperature settings.

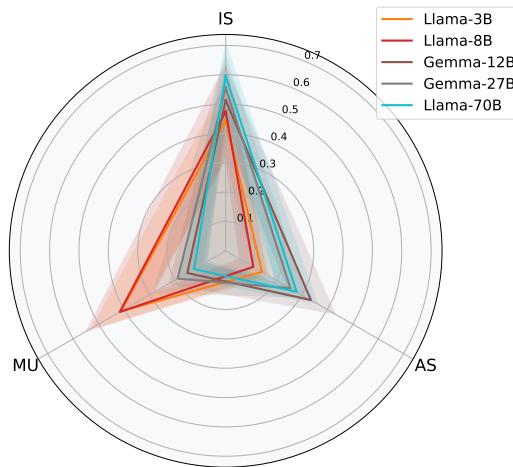


Figure 9: Variance decomposition estimated with temperature 0.2

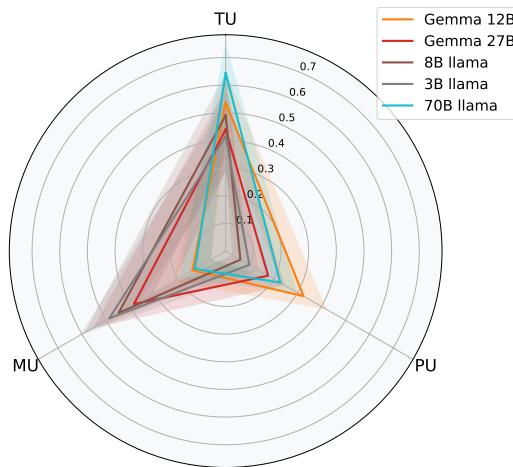


Figure 10: Variance decomposition estimated with temperature 0.5

⁶The Gemma 27B model was run at a temperature of 0.55 due to instability at 0.5 on OpenRouter.

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

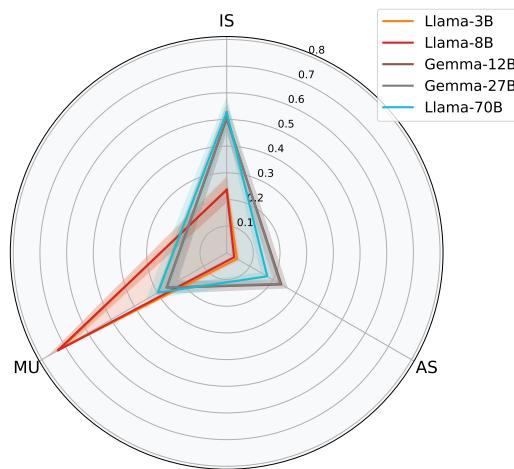


Figure 11: Variance decomposition estimated with temperature 1.

1512 **I PROMPTS**
15131514 **I.1 EXAMPLE QUESTIONS**
15151516 **EXAMPLE QUESTION TEMPLATES**
15171518 **PREVALENCE / COUNTS**
15191520

1. How many {placeholder} are there in a typical city?
2. Approximately how many {placeholder} are produced each farm?
3. What is the total number of {placeholder} registered voters?
4. Roughly how many {placeholder} exist worldwide?
5. How many new {placeholder} were added in 2022?

1526 **AVERAGES / MEANS**
15271528

1. What is the average {placeholder} per high school student?
2. What is the mean {placeholder} recorded each year?
3. On average, how much {placeholder} does an individual consume in a day?
4. What is the typical {placeholder} per unit of output?
5. What is the long-run average {placeholder} for a US county?

1534 **RATES / SHARES / PERCENTAGES**
15351536

1. By what percentage did {placeholder} change between 2010 and 2020?
2. What fraction of total {scope} is accounted for by {placeholder}?
3. What is the annual growth rate of {placeholder}?
4. What proportion of US population owns at least one {placeholder}?
5. What share of all household expenditures goes to {placeholder}?

1543 **TOTALS / AGGREGATES**
15421544

1. Estimate the total {placeholder} required over 5 years horizon.
2. What is the cumulative {placeholder} expected over the next 5 years?
3. What is the projected lifetime total of {placeholder} for a typical student?
4. What is the overall stock of {placeholder} currently in use?
5. How much {placeholder} will be needed to cover national healthcare system for the next decade?

1551 **INTENSITIES / PER-UNIT METRICS**
15521553

1. What is the {placeholder} per unit of output?
2. How much {placeholder} is required per kilometre travelled?
3. What is the {placeholder} per capita in Japan?
4. What is the carbon intensity measured as {placeholder} per kWh?
5. What is the average {placeholder} per square metre?

1559 **MAXIMA / MINIMA / RECORDS**
15601561

1. What is the maximum recorded {placeholder} in a single year?
2. What is the historical peak {placeholder} observed since 2020?
3. What is the lowest recorded {placeholder}?
4. What is the record-high {placeholder} achieved by a single entity?
5. What is the upper bound of {placeholder} under current regulations?

1566 COSTS / MONETARY

1567

- 1568 1. What is the average cost of {placeholder} per square feet?
- 1569 2. What is the expected budget share spent on {placeholder} each decade?
- 1570 3. What is the median price paid for {placeholder} in the US?
- 1571 4. How much investment is required for one unit of {placeholder}?
- 1572 5. What is the total expenditure on {placeholder} in 2015?

1575 DURATIONS / TIMES

1576

- 1577 1. What is the average time needed to complete {placeholder}?
- 1578 2. How long does it typically take for {placeholder} to reach completion?
- 1579 3. What is the mean lifetime of a {placeholder}?
- 1580 4. What is the expected waiting time until {placeholder} occurs?
- 1581 5. What is the typical duration of {placeholder} in the manufacturing sector?

1584 DENSITIES / CONCENTRATIONS

1585

- 1586 1. What is the density of {placeholder} per square kilometre?
- 1587 2. What is the concentration of {placeholder} per litre in the sample?
- 1588 3. What is the average number of {placeholder} per household?
- 1589 4. What is the typical {placeholder} per lane-kilometre of road?
- 1590 5. What is the median {placeholder} per employee in the sector?

1593 PROBABILISTIC / RISK

1594

- 1595 1. What is the probability that {placeholder} occurs within a given year?
- 1596 2. What is the expected frequency of {placeholder} per decade?
- 1597 3. What is the chance of observing at least one {placeholder} in a week?
- 1598 4. What is the return probability to {placeholder} after 5 years?
- 1599 5. What is the expected failure rate expressed as {placeholder} per 1,000 units?

1602 RESOURCE / INPUT COEFFICIENTS

1603

- 1604 1. How much {placeholder} is consumed per tonne of output?
- 1605 2. What is the marginal {placeholder} required for one additional unit?
- 1606 3. What is the input output coefficient of {placeholder} to gross output?
- 1607 4. What is the elasticity of {placeholder} with respect to price?
- 1608 5. What is the shadow cost of one unit of {placeholder}?

1611 VOLUME & CAPACITY

1612

- 1613 1. How many {placeholder} would it take to fill a standard school bus?
- 1614 2. What is the total volume of {placeholder} that flows over Niagara Falls in a single day?
- 1615 3. If all the {placeholder} consumed in the United States in a year were put in a single container, how large would it be?
- 1616 4. What is the total annual production of {placeholder} in France, in liters?
- 1617 5. How many bathtubs could you fill with the amount of {placeholder} consumed globally each day?

1620 WEIGHT & MASS

1621

1622 1. What is the total weight of all the {placeholder} on Earth?

1623 2. Estimate the total mass of all {placeholder} currently in the Netherlands.

1624 3. What is the weight of all the {placeholder} produced by New York City each week?

1625 4. What is the total weight of all the {placeholder} in the state of Texas?

1626 5. Estimate the total mass of all {placeholder} currently airborne over the United States.

1627

1628

1629 LENGTH, DISTANCE & AREA

1630

1631 1. What is the total length, in miles, of all the {placeholder} in Germany?

1632 2. If you laid every {placeholder} eaten in America on July 4th end-to-end, how far would the line stretch?

1633

1634 3. What is the total surface area of all the {placeholder} in China?

1635 4. How many times does an average {placeholder} rotate during its operational lifetime?

1636

1637 5. Estimate the total length of all the {placeholder} sold in North America each holiday season.

1638

1639 FINANCIAL & ECONOMIC

1640

1641 1. How much money, in loose change, is currently in all the {placeholder} in the United States?

1642 2. What is the total cost to fuel all the {placeholder} in California for one day?

1643 3. What is the total annual revenue of all the {placeholder} operating in the United States?

1644

1645 4. How much money is spent on {placeholder} in Canada annually?

1646

1647 5. What is the total market value of all items listed as {placeholder} on eBay worldwide?

1648 TIME & DURATION

1649

1650 1. How many total hours do all people in the United States spend doing {placeholder} each year?

1651 2. How long would it take one person to watch every {placeholder} on YouTube?

1652 3. On average, how many times a day does a person in Japan check their {placeholder}?

1653

1654 4. Estimate the total person-years spent on {placeholder} globally each day.

1655

1656 5. What is the average wait time for a {placeholder} in London during peak hours?

1657 RATES & FREQUENCY

1658

1659 1. Estimate the total number of {placeholder} sent in India every day.

1660 2. How many {placeholder} are sold in the United Kingdom each year?

1661 3. How many {placeholder} are fixed in Chicago each year?

1662

1663 4. What is the consumption rate of {placeholder}, in units per second, in the United States on a Friday night?

1664

1665 5. How many {placeholder} are uploaded to Instagram worldwide every minute?

1666 POPULATION & PROFESSION

1667

1668 1. How many {placeholder} are there in the state of Illinois?

1669 2. Estimate the number of {placeholder} in Brazil.

1670

1671 3. What is the total number of {placeholder} on all the people in Japan?

1672

1673 4. Estimate the total number of people currently airborne in {placeholder} around the world.

5. How many {placeholder} work in Paris?

1674

EVERYDAY OBJECTS & CONSUMPTION

1675

1. How many pairs of {placeholder} does the average person in America own in their lifetime?
2. Estimate the total number of {placeholder} used by all babies in the United States in one year.
3. What is the total amount of {placeholder} consumed in the United States annually?
4. How many gallons of {placeholder} are used by the U.S. newspaper industry annually?
5. How many words does the average person read per day on their {placeholder}?

1682

INFRASTRUCTURE & URBAN

1683

1. What is the total number of {placeholder} in all the buildings of downtown Manhattan?
2. How many {placeholder} are there in the entire city of Tokyo?
3. Estimate the total number of {placeholder} in New York's Central Park?
4. What is the total number of {placeholder} in the Empire State Building?
5. Estimate the total annual electricity consumption of all the {placeholder} in the world.

1689

CREATIVE & ABSTRACT

1690

1. How many individual {placeholder} are on a professional soccer field?
2. How many {placeholder} could you fit in Grand Central Terminal's main concourse?
3. What is the total number of {placeholder} manufactured globally in a single day?
4. What is the total population of {placeholder} in Venice?
5. How many {placeholder} would it take to build a chain to the Moon?

1699

PROBABILISTIC & ODDS

1700

1. What is the probability that a randomly selected person in the United States has {placeholder}?
2. What are the odds of a flight being delayed at {placeholder} International Airport?
3. What is the daily probability that the {placeholder} in a major city experiences a major outage?
4. What is the chance that a new {placeholder} in the United States fails within its first year?
5. If you pick a random word from the New York Times, what is the probability it is the word '{placeholder}'?
6. What is the likelihood of experiencing {placeholder} in London on a given day in July?
7. What is the probability that a car driving one mile on a US highway will get a {placeholder}?
8. Estimate the chance that a randomly selected email in an average inbox is {placeholder}.
9. What is the annual probability of a {placeholder} causing significant damage in California?
10. What are the odds that a randomly chosen {placeholder} from a large supermarket is expired?

1714

J PROMPT SPECIFICATIONS AND DATA COLLECTION PROCEDURE

1716

1717 This section describes all prompts and control settings used in data collection.

1718

J.1 SYSTEM PROMPT USED FOR ANSWER GENERATION

1720

1721 For every question posed to a model (the *estimation task*), we attach the following system message
 1722 and then a user message containing the question text:

1723

1724 You are a helpful assistant designed to answer users
 1725 questions that involve estimating real-world quantities.
 1726 When asked for a numerical value (e.g., average, frequency,
 1727 duration, or count), always provide your best-guess
 1728 estimate, even if you lack exact data. Avoid generic
 1729 refusals like "I don't have that information." If needed,

1728 rely on general knowledge, plausible assumptions, or known
 1729 ranges from similar contexts. The goal is to be useful by
 1730 offering a grounded and reasoned numerical estimate. Return
 1731 a single numeric value in your response, followed by the
 1732 appropriate unit of measurement (e.g., 3 days, 3 kg, 10 mg).
 1733

1734 J.2 BACK-TRANSLATION PROMPTING AND DIVERSITY PROCEDURE

1735 We generate alternative wordings of a seed question via back-translation. Each back-translation hop
 1736 uses the same system translator instruction:
 1737

1738 You are a professional translator. Translate the user text
 1739 to {target_lang}, preserving {placeholder}.
 1740 Preserve meaning exactly. Do not add, remove facts and
 1741 information.
 1742 Keep sentence boundaries and speaker perspective the same.
 1743 Return ***only*** the translation--no commentary.

1744 The corresponding user content is the current question snippet (potentially containing a literal
 1745 {placeholder} token). We first translate English → intermediate language(s), then back to
 1746 English using the same instruction with {target_lang} = English. Placeholders enclosed in
 1747 curly braces are preserved verbatim through all hops.
 1748

1749 **Language chain.** We use a chain of H languages (default $H = 2$) selected to encourage script and
 1750 typological variation. One of {Chinese, Japanese, Arabic} is always included; the remaining hop(s)
 1751 are sampled from: { Chinese, Japanese, Arabic, Korean, Portuguese, Spanish, German, Russian,
 1752 Italian, French, Hindi }. (Exact draws are randomized per back-translation.)

1753 EQUIVALENCE FILTER.

1754 **system:** You are an expert at judging whether two English
 1755 sentences mean exactly the same thing.
 1756 **user:** Do these two sentences convey the same meaning and
 1757 intent?
 1758 Sentence A: "..."
 1759 Sentence B: "..."
 1760 Respond only with "Yes" or "No".
 1761

1762 When a concrete value for {placeholder} is available, the comparison is done after substituting
 1763 that value in both sentences (the raw candidate must still contain the literal {placeholder} token).
 1764

1765 SPEAKER/PERSPECTIVE FILTER.

1766 **system:** You are an expert in grammar and meaning. Your
 1767 job is to assess whether two English sentences use the same
 1768 grammatical subject or speaker perspective. That includes
 1769 whether both use "I", "we", "you", passive voice, or the
 1770 same named subject (e.g., "the city", "Jiji").
 1771 **user:** Do these two sentences use the same speaker or
 1772 subject perspective?
 1773 Sentence A: "..."
 1774 Sentence B: "..."
 1775 Respond only with "Yes" or "No". If the speaker changes or
 1776 a name is added or removed, respond "No".
 1777
 1778
 1779
 1780
 1781

1782 J.2.1 LIST OF QUESTION PROMPTS AND EXTRACTIONS
17831784 HEALTH & NUTRITION
17851786 **Question:** On average, what is the energy expenditure from {placeholder} for an adult during one
1787 hour, in kilocalories per hour?1788 **Values:** ["walking", "running", "cycling", "swimming", "yoga", "dancing", "gardening", "standing",
1789 "typing", "reading", "cleaning", "driving"], You extract a single numeric value from an answer string.
1790 Return only the number and ensure it matches the requested unit.1791 **Extraction Prompt:** Extract a single numeric value representing kilocalories per hour. Return
1792 "None" if there is no numerical value.

1793 Answer text: {answer_text}

1795 **Question:** Estimate the average amount of dietary fiber contained in one serving of {placeholder}, in
1796 grams per serving.1797 **Values:** ["black beans", "oatmeal", "whole wheat bread", "broccoli", "chia seeds", "raspberries",
1798 "almonds", "sweet potato", "green peas", "avocado", "brown rice", "carrots"], You extract a single
1799 numeric value from an answer string. Return only the number and ensure it matches the requested
1800 unit.1801 **Extraction Prompt:** Extract a single numeric value representing grams per serving. Return "None"
1802 if there is no numerical value.

1803 Answer text: {answer_text}

1805 **Question:** Estimate the average number of times a typical household uses a {placeholder} per week,
1806 in uses per week.1807 **Values:** ["microwave", "dishwasher", "vacuum cleaner", "washing machine", "television", "oven",
1808 "toaster", "coffee maker", "blender", "refrigerator door", "air conditioner", "hair dryer"], You extract
1809 a single numeric value from an answer string. Return only the number and ensure it matches the
1810 requested unit.1811 **Extraction Prompt:** Extract a single numeric value representing the number of uses per week; unit:
1812 'uses per week'. Return "None" if there is no numerical value.

1813 Answer text: {answer_text}

1816 **Question:** Estimate the average annual revenue generated by {placeholder} in the health and nutrition
1817 industry in the United States, in dollars per year.1818 **Values:** ["dietary supplements", "organic food products", "vitamin sales", "weight loss programs",
1819 "nutrition consulting services", "sports nutrition drinks", "meal replacement shakes", "functional
1820 foods", "natural health stores", "health coaching services", "nutritional app subscriptions", "protein
1821 powder brands"], You extract a single numeric value from an answer string. Return only the number
1822 and ensure it matches the requested unit.1823 **Extraction Prompt:** Extract a single numeric value reported in dollars per year. Return "None" if
1824 there is no numerical value.

1825 Answer text: {answer_text}

1827 **Question:** Estimate the average body mass index (BMI) of {placeholder} in the United States, in
1828 kilograms per square meter (kg/m^2).1829 **Values:** ["adults", "children", "teenagers", "elderly men", "pregnant women", "college students",
1830 "African Americans", "Asian Americans", "male athletes", "female nurses", "Hispanic women",
1831 "middle-aged adults"], You extract a single numeric value from an answer string. Return only the
1832 number and ensure it matches the requested unit.1833 **Extraction Prompt:** Extract a single numeric value and ensure the unit is kilograms per square meter
1834 (kg/m^2). Return "None" if there is no numerical value.

1835 Answer text: {answer_text}

1836

1837

1838 **Question:** Estimate the average shelf life of {placeholder} in a typical American grocery store, in
1839 days per item.

1840 **Values:** ["milk", "eggs", "lettuce", "yogurt", "chicken breast", "ground beef", "bread loaf", "apples",
1841 "bananas", "cheese block", "carrots", "spinach"], You extract a single numeric value from an answer
1842 string. Return only the number and ensure it matches the requested unit.

1843 **Extraction Prompt:** Extract a single numeric value representing average shelf life; unit must be
1844 'days per item'. Return "None" if there is no numerical value.

1845 Answer text: {answer_text}

1846

1847

1848 **Question:** Estimate the average amount of vitamin C contained in one medium-sized {placeholder},
1849 in milligrams per item.

1850 **Values:** ["orange", "kiwi", "strawberry", "broccoli floret", "bell pepper", "tomato", "grapefruit",
1851 "mango", "papaya", "brussels sprout", "cabbage leaf", "pineapple slice"], You extract a single numeric
1852 value from an answer string. Return only the number and ensure it matches the requested unit.

1853 **Extraction Prompt:** Extract a single numeric value and report it in milligrams per item. Return
1854 "None" if there is no numerical value.

1855 Answer text: {answer_text}

1856

1857

1858 **Question:** Estimate the average annual production of {placeholder} in the United States, in metric
1859 tons per year.

1860 **Values:** ["corn", "soybeans", "wheat", "cotton", "rice", "sugar beets", "potatoes", "tomatoes", "coal",
1861 "steel", "aluminum", "cement"], You extract a single numeric value from an answer string. Return
1862 only the number and ensure it matches the requested unit.

1863 **Extraction Prompt:** Extract a single numeric value representing annual production. Only report in
1864 'metric tons per year'. Return "None" if there is no numerical value.

1865 Answer text: {answer_text}

1866

1867

1868 **Question:** Estimate the average maintenance cost of a {placeholder} used in a typical American
1869 hospital, in dollars per device per year.

1870 **Values:** ["MRI machine", "X-ray machine", "ultrasound scanner", "ventilator", "defibrillator", "ECG
1871 monitor", "infusion pump", "anesthesia machine", "patient monitor", "CT scanner", "sterilizer",
1872 "dialysis machine"], You extract a single numeric value from an answer string. Return only the
1873 number and ensure it matches the requested unit.

1874 **Extraction Prompt:** Extract a single numeric value and report it using the unit 'dollars per device
1875 per year'. Return "None" if there is no numerical value.

1876 Answer text: {answer_text}

1877

1878

1879 **Question:** Estimate the average failure rate for {placeholder} in clinical nutrition settings, in percent
1880 (%) per year.

1881 **Values:** ["infusion pumps", "enteral feeding tubes", "peripheral IV catheters", "central venous
1882 catheters", "parenteral nutrition bags", "glucometers", "electronic medication carts", "feeding pumps",
1883 "blood glucose monitors", "nutrition software systems", "IV fluid warmers", "nutritional supplement
1884 dispensers"], You extract a single numeric value from an answer string. Return only the number and
1885 ensure it matches the requested unit.

1886 **Extraction Prompt:** Extract a single numeric value and report it in percent (%) per year. Return
1887 "None" if there is no numerical value.

1888 Answer text: {answer_text}

1889

1888 **Question:** Estimate the average annual revenue generated by {placeholder} per grocery store in the
1889 United States, in dollars per year.

1890 **Values:** ["fresh produce sales", "dairy products", "bakery items", "meat department", "seafood sales",
 1891 "beverage section", "frozen foods", "prepared meals", "snack foods", "organic products", "household
 1892 goods", "health and beauty aids"], You extract a single numeric value from an answer string. Return
 1893 only the number and ensure it matches the requested unit.

1894 **Extraction Prompt:** Extract a single numeric value expressed in 'dollars per year'. Return "None" if
 1895 there is no numerical value.

1896 Answer text: {answer_text}

1898 **Question:** Estimate the average annual treatment cost, in dollars per year, for a patient with {placeholder}.

1899 **Values:** ["diabetes", "hypertension", "asthma", "rheumatoid arthritis", "multiple sclerosis", "chronic
 1900 kidney disease", "heart failure", "COPD", "psoriasis", "Parkinson's disease", "HIV infection", "breast
 1901 cancer"], You extract a single numeric value from an answer string. Return only the number and
 1902 ensure it matches the requested unit.

1903 **Extraction Prompt:** Extract a single numeric value representing cost and ensure the unit is 'dollars
 1904 per year'. Return "None" if there is no numerical value.

1905 Answer text: {answer_text}

1908 **Question:** What is the estimated annual economic cost of {placeholder} in the United States,
 1909 measured in dollars per year?

1910 **Values:** ["Diet-related chronic diseases", "foodborne illnesses", "drug abuse", "workplace injuries",
 1911 "medical errors", "Childhood obesity", "Adult obesity", "Micronutrient deficiencies", "chronic
 1912 diseases", "environmental pollution", "Type 2 diabetes", "High sodium consumption"], You extract
 1913 a single numeric value from an answer string. Return only the number and ensure it matches the
 1914 requested unit.

1915 **Extraction Prompt:** Extract a single numeric value and report it in 'dollars per year'. Return "None"
 1916 if there is no numerical value.

1917 Answer text: {answer_text}

1920 **Question:** What is the estimated average indoor concentration of {placeholder} in residential homes,
 1921 measured in micrograms per cubic meter?

1922 **Values:** ["particulate matter", "formaldehyde", "benzene", "toluene", "nitrogen dioxide", "carbon
 1923 monoxide", "ammonia", "radon", "ozone", "acetaldehyde", "volatile organic compounds", "mold
 1924 spores"], You extract a single numeric value from an answer string. Return only the number and
 1925 ensure it matches the requested unit.

1926 **Extraction Prompt:** Extract a single numeric value. The unit must be micrograms per cubic meter.
 1927 Return "None" if there is no numerical value.

1928 Answer text: {answer_text}

1930 **Question:** What is the average number of meals prepared using {placeholder} per household per
 1931 month? (unit: meals per month)

1932 **Values:** ["weekdays", "weekends", "holidays", "summer", "winter", "schooldays", "vacation", "festivals",
 1933 "spring", "autumn", "busy days", "quiet days"], You extract a single numeric value from an
 1934 answer string. Return only the number and ensure it matches the requested unit.

1935 **Extraction Prompt:** Extract a single numeric value representing the average number of meals
 1936 prepared using the specified time window/type, reported in meals per month. Return "None" if there
 1937 is no numerical value.

1938 Answer text: {answer_text}

1940 **Question:** What is the estimated annual revenue generated from {placeholder} in the United States,
 1941 measured in dollars per year?

1942 **Values:** ["online advertising", "pharmaceutical sales", "movie ticket sales", "automobile manu-

1944 facturing", "telecommunications services", "video game industry", "fast food chains", "streaming
 1945 subscriptions", "health insurance premiums", "retail e-commerce", "professional sports leagues",
 1946 "music industry"], You extract a single numeric value from an answer string. Return only the number
 1947 and ensure it matches the requested unit.

1948 **Extraction Prompt:** Extract a single numeric value representing the annual revenue from {placeholder}
 1949 in dollars per year. Only report a number and 'dollars per year' as the unit. Return "None" if
 1950 there is no numerical value.

1951 Answer text: {answer_text}

1953 **Question:** What fraction of total public health spending is allocated to {placeholder}, measured in
 1954 percent (%)?

1956 **Values:** ["mental health services", "immunization programs", "maternal care", "chronic disease
 1957 management", "substance abuse prevention", "emergency preparedness", "HIV/AIDS treatment",
 1958 "elderly care", "primary care initiatives", "health education campaigns", "rural health outreach",
 1959 "tuberculosis control"], You extract a single numeric value from an answer string. Return only the
 1960 number and ensure it matches the requested unit.

1961 **Extraction Prompt:** Extract a single numeric value representing the fraction as a percentage (%).
 1962 Return "None" if there is no numerical value.

1963 Answer text: {answer_text}

1964 **Question:** What is the average annual interest paid on {placeholder} by U.S. households, measured
 1965 in dollars per year?

1967 **Values:** ["credit cards", "mortgages", "auto loans", "student loans", "personal loans", "home equity
 1968 lines", "payday loans", "installment loans", "private loans", "business loans", "medical debt", "store
 1969 credit accounts"], You extract a single numeric value from an answer string. Return only the number
 1970 and ensure it matches the requested unit.

1971 **Extraction Prompt:** Extract a single numeric value that answers the question, reported in dollars per
 1972 year. Return "None" if there is no numerical value.

1973 Answer text: {answer_text}

1975 **Question:** What is the average daily calorie intake for {placeholder} in kilocalories per day?

1976 **Values:** ["adult men", "adult women", "teenagers", "infants", "preschool children", "pregnant women",
 1977 "lactating mothers", "elderly adults", "athletes", "office workers", "manual laborers", "vegetarians"],
 1978 You extract a single numeric value from an answer string. Return only the number and ensure it
 1979 matches the requested unit.

1980 **Extraction Prompt:** Extract a single numeric value representing daily calorie intake and report the
 1981 unit as 'kilocalories per day'. Return "None" if there is no numerical value.

1982 Answer text: {answer_text}

1984 **Question:** What is the annual growth rate of cost of {placeholder} in the healthcare sector, measured
 1985 in percent (%) per year?

1987 **Values:** ["prescription drugs", "medical devices", "hospital services", "physician fees", "nursing
 1988 care", "diagnostic tests", "surgical procedures", "insurance premiums", "emergency care", "laboratory
 1989 services", "imaging services", "rehabilitation therapy"], You extract a single numeric value from an
 1990 answer string. Return only the number and ensure it matches the requested unit.

1991 **Extraction Prompt:** Extract a single numeric value representing an annual growth rate, reported in
 1992 percent (%) per year. Return "None" if there is no numerical value.

1993 Answer text: {answer_text}

1995 **Question:** Estimate the number of {placeholder} operating in New York City on an average weekday,
 1996 measured in vehicles per day.

1997 **Values:** ["taxis", "buses", "delivery trucks", "rideshare cars", "garbage trucks", "ambulances", "fire

1998 engines", "limousines", "school buses", "police cars", "construction vehicles", "motorcycles"], You
 1999 extract a single numeric value from an answer string. Return only the number and ensure it matches
 2000 the requested unit.

2001 **Extraction Prompt:** Extract a single numeric value and report it using the unit: vehicles per day.
 2002 Return "None" if there is no numerical value.

2003 Answer text: {answer_text}

2006 **Question:** What is the median distance traveled annually by a {placeholder} in the United States,
 2007 measured in kilometers per year?

2008 **Values:** ["passenger car", "pickup truck", "SUV", "motorcycle", "school bus", "delivery van",
 2009 "taxicab", "minivan", "city bus", "semi-truck", "ambulance", "fire engine"], You extract a single
 2010 numeric value from an answer string. Return only the number and ensure it matches the requested
 2011 unit.

2012 **Extraction Prompt:** Extract a single numeric value representing the median annual distance traveled
 2013 by a {placeholder}. The allowed unit is kilometers per year. Return "None" if there is no numerical
 2014 value.

2015 Answer text: {answer_text}

2017 **Question:** What is the average distance traveled per day by a {placeholder} in urban areas, measured
 2018 in kilometers per day?

2019 **Values:** ["taxi", "bus", "bicycle", "electric scooter", "motorcycle", "delivery van", "ride-share car",
 2020 "ambulance", "garbage truck", "fire engine", "postal vehicle", "private car"], You extract a single
 2021 numeric value from an answer string. Return only the number and ensure it matches the requested
 2022 unit.

2023 **Extraction Prompt:** Extract a single numeric value for average daily distance and use 'kilometers
 2024 per day' as the unit. Return "None" if there is no numerical value.

2025 Answer text: {answer_text}

2027 **Question:** Estimate the total mass of all {placeholder} currently stored in U.S. hospitals, measured
 2028 in kilograms.

2029 **Values:** ["MRI machines", "CT scanners", "ultrasound devices", "X-ray tubes", "ventilators", "in-
 2030 fusion pumps", "dialysis machines", "defibrillators", "anesthesia workstations", "surgical robots",
 2031 "incubators", "ECG monitors"], You extract a single numeric value from an answer string. Return
 2032 only the number and ensure it matches the requested unit.

2033 **Extraction Prompt:** Extract a single numeric value and report it in kilograms. Return "None" if
 2034 there is no numerical value.

2035 Answer text: {answer_text}

2037 **Question:** How many servings of {placeholder} are typically consumed by an adult in the United
 2038 States per week, measured in servings per week?

2039 **Values:** ["breakfast", "lunch", "dinner", "snack", "dessert", "vegetables", "fruits", "meat", "fish",
 2040 "dairy products", "grains", "salads"], You extract a single numeric value from an answer string.
 2041 Return only the number and ensure it matches the requested unit.

2042 **Extraction Prompt:** Extract a single numeric value and specify the unit as 'servings per week'.
 2043 Return "None" if there is no numerical value.

2044 Answer text: {answer_text}

2047 **Question:** On average, how many minutes per week does a person spend on {placeholder}?

2048 **Values:** ["exercise", "reading", "cooking", "commuting", "watching television", "cleaning", "work-
 2049 ing", "shopping", "studying", "socializing", "gardening", "sleeping"], You extract a single numeric
 2050 value from an answer string. Return only the number and ensure it matches the requested unit.

2052
 2053 **Extraction Prompt:** Extract a single numeric value for time spent per week. Allowed unit: minutes
 2054 per week. Return "None" if there is no numerical value.

2055 Answer text: {answer_text}

2056

2057

2058 LOGISTICS

2059

2060 **Question:** What is the estimated number of {placeholder} employed in warehouse logistics world-
 2061 wide, reported as individuals?

2062 **Values:** ["forklift operators", "inventory managers", "shipping coordinators", "order pickers", "ware-
 2063 house supervisors", "logistics analysts", "packers", "receiving clerks", "distribution managers",
 2064 "material handlers", "customs brokers", "freight forwarders"], You extract a single numeric value
 2065 from an answer string. Return only the number and ensure it matches the requested unit.

2066 **Extraction Prompt:** Extract a single numeric value representing individuals. The allowed unit is
 2067 'individuals'. Return "None" if there is no numerical value.

2068 Answer text: {answer_text}

2069

2070

2071 **Question:** Estimate the total annual number of {placeholder} incidents reported in major logistics
 2072 networks worldwide, measured in cases per year.

2073 **Values:** ["cargo theft", "lost shipment", "delayed delivery", "damaged goods", "customs violation",
 2074 "fraudulent invoice", "piracy attack", "cyber breach", "hazardous spill", "stolen container", "misrouted
 2075 package", "supply chain disruption"], You extract a single numeric value from an answer string.
 2076 Return only the number and ensure it matches the requested unit.

2077 **Extraction Prompt:** Extract a single numeric value representing the estimated total annual number
 2078 of {placeholder} incidents, reported in cases per year. Return "None" if there is no numerical value.

2079 Answer text: {answer_text}

2080

2081 **Question:** What is the average financial loss caused by {placeholder} in global supply chains per
 2082 year, reported in US dollars per year?

2083 **Values:** ["cyberattacks", "natural disasters", "port congestion", "trade wars", "labor strikes", "piracy",
 2084 "regulatory changes", "pandemics", "supplier insolvency", "transportation delays", "counterfeit
 2085 goods", "customs bottlenecks"], You extract a single numeric value from an answer string. Return
 2086 only the number and ensure it matches the requested unit.

2087 **Extraction Prompt:** Extract a single numeric value and ensure it is reported in US dollars per year.
 2088 Return "None" if there is no numerical value.

2089 Answer text: {answer_text}

2090

2091

2092 **Question:** What is the historical peak number of {placeholder} operating in global logistics, measured
 2093 in vehicles per year?

2094 **Values:** ["container ships", "cargo planes", "delivery trucks", "freight trains", "oil tankers", "bulk
 2095 carriers", "vans", "electric trucks", "autonomous vehicles", "motorcycles", "reefer trucks", "intermodal
 2096 trailers"], You extract a single numeric value from an answer string. Return only the number and
 2097 ensure it matches the requested unit.

2098 **Extraction Prompt:** Extract a single numeric value for the a quantity in the response. Return "None"
 2099 if there is no numerical value.

2100 Answer text: {answer_text}

2101

2102 **Question:** How much revenue is generated from {placeholder} in the logistics industry worldwide
 2103 each year, measured in US dollars per year?

2104 **Values:** ["freight forwarding", "warehousing", "customs brokerage", "last-mile delivery", "cold chain
 2105 logistics", "e-commerce fulfillment", "express shipping", "third-party logistics services", "reverse
 logistics", "container leasing", "transportation management systems", "supply chain consulting"],

2106 You extract a single numeric value from an answer string. Return only the number and ensure it
 2107 matches the requested unit.

2108 **Extraction Prompt:** Extract a single numeric value representing annual revenue. The allowed unit is
 2109 US dollars per year. Return "None" if there is no numerical value.

2110 Answer text: {answer_text}

2112
 2113 **Question:** What is the probability that a shipping container in transit will experience {placeholder},
 2114 measured in percent per shipment?

2115 **Values:** ["mold growth", "water damage", "infestation", "corrosion", "contamination", "spoilage",
 2116 "condensation", "bacterial infection", "fungal contamination", "rusting", "odorous emission", "rot"],
 2117 You extract a single numeric value from an answer string. Return only the number and ensure it
 2118 matches the requested unit.

2119 **Extraction Prompt:** Extract a single numeric value representing the probability, reported in percent
 2120 per shipment. Return "None" if there is no numerical value.

2121 Answer text: {answer_text}

2122
 2123 **Question:** What is the consumption rate of {placeholder} in a major logistics hub, measured in units
 2124 per hour?

2125 **Values:** ["pallets", "shipping containers", "fuel drums", "packaging materials", "barcode labels",
 2126 "forklift batteries", "loading crates", "delivery vans", "conveyor belts", "storage bins", "handheld
 2127 scanners", "sorting trays"], You extract a single numeric value from an answer string. Return only the
 2128 number and ensure it matches the requested unit.

2129 **Extraction Prompt:** Extract a single numeric value representing the consumption rate and specify
 2130 the unit as 'units per hour'. Return "None" if there is no numerical value.

2131 Answer text: {answer_text}

2132
 2133 **Question:** Estimate the number of {placeholder} operating in Japan at any given moment, measured
 2134 in vehicles.

2135 **Values:** ["taxis", "buses", "trains", "ambulances", "fire trucks", "police cars", "delivery vans",
 2136 "motorcycles", "rental cars", "garbage trucks", "private cars", "ride-sharing vehicles"], You extract
 2137 a single numeric value from an answer string. Return only the number and ensure it matches the
 2138 requested unit.

2139 **Extraction Prompt:** Extract a single numeric value and report it in vehicles. Return "None" if there
 2140 is no numerical value.

2141 Answer text: {answer_text}

2142
 2143 **Question:** Estimate the total energy consumption attributed to {placeholder} in large logistics centers
 2144 worldwide each year, measured in megawatt-hours per year.

2145 **Values:** ["lighting", "heating", "cooling", "ventilation", "material handling equipment", "refrigeration",
 2146 "security systems", "conveyor belts", "automated sorting systems", "charging electric vehicles",
 2147 "water heating", "data processing centers"], You extract a single numeric value from an answer string.
 2148 Return only the number and ensure it matches the requested unit.

2149 **Extraction Prompt:** Extract a single numeric value representing annual energy consumption, using
 2150 megawatt-hours per year as the unit. Return "None" if there is no numerical value.

2151 Answer text: {answer_text}

2152
 2153 **Question:** What is the probability that {placeholder} will experience a major logistics disruption in a
 2154 given year, measured in percent per year?

2155 **Values:** ["regional warehouse", "supply chain", "distribution center", "retail outlet", "manufacturing
 2156 facility", "logistics network", "shipping hub", "inventory system", "transport fleet", "customs terminal",
 2157 "port authority", "fulfillment center"], You extract a single numeric value from an answer string.
 2158 Return only the number and ensure it matches the requested unit.

2160
 2161 **Extraction Prompt:** Extract a single numeric value representing probability. The allowed unit is
 2162 percent per year. Return "None" if there is no numerical value.

2163 Answer text: {answer_text}

2164
 2165 **Question:** Estimate the average interest rate charged for {placeholder} used by logistics companies
 2166 in percent per year.

2167 **Values:** ["working capital loans", "equipment leases", "revolving credit lines", "invoice factoring",
 2168 "asset-based loans", "vehicle financing", "commercial mortgages", "bridge loans", "trade credit
 2169 facilities", "term loans", "letters of credit", "fleet leasing agreements"], You extract a single numeric
 2170 value from an answer string. Return only the number and ensure it matches the requested unit.

2171 **Extraction Prompt:** Extract a single numeric value representing the average interest rate, and specify
 2172 the unit as percent per year (%/year). Return "None" if there is no numerical value.

2173 Answer text: {answer_text}

2174
 2175 **Question:** What is the average annual salary earned by {placeholder} working in the logistics
 2176 industry, measured in US dollars per year?

2177 **Values:** ["warehouse manager", "forklift operator", "supply chain analyst", "logistics coordinator",
 2178 "inventory specialist", "transportation manager", "customs broker", "shipping clerk", "freight dis-
 2179 patcher", "delivery driver", "operations supervisor", "procurement officer"], You extract a single
 2180 numeric value from an answer string. Return only the number and ensure it matches the requested
 2181 unit.

2182 **Extraction Prompt:** Extract a single numeric value representing average annual salary, reported in
 2183 US dollars per year. Return "None" if there is no numerical value.

2184 Answer text: {answer_text}

2185
 2186 **Question:** How much maintenance cost for {placeholder} is incurred per kilometer traveled by a
 2187 delivery truck, measured in US dollars per kilometer?

2188 **Values:** ["engine repairs", "tire replacement", "oil changes", "brake servicing", "transmission main-
 2189 tenance", "suspension repairs", "coolant system upkeep", "battery replacement", "exhaust system
 2190 repair", "air conditioning maintenance", "electrical system servicing", "fuel system cleaning"], You
 2191 extract a single numeric value from an answer string. Return only the number and ensure it matches
 2192 the requested unit.

2193 **Extraction Prompt:** Extract a single numeric value reported in US dollars per kilometer. Return
 2194 "None" if there is no numerical value.

2195 Answer text: {answer_text}

2196
 2197 **Question:** What is the total annual value of {placeholder} issued to logistics companies worldwide,
 2198 measured in US dollars per year?

2199 **Values:** ["trade finance", "invoice factoring", "letters of credit", "equipment leases", "supply chain
 2200 loans", "working capital loans", "commercial paper", "asset-backed securities", "revolving credit
 2201 facilities", "warehouse receipts financing", "export credits", "fleet insurance policies"], You extract
 2202 a single numeric value from an answer string. Return only the number and ensure it matches the
 2203 requested unit.

2204 **Extraction Prompt:** Extract a single numeric value and report it in 'US dollars per year'. Return
 2205 "None" if there is no numerical value.

2206 Answer text: {answer_text}

2207
 2208 **Question:** What is the annual growth rate of revenue from {placeholder} in the logistics sector,
 2209 measured in percent per year?

2210 **Values:** ["freight forwarding", "warehousing services", "last-mile delivery", "customs brokerage",
 2211 "cold chain logistics", "express shipping", "reverse logistics", "e-commerce fulfillment", "intermodal
 2212 transport", "fleet management", "supply chain consulting", "contract logistics"], You extract a single
 2213

2214 numeric value from an answer string. Return only the number and ensure it matches the requested
 2215 unit.

2216 **Extraction Prompt:** Extract a single numeric value representing an annual growth rate. The allowed
 2217 unit is percent per year. Return "None" if there is no numerical value.

2218 Answer text: {answer_text}

2221 **Question:** Estimate the total number of gallons of {placeholder} consumed by the global logistics
 2222 industry each year.

2223 **Values:** ["diesel", "gasoline", "jet fuel", "marine fuel", "biodiesel", "hydrogen", "liquefied natural
 2224 gas", "ethanol blend", "synthetic fuel", "renewable diesel", "compressed natural gas", "aviation fuel"], You extract a single numeric value from an answer string. Return only the number and ensure it
 2225 matches the requested unit.

2226 **Extraction Prompt:** Extract a single numeric value and ensure the unit is gallons per year. Return
 2227 "None" if there is no numerical value.

2228 Answer text: {answer_text}

2231 **Question:** Approximately how many {placeholder} are loaded onto a cargo ship per voyage? (units:
 2232 items per voyage)

2233 **Values:** ["containers", "cranes", "forklifts", "pallets", "vehicles", "generators", "refrigerators", "trac-
 2234 tors", "bulldozers", "excavators", "computers", "machinery"], You extract a single numeric value
 2235 from an answer string. Return only the number and ensure it matches the requested unit.

2236 **Extraction Prompt:** Extract a single numeric value representing the number of items loaded per
 2237 voyage, using 'items per voyage' as the unit. Return "None" if there is no numerical value.

2238 Answer text: {answer_text}

2241 **Question:** What is the average number of crates of {placeholder} delivered to supermarkets in New
 2242 York City per day? (units: crates per day)

2243 **Values:** ["apples", "oranges", "bananas", "lettuce", "tomatoes", "potatoes", "carrots", "onions",
 2244 "grapes", "spinach", "broccoli", "peppers"], You extract a single numeric value from an answer string.
 2245 Return only the number and ensure it matches the requested unit.

2246 **Extraction Prompt:** Extract a single numeric value and ensure the unit is 'crates per day'. Return
 2247 "None" if there is no numerical value.

2248 Answer text: {answer_text}

2250 **Question:** Estimate the total weight of {placeholder} transported by trucks across Europe in metric
 2251 tons per year.

2252 **Values:** ["construction materials", "fresh produce", "electronic goods", "automobiles", "industrial
 2253 machinery", "textiles", "petroleum products", "furniture", "pharmaceuticals", "beverages", "household
 2254 appliances", "steel"], You extract a single numeric value from an answer string. Return only the
 2255 number and ensure it matches the requested unit.

2256 **Extraction Prompt:** Extract a single numeric value representing total weight, and ensure the unit is
 2257 metric tons per year. Return "None" if there is no numerical value.

2258 Answer text: {answer_text}

2261 **Question:** How long does it typically take for {placeholder} to reach completion in days per
 2262 shipment?

2263 **Values:** ["customs clearance", "order processing", "quality inspection", "inventory restocking",
 2264 "freight consolidation", "packaging preparation", "route planning", "document verification", "payment
 2265 confirmation", "load scheduling", "cargo unloading", "delivery coordination"], You extract a single
 2266 numeric value from an answer string. Return only the number and ensure it matches the requested
 2267 unit.

2268 **Extraction Prompt:** Extract a single numeric value representing the typical completion time for
 2269 **{placeholder}**, reported in days per shipment. Return "None" if there is no numerical value.
 2270

2271 Answer text: {answer_text}

2272 **Question:** What percentage of warehouse workers experience {placeholder} each year due to repetitive
 2273 lifting? (units: percent per year)

2274 **Values:** ["lower back pain", "shoulder strain", "tendinitis", "herniated discs", "carpal tunnel syn-
 2275 drome", "muscle fatigue", "sprains", "rotator cuff injuries", "joint inflammation", "elbow pain",
 2276 "chronic soreness", "ligament tears"], You extract a single numeric value from an answer string.
 2277 Return only the number and ensure it matches the requested unit.

2278 **Extraction Prompt:** Extract a single numeric value. Unit must be 'percent per year'. Return "None"
 2279 if there is no numerical value.

2280 Answer text: {answer_text}

2281 **Question:** What is the average amount of {placeholder} transported by a single refrigerated truck in
 2282 kilograms per trip?

2283 **Values:** ["beef", "chicken", "lettuce", "milk", "cheese", "yogurt", "apples", "broccoli", "carrots", "ice
 2284 cream", "fish fillets", "tomatoes"], You extract a single numeric value from an answer string. Return
 2285 only the number and ensure it matches the requested unit.

2286 **Extraction Prompt:** Extract one numeric value representing the average quantity, using kilograms
 2287 per trip as the unit. Return "None" if there is no numerical value.

2288 Answer text: {answer_text}

2289 **Question:** Estimate the total number of {placeholder} handled by a medium-sized logistics company
 2290 in one month (units: shipments per month).

2291 **Values:** ["overnight shipments", "express packages", "international deliveries", "standard parcels",
 2292 "bulk consignments", "fragile items", "return shipments", "same-day deliveries", "temperature-
 2293 controlled goods", "e-commerce orders", "seasonal shipments", "high-value packages"], You extract
 2294 a single numeric value from an answer string. Return only the number and ensure it matches the
 2295 requested unit.

2296 **Extraction Prompt:** Extract a single numeric value representing the estimated number of shipments
 2297 handled per month. The allowed unit is 'shipments per month'. Return "None" if there is no numerical
 2298 value.

2299 Answer text: {answer_text}

2300 **Question:** What is the annual turnover rate for {placeholder} working in warehouse logistics,
 2301 measured in percent per year?

2302 **Values:** ["female employees", "male employees", "temporary staff", "full-time workers", "part-time
 2303 workers", "seasonal workers", "shift supervisors", "older workers", "younger employees", "new hires",
 2304 "contractors", "management staff"], You extract a single numeric value from an answer string. Return
 2305 only the number and ensure it matches the requested unit.

2306 **Extraction Prompt:** Extract a single numeric value representing an annual turnover rate, reported in
 2307 percent per year. Return "None" if there is no numerical value.

2308 Answer text: {answer_text}

2309 **Question:** If all the {placeholder} incidents reported by logistics companies in one year were placed
 2310 into a single file, how many pages would it contain? (units: pages per year)

2311 **Values:** ["vehicle breakdown", "missed delivery", "cargo theft", "shipment delay", "lost package",
 2312 "inventory discrepancy", "equipment malfunction", "damaged goods", "routing error", "documentation
 2313 error", "fuel shortage", "container misplacement"], You extract a single numeric value from an
 2314 answer string. Return only the number and ensure it matches the requested unit.

2322 **Extraction Prompt:** Extract the estimated total number of pages and report the value in 'pages per
 2323 year'. Return "None" if there is no numerical value.
 2324

2325 Answer text: {answer_text}

2327 **Question:** Estimate the total number of cases of {placeholder} reported among long-haul truck
 2328 drivers in the United States per year (units: cases per year).
 2329

2330 **Values:** ["sleep apnea", "hypertension", "diabetes", "depression", "obesity", "back pain", "lung
 2331 cancer", "hepatitis C", "skin infections", "substance abuse", "cardiovascular disease", "chronic
 2332 fatigue"], You extract a single numeric value from an answer string. Return only the number and
 2333 ensure it matches the requested unit.
 2334

2335 **Extraction Prompt:** Extract a single numeric value representing the total annual cases, using 'cases
 2336 per year' as the unit. Return "None" if there is no numerical value.
 2337

2338 Answer text: {answer_text}

2339 **Question:** Estimate the total market value of all {placeholder} currently stored in U.S. warehouses,
 2340 measured in US dollars.
 2341

2342 **Values:** ["soybeans", "automobiles", "furniture", "pharmaceuticals", "electronics", "apparel",
 2343 "petroleum", "coffee beans", "lumber", "steel coils", "corn", "copper wire"], You extract a sin-
 2344 gle numeric value from an answer string. Return only the number and ensure it matches the requested
 2345 unit.
 2346

2347 **Extraction Prompt:** Extract a single numeric value representing the estimated total market value.
 2348 The allowed unit is US dollars. Return "None" if there is no numerical value.
 2349

2350 Answer text: {answer_text}

2351 **Question:** What is the historical peak number of {placeholder} operating in global logistics, measured
 2352 in vehicles per year?
 2353

2354 **Values:** ["container ships", "cargo planes", "delivery trucks", "freight trains", "oil tankers", "bulk
 2355 carriers", "vans", "electric trucks", "autonomous vehicles", "motorcycles", "reefer trucks", "intermodal
 2356 trailers"], You extract a single numeric value from an answer string. Return only the number and
 2357 ensure it matches the requested unit.
 2358

2359 **Extraction Prompt:** Extract a single numeric value followed by 'vehicles per year'. Return "None"
 2360 if there is no numerical value.
 2361

2362 Answer text: {answer_text}

2363 **Question:** How much revenue is generated from {placeholder} in the logistics industry worldwide
 2364 each year, measured in US dollars per year?
 2365

2366 **Values:** ["freight forwarding", "warehousing", "customs brokerage", "last-mile delivery", "cold chain
 2367 logistics", "e-commerce fulfillment", "express shipping", "third-party logistics services", "reverse
 2368 logistics", "container leasing", "transportation management systems", "supply chain consulting"],
 2369 You extract a single numeric value from an answer string. Return only the number and ensure it
 2370 matches the requested unit.
 2371

2372 **Extraction Prompt:** Extract a single numeric value representing annual revenue. The allowed unit is
 2373 US dollars per year. Return "None" if there is no numerical value.
 2374

2375 Answer text: {answer_text}

2376 **Question:** What is the probability that a shipping container in transit will experience {placeholder},
 2377 measured in percent per shipment?
 2378

2379 **Values:** ["mold growth", "water damage", "infestation", "corrosion", "contamination", "spoilage",
 2380 "condensation", "bacterial infection", "fungal contamination", "rusting", "odorous emission", "rot"],
 2381 You extract a single numeric value from an answer string. Return only the number and ensure it
 2382 matches the requested unit.
 2383

2376 **Extraction Prompt:** Extract a single numeric value representing the probability, reported in percent
 2377 per shipment. Return "None" if there is no numerical value.
 2378

2379 Answer text: {answer_text}

2380
 2381 **Question:** What is the consumption rate of {placeholder} in a major logistics hub, measured in units
 2382 per hour?
 2383

2384 **Values:** ["pallets", "shipping containers", "fuel drums", "packaging materials", "barcode labels",
 2385 "forklift batteries", "loading crates", "delivery vans", "conveyor belts", "storage bins", "handheld
 2386 scanners", "sorting trays"], You extract a single numeric value from an answer string. Return only the
 2387 number and ensure it matches the requested unit.

2388 **Extraction Prompt:** Extract a single numeric value representing the consumption rate and specify
 2389 the unit as 'units per hour'. Return "None" if there is no numerical value.

2390 Answer text: {answer_text}

2391
 2392 **Question:** Estimate the number of {placeholder} operating in Japan at any given moment, measured
 2393 in vehicles.
 2394

2395 **Values:** ["taxis", "buses", "trains", "ambulances", "fire trucks", "police cars", "delivery vans",
 2396 "motorcycles", "rental cars", "garbage trucks", "private cars", "ride-sharing vehicles"], You extract
 2397 a single numeric value from an answer string. Return only the number and ensure it matches the
 2398 requested unit.

2399 **Extraction Prompt:** Extract a single numeric value and report it in vehicles. Return "None" if there
 2400 is no numerical value.

2401 Answer text: {answer_text}

2402
 2403 **Question:** Estimate the total energy consumption attributed to {placeholder} in large logistics centers
 2404 worldwide each year, measured in megawatt-hours per year.
 2405

2406 **Values:** ["lighting", "heating", "cooling", "ventilation", "material handling equipment", "refrigeration",
 2407 "security systems", "conveyor belts", "automated sorting systems", "charging electric vehicles",
 2408 "water heating", "data processing centers"], You extract a single numeric value from an answer string.
 2409 Return only the number and ensure it matches the requested unit.

2410 **Extraction Prompt:** Extract a single numeric value representing annual energy consumption, using
 2411 megawatt-hours per year as the unit. Return "None" if there is no numerical value.

2412 Answer text: {answer_text}

2413
 2414 **Question:** What is the probability that {placeholder} will experience a major logistics disruption in a
 2415 given year, measured in percent per year?
 2416

2417 **Values:** ["regional warehouse", "supply chain", "distribution center", "retail outlet", "manufacturing
 2418 facility", "logistics network", "shipping hub", "inventory system", "transport fleet", "customs terminal",
 2419 "port authority", "fulfillment center"], You extract a single numeric value from an answer string.
 2420 Return only the number and ensure it matches the requested unit.

2421 **Extraction Prompt:** Extract a single numeric value representing probability. The allowed unit is
 2422 percent per year. Return "None" if there is no numerical value.

2423 Answer text: {answer_text}

2424
 2425 **Question:** Estimate the average interest rate charged for {placeholder} used by logistics companies
 2426 in percent per year.
 2427

2428 **Values:** ["working capital loans", "equipment leases", "revolving credit lines", "invoice factoring",
 2429 "asset-based loans", "vehicle financing", "commercial mortgages", "bridge loans", "trade credit
 2430 facilities", "term loans", "letters of credit", "fleet leasing agreements"], You extract a single numeric
 2431 value from an answer string. Return only the number and ensure it matches the requested unit.

2432 **Extraction Prompt:** Extract a single numeric value representing the average interest rate, and specify
 2433 the unit as percent per year (%/year). Return "None" if there is no numerical value.

2430 Answer text: {answer_text}

2431

2432

2433 **Question:** What is the average annual salary earned by {placeholder} working in the logistics
2434 industry, measured in US dollars per year?

2435 **Values:** ["warehouse manager", "forklift operator", "supply chain analyst", "logistics coordinator",
2436 "inventory specialist", "transportation manager", "customs broker", "shipping clerk", "freight dis-
2437 patcher", "delivery driver", "operations supervisor", "procurement officer"], You extract a single
2438 numeric value from an answer string. Return only the number and ensure it matches the requested
2439 unit.

2440 **Extraction Prompt:** Extract a single numeric value representing average annual salary, reported in
2441 US dollars per year. Return "None" if there is no numerical value.

2442

2443 Answer text: {answer_text}

2444

2445 **Question:** How much maintenance cost for {placeholder} is incurred per kilometer traveled by a
2446 delivery truck, measured in US dollars per kilometer?

2447 **Values:** ["engine repairs", "tire replacement", "oil changes", "brake servicing", "transmission main-
2448 tenance", "suspension repairs", "coolant system upkeep", "battery replacement", "exhaust system
2449 repair", "air conditioning maintenance", "electrical system servicing", "fuel system cleaning"], You
2450 extract a single numeric value from an answer string. Return only the number and ensure it matches
2451 the requested unit.

2452 **Extraction Prompt:** Extract a single numeric value reported in US dollars per kilometer. Return
2453 "None" if there is no numerical value.

2454

2455 Answer text: {answer_text}

2456

2457 **Question:** What is the total annual value of {placeholder} issued to logistics companies worldwide,
2458 measured in US dollars per year?

2459 **Values:** ["trade finance", "invoice factoring", "letters of credit", "equipment leases", "supply chain
2460 loans", "working capital loans", "commercial paper", "asset-backed securities", "revolving credit
2461 facilities", "warehouse receipts financing", "export credits", "fleet insurance policies"], You extract
2462 a single numeric value from an answer string. Return only the number and ensure it matches the
2463 requested unit.

2464 **Extraction Prompt:** Extract a single numeric value and report it in 'US dollars per year'. Return
2465 "None" if there is no numerical value.

2466

2467

2468 **Question:** What is the annual growth rate of revenue from {placeholder} in the logistics sector,
2469 measured in percent per year?

2470 **Values:** ["freight forwarding", "warehousing services", "last-mile delivery", "customs brokerage",
2471 "cold chain logistics", "express shipping", "reverse logistics", "e-commerce fulfillment", "intermodal
2472 transport", "fleet management", "supply chain consulting", "contract logistics"], You extract a single
2473 numeric value from an answer string. Return only the number and ensure it matches the requested
2474 unit.

2475 **Extraction Prompt:** Extract a single numeric value representing an annual growth rate. The allowed
2476 unit is percent per year. Return "None" if there is no numerical value.

2477

2478

2479 **Question:** Estimate the total number of gallons of {placeholder} consumed by the global logistics
2480 industry each year.

2481 **Values:** ["diesel", "gasoline", "jet fuel", "marine fuel", "biodiesel", "hydrogen", "liquefied natural
2482 gas", "ethanol blend", "synthetic fuel", "renewable diesel", "compressed natural gas", "aviation fuel"], You
2483 extract a single numeric value from an answer string. Return only the number and ensure it
matches the requested unit.

2484 **Extraction Prompt:** Extract a single numeric value and ensure the unit is gallons per year. Return
 2485 "None" if there is no numerical value.
 2486

2487 Answer text: {answer_text}

2489 **Question:** Approximately how many {placeholder} are loaded onto a cargo ship per voyage? (units:
 2490 items per voyage)

2491 **Values:** ["containers", "cranes", "forklifts", "pallets", "vehicles", "generators", "refrigerators", "trac-
 2492 tors", "bulldozers", "excavators", "computers", "machinery"], You extract a single numeric value
 2493 from an answer string. Return only the number and ensure it matches the requested unit.

2494 **Extraction Prompt:** Extract a single numeric value representing the number of items loaded per
 2495 voyage, using 'items per voyage' as the unit. Return "None" if there is no numerical value.

2496 Answer text: {answer_text}

2498 **Question:** What is the average number of crates of {placeholder} delivered to supermarkets in New
 2499 York City per day? (units: crates per day)

2500 **Values:** ["apples", "oranges", "bananas", "lettuce", "tomatoes", "potatoes", "carrots", "onions",
 2501 "grapes", "spinach", "broccoli", "peppers"], You extract a single numeric value from an answer string.
 2502 Return only the number and ensure it matches the requested unit.

2503 **Extraction Prompt:** Extract a single numeric value and ensure the unit is 'crates per day'. Return
 2504 "None" if there is no numerical value.

2505 Answer text: {answer_text}

2508 **Question:** Estimate the total weight of {placeholder} transported by trucks across Europe in metric
 2509 tons per year.

2510 **Values:** ["construction materials", "fresh produce", "electronic goods", "automobiles", "industrial
 2511 machinery", "textiles", "petroleum products", "furniture", "pharmaceuticals", "beverages", "household
 2512 appliances", "steel"], You extract a single numeric value from an answer string. Return only the
 2513 number and ensure it matches the requested unit.

2514 **Extraction Prompt:** Extract a single numeric value representing total weight, and ensure the unit is
 2515 metric tons per year. Return "None" if there is no numerical value.

2516 Answer text: {answer_text}

2518 **Question:** How long does it typically take for {placeholder} to reach completion in days per
 2519 shipment?

2520 **Values:** ["customs clearance", "order processing", "quality inspection", "inventory restocking",
 2521 "freight consolidation", "packaging preparation", "route planning", "document verification", "payment
 2522 confirmation", "load scheduling", "cargo unloading", "delivery coordination"], You extract a single
 2523 numeric value from an answer string. Return only the number and ensure it matches the requested
 2524 unit.

2525 **Extraction Prompt:** Extract a single numeric value representing the typical completion time for
 2526 {placeholder}, reported in days per shipment. Return "None" if there is no numerical value.

2527 Answer text: {answer_text}

2530 **Question:** What percentage of warehouse workers experience {placeholder} each year due to repeti-
 2531 tive lifting? (units: percent per year)

2532 **Values:** ["lower back pain", "shoulder strain", "tendinitis", "herniated discs", "carpal tunnel syn-
 2533 drome", "muscle fatigue", "sprains", "rotator cuff injuries", "joint inflammation", "elbow pain",
 2534 "chronic soreness", "ligament tears"], You extract a single numeric value from an answer string.
 2535 Return only the number and ensure it matches the requested unit.

2536 **Extraction Prompt:** Extract a single numeric value. Unit must be 'percent per year'. Return "None"
 2537 if there is no numerical value.

2538 Answer text: {answer_text}

2539

2540

2541 **Question:** What is the average amount of {placeholder} transported by a single refrigerated truck in
2542 kilograms per trip?

2543 **Values:** ["beef", "chicken", "lettuce", "milk", "cheese", "yogurt", "apples", "broccoli", "carrots", "ice
2544 cream", "fish fillets", "tomatoes"], You extract a single numeric value from an answer string. Return
2545 only the number and ensure it matches the requested unit.

2546 **Extraction Prompt:** Extract one numeric value representing the average quantity, using kilograms
2547 per trip as the unit. Return "None" if there is no numerical value.

2548 Answer text: {answer_text}

2549

2550

2551 **Question:** Estimate the total number of {placeholder} handled by a medium-sized logistics company
2552 in one month (units: shipments per month).

2553 **Values:** ["overnight shipments", "express packages", "international deliveries", "standard parcels",
2554 "bulk consignments", "fragile items", "return shipments", "same-day deliveries", "temperature-
2555 controlled goods", "e-commerce orders", "seasonal shipments", "high-value packages"], You extract
2556 a single numeric value from an answer string. Return only the number and ensure it matches the
2557 requested unit.

2558 **Extraction Prompt:** Extract a single numeric value representing the estimated number of shipments
2559 handled per month. The allowed unit is 'shipments per month'. Return "None" if there is no numerical
2560 value.

2561 Answer text: {answer_text}

2562

2563

2564 **Question:** What is the annual turnover rate for {placeholder} working in warehouse logistics,
2565 measured in percent per year?

2566 **Values:** ["female employees", "male employees", "temporary staff", "full-time workers", "part-time
2567 workers", "seasonal workers", "shift supervisors", "older workers", "younger employees", "new hires",
2568 "contractors", "management staff"], You extract a single numeric value from an answer string. Return
2569 only the number and ensure it matches the requested unit.

2570 **Extraction Prompt:** Extract a single numeric value representing an annual turnover rate, reported in
2571 percent per year. Return "None" if there is no numerical value.

2572 Answer text: {answer_text}

2573

2574

2575 **Question:** If all the {placeholder} incidents reported by logistics companies in one year were placed
2576 into a single file, how many pages would it contain? (units: pages per year)

2577 **Values:** ["vehicle breakdown", "missed delivery", "cargo theft", "shipment delay", "lost package",
2578 "inventory discrepancy", "equipment malfunction", "damaged goods", "routing error", "documentation
2579 error", "fuel shortage", "container misplacement"], You extract a single numeric value from an
2580 answer string. Return only the number and ensure it matches the requested unit.

2581 **Extraction Prompt:** Extract the estimated total number of pages and report the value in 'pages per
2582 year'. Return "None" if there is no numerical value.

2583 Answer text: {answer_text}

2584

2585

2586 **Question:** Estimate the total number of cases of {placeholder} reported among long-haul truck
2587 drivers in the United States per year (units: cases per year).

2588 **Values:** ["sleep apnea", "hypertension", "diabetes", "depression", "obesity", "back pain", "lung
2589 cancer", "hepatitis C", "skin infections", "substance abuse", "cardiovascular disease", "chronic
2590 fatigue"], You extract a single numeric value from an answer string. Return only the number and
2591 ensure it matches the requested unit.

2592 **Extraction Prompt:** Extract a single numeric value representing the total annual cases, using 'cases
2593 per year' as the unit. Return "None" if there is no numerical value.

2594 Answer text: {answer_text}

2592

2593

2594 **Question:** Estimate the total market value of all {placeholder} currently stored in U.S. warehouses,
 2595 measured in US dollars.

2596 **Values:** ["soybeans", "automobiles", "furniture", "pharmaceuticals", "electronics", "apparel",
 2597 "petroleum", "coffee beans", "lumber", "steel coils", "corn", "copper wire"], You extract a sin-
 2598 gle numeric value from an answer string. Return only the number and ensure it matches the requested
 2599 unit.

2600 **Extraction Prompt:** Extract a single numeric value representing the estimated total market value.
 2601 The allowed unit is US dollars. Return "None" if there is no numerical value.

2602 Answer text: {answer_text}

2603

2604

2605 PERSONAL FINANCE

2606

2607 **Question:** On average, how many grams of {placeholder} does a U.S. adult consume per day?

2608 **Values:** ["protein", "fiber", "sugar", "fat", "carbohydrate", "sodium", "cholesterol", "calcium",
 2609 "potassium", "magnesium", "iron", "vitamin C"], You extract a single numeric value from an answer
 2610 string. Return only the number and ensure it matches the requested unit.

2611 **Extraction Prompt:** Extract a single numeric value representing the amount consumed per day,
 2612 reported in grams. Return "None" if there is no numerical value.

2613 Answer text: {answer_text}

2614

2615 **Question:** What is the average monthly electricity usage attributed specifically to {placeholder} in a
 2616 typical U.S. household (kWh per month)?

2617 **Values:** ["refrigerator", "air conditioning", "lighting", "water heater", "clothes dryer", "dishwasher",
 2618 "television", "microwave oven", "freezer", "space heating", "computer equipment", "washing
 2619 machine"], You extract a single numeric value from an answer string. Return only the number and ensure
 2620 it matches the requested unit.

2621 **Extraction Prompt:** Extract a single numeric value and the exact allowed unit: kWh per month.
 2622 Return "None" if there is no numerical value.

2623 Answer text: {answer_text}

2624

2625 **Question:** On average, how much do households in the United States spend on transportation for
 2626 {placeholder} per year (dollars per year)?

2627 **Values:** ["retirees", "single adults", "families with children", "urban residents", "rural households",
 2628 "college students", "low-income families", "high-income households", "immigrants", "senior citizens",
 2629 "military families", "recent graduates"], You extract a single numeric value from an answer string.
 2630 Return only the number and ensure it matches the requested unit.

2631 **Extraction Prompt:** Extract a single numeric value representing dollars spent per year; the allowed
 2632 unit is 'dollars per year'. Return "None" if there is no numerical value.

2633 Answer text: {answer_text}

2634

2635 **Question:** What is the average financial recovery time for a household after experiencing {placeholder}
 2636 in the United States (months per incident)?

2637 **Values:** ["job loss", "medical emergency", "natural disaster", "house fire", "identity theft", "car
 2638 accident", "flooding", "burglary", "divorce", "eviction", "cyberattack", "bankruptcy"], You extract
 2639 a single numeric value from an answer string. Return only the number and ensure it matches the
 2640 requested unit.

2641 **Extraction Prompt:** Extract a single numeric value representing the average number of months
 2642 required for financial recovery per incident; unit: months per incident. Return "None" if there is no
 2643 numerical value.

2644 Answer text: {answer_text}

2645

2646

2647

Question: Estimate the total number of active {placeholder} accounts in the United States (accounts).
Values: ["monthly", "daily", "weekly", "annual", "student", "business", "retail", "savings", "checking", "joint", "corporate", "online"], You extract a single numeric value from an answer string. Return only the number and ensure it matches the requested unit.

Extraction Prompt: Extract a single numeric value representing the estimated total number of active {placeholder} accounts in the United States. Use 'accounts' as the unit. Return "None" if there is no numerical value.

2655 Answer text: {answer_text}

2656

2657

Question: Estimate the average annual child care expense for households with {placeholder} in the United States (dollars per year).

Values: ["single mothers", "two working parents", "military families", "immigrant families", "foster children", "parents under 25", "rural residents", "urban households", "Latino families", "Asian American parents", "low-income households", "same-sex couples"], You extract a single numeric value from an answer string. Return only the number and ensure it matches the requested unit.

Extraction Prompt: Extract a single numeric value representing average annual child care expense. The allowed unit is 'dollars per year.' Return "None" if there is no numerical value.

2665 Answer text: {answer_text}

2666

2667

Question: What is the average monthly income earned from {placeholder} by a typical household in the United States (dollars per month)?

Values: ["rental properties", "dividends", "social security", "freelance work", "pension", "stock investments", "side business", "online sales", "royalties", "child support payments", "interest income", "government assistance"], You extract a single numeric value from an answer string. Return only the number and ensure it matches the requested unit.

Extraction Prompt: Extract a single numeric value. The unit must be dollars per month. Return "None" if there is no numerical value.

2676 Answer text: {answer_text}

2677

2678

Question: What is the average monthly expenditure on {placeholder} for a single adult in the United States (dollars per month)?

Values: ["groceries", "clothing", "toiletries", "household supplies", "electronics", "furniture", "pet food", "medications", "personal care products", "cleaning products", "books", "stationery"], You extract a single numeric value from an answer string. Return only the number and ensure it matches the requested unit.

Extraction Prompt: Extract a single numeric value representing average monthly expenditure, reported in dollars per month. Return "None" if there is no numerical value.

2686 Answer text: {answer_text}

2687

2688

Question: Estimate the total annual amount spent on replacement of {placeholder} in the United States (dollars per year).

Values: ["automobile tires", "roof shingles", "water heaters", "air conditioners", "refrigerators", "light bulbs", "cell phones", "laptop computers", "washing machines", "televisions", "furnaces", "car batteries"], You extract a single numeric value from an answer string. Return only the number and ensure it matches the requested unit.

Extraction Prompt: Extract a single numeric value and ensure it is reported in dollars per year. Return "None" if there is no numerical value.

2697 Answer text: {answer_text}

2698

2699

2700 **Question:** Estimate the average annual household spending on internet services for {placeholder} in
 2701 the United States (dollars per year).

2702 **Values:** ["urban families", "rural households", "millennials", "retirees", "college students", "single-
 2703 parent families", "low-income households", "high-income households", "suburban residents", "tech
 2704 enthusiasts", "remote workers", "senior citizens"], You extract a single numeric value from an answer
 2705 string. Return only the number and ensure it matches the requested unit.

2706 **Extraction Prompt:** Extract a single numeric value representing annual household spending. Use
 2707 'dollars per year' as the unit. Return "None" if there is no numerical value.

2708 Answer text: {answer_text}

2710 **Question:** What is the total number of operational {placeholder} currently in use by households in
 2711 the United States (units)?

2713 **Values:** ["refrigerators", "microwave ovens", "dishwashers", "washing machines", "televisions", "air
 2714 conditioners", "water heaters", "clothes dryers", "vacuum cleaners", "computers", "smoke detectors",
 2715 "space heaters"], You extract a single numeric value from an answer string. Return only the number
 2716 and ensure it matches the requested unit.

2716 **Extraction Prompt:** Extract a single integer value representing the total count. The allowed unit is
 2717 'units'. Return "None" if there is no numerical value.

2718 Answer text: {answer_text}

2721 **Question:** Estimate the total annual electricity consumption of all the {placeholder} in the United
 2722 States (kWh per year).

2723 **Values:** ["refrigerators", "air conditioners", "washing machines", "televisions", "microwave ovens",
 2724 "dishwashers", "water heaters", "clothes dryers", "computers", "freezers", "electric stoves", "space
 2725 heaters"], You extract a single numeric value from an answer string. Return only the number and
 2726 ensure it matches the requested unit.

2727 **Extraction Prompt:** Extract a single numeric value for total annual electricity consumption. The
 2728 allowed unit is kWh per year. Return "None" if there is no numerical value.

2729 Answer text: {answer_text}

2731 **Question:** What is the average monthly payment required to maintain a {placeholder} in the United
 2732 States (dollars per month)?

2734 **Values:** ["mortgage", "car lease", "health insurance plan", "cell phone plan", "student loan", "gym
 2735 membership", "internet subscription", "rent", "childcare service", "homeowners insurance policy",
 2736 "streaming service subscription", "utility bill"], You extract a single numeric value from an answer
 2737 string. Return only the number and ensure it matches the requested unit.

2737 **Extraction Prompt:** Extract a single numeric value and report it in dollars per month. Return "None"
 2738 if there is no numerical value.

2739 Answer text: {answer_text}

2742 **Question:** Estimate the average financial loss from one incident of {placeholder} for a household in
 2743 the United States (dollars per incident).

2744 **Values:** ["burglary", "fire", "flooding", "identity theft", "car theft", "vandalism", "cyberattack",
 2745 "medical emergency", "appliance failure", "water leak", "windstorm damage", "earthquake"], You
 2746 extract a single numeric value from an answer string. Return only the number and ensure it matches
 2747 the requested unit.

2748 **Extraction Prompt:** Extract a single numeric value in dollars per incident, e.g., '700 dollars per
 2749 incident'. Return "None" if there is no numerical value.

2750 Answer text: {answer_text}

2752 **Question:** Estimate the total annual spending on repairs for {placeholder} by households in the
 2753 United States (dollars per year).

2754
 2755 **Values:** ["roofing", "plumbing systems", "heating systems", "air conditioning units", "water heaters",
 2756 "kitchen appliances", "garage doors", "windows", "electrical wiring", "flooring surfaces", "exterior
 2757 siding", "bathroom fixtures"], You extract a single numeric value from an answer string. Return only
 the number and ensure it matches the requested unit.

2758 **Extraction Prompt:** Extract a single numeric value representing total annual spending. The allowed
 2759 unit is 'dollars per year'. Return "None" if there is no numerical value.

2760 Answer text: {answer_text}

2762
 2763 **Question:** Estimate the total annual spending on {placeholder} for all households in the United
 2764 States (dollars per year).

2765 **Values:** ["toilet paper", "laundry detergent", "pet food", "paper towels", "coffee beans", "bottled
 2766 water", "diapers", "cleaning supplies", "milk", "bread", "trash bags", "dish soap"], You extract a single
 2767 numeric value from an answer string. Return only the number and ensure it matches the requested
 unit.

2768 **Extraction Prompt:** Extract a single numeric value and report it in dollars per year. Return "None"
 2769 if there is no numerical value.

2770 Answer text: {answer_text}

2772
 2773 **Question:** Estimate the total number of miles traveled per year by all {placeholder} in the United
 2774 States (miles per year).

2775 **Values:** ["cars", "pickup trucks", "motorcycles", "school buses", "delivery vans", "semi trucks", "city
 2776 buses", "SUVs", "minivans", "ambulances", "fire trucks", "taxis"], You extract a single numeric value
 2777 from an answer string. Return only the number and ensure it matches the requested unit.

2778 **Extraction Prompt:** Extract a single numeric value representing total miles per year. The allowed
 2779 unit is 'miles per year'. Return "None" if there is no numerical value.

2780 Answer text: {answer_text}

2782
 2783 **Question:** Estimate the total annual federal budget allocation for {placeholder} in the United States
 2784 (dollars per year).

2785 **Values:** ["Medicaid", "defense spending", "education", "infrastructure", "scientific research", "vet-
 2786 erans benefits", "environmental protection", "homeland security", "agriculture subsidies", "public
 2787 transportation", "student loans", "healthcare"], You extract a single numeric value from an answer
 2788 string. Return only the number and ensure it matches the requested unit.

2789 **Extraction Prompt:** Extract a single numeric value representing an annual amount, reported in
 'dollars per year.' Return "None" if there is no numerical value.

2790 Answer text: {answer_text}

2792
 2793 **Question:** What is the projected lifetime total spent on {placeholder} for an average adult in the
 2794 United States (dollars over a lifetime)?

2795 **Values:** ["groceries", "healthcare", "education", "transportation", "housing", "vacations", "clothing",
 2796 "dining out", "insurance premiums", "pet care", "entertainment", "childcare"], You extract a single
 2797 numeric value from an answer string. Return only the number and ensure it matches the requested
 2798 unit.

2799 **Extraction Prompt:** Extract a single numeric value representing dollars over a lifetime. Return
 2800 "None" if there is no numerical value.

2801 Answer text: {answer_text}

2803
 2804 **Question:** What is the average annual insurance payout for claims related to {placeholder} in the
 2805 United States (dollars per year)?

2806 **Values:** ["diabetes", "cancer", "stroke", "heart attack", "asthma", "arthritis", "COPD", "hypertension",
 2807 "kidney failure", "depression", "HIV/AIDS", "multiple sclerosis"], You extract a single numeric value
 from an answer string. Return only the number and ensure it matches the requested unit.

2808
 2809 **Extraction Prompt:** Extract a single numeric value representing average annual insurance payout for
 2810 claims related to {placeholder}, reported in dollars per year. Return "None" if there is no numerical
 2811 value.

2812 Answer text: {answer_text}

2813
 2814 **Question:** What is the average amount of {placeholder} contributed to retirement accounts per
 2815 working adult in the United States each year (dollars per year)?

2816 **Values:** ["income", "salary", "wages", "bonus", "commission", "overtime pay", "dividends", "inter-
 2817 est", "tax refund", "gift money", "inheritance", "profit"], You extract a single numeric value from an
 2818 answer string. Return only the number and ensure it matches the requested unit.

2819 **Extraction Prompt:** Extract a single numeric value representing the average annual contribution and
 2820 include the unit 'dollars per year'. Return "None" if there is no numerical value.

2821 Answer text: {answer_text}

2822
 2823 **Question:** What is the average annual revenue generated from {placeholder} by a small business in
 2824 the United States (dollars per year)?

2825 **Values:** ["online sales", "consulting services", "product subscriptions", "advertising", "affiliate
 2826 marketing", "retail operations", "event hosting", "franchise fees", "membership dues", "service
 2827 contracts", "licensing agreements", "training workshops"], You extract a single numeric value from
 2828 an answer string. Return only the number and ensure it matches the requested unit.

2829 **Extraction Prompt:** Extract a single numeric value representing annual revenue. The allowed unit is
 2830 dollars per year. Return "None" if there is no numerical value.

2831 Answer text: {answer_text}

2832
 2833 **Question:** What is the average annual depreciation rate of a {placeholder} in the United States (%
 2834 per year)?

2835 **Values:** ["sedan", "pickup truck", "SUV", "motorcycle", "RV", "commercial van", "luxury car",
 2836 "electric vehicle", "hybrid car", "sports car", "minivan", "cargo trailer"], You extract a single numeric
 2837 value from an answer string. Return only the number and ensure it matches the requested unit.

2838 **Extraction Prompt:** Extract a single numeric value representing an annual depreciation rate, using
 2839 percent per year (% per year) as the unit. Return "None" if there is no numerical value.

2840 Answer text: {answer_text}

2841
 2842 **Question:** What is the average annual household spending on managing {placeholder} in the United
 2843 States (dollars per year)?

2844 **Values:** ["diabetes", "asthma", "hypertension", "arthritis", "obesity", "depression", "allergies", "can-
 2845 cer", "heart disease", "migraine", "eczema", "osteoporosis"], You extract a single numeric value from
 2846 an answer string. Return only the number and ensure it matches the requested unit.

2847 **Extraction Prompt:** Extract a single numeric value representing annual household spending. Unit
 2848 must be 'dollars per year'. Return "None" if there is no numerical value.

2849 Answer text: {answer_text}

2850
 2851 **Question:** Estimate the total annual household consumption of {placeholder} in the United States
 2852 (kilograms per year).

2853 **Values:** ["sugar", "rice", "beef", "cheese", "chicken", "potatoes", "wheat flour", "eggs", "milk
 2854 powder", "pasta", "apples", "fish"], You extract a single numeric value from an answer string. Return
 2855 only the number and ensure it matches the requested unit.

2856 **Extraction Prompt:** Extract a single numeric value representing annual household consumption.
 2857 Only use 'kilograms per year' as the unit. Return "None" if there is no numerical value.

2858 Answer text: {answer_text}

2862 **Question:** Estimate the total number of {placeholder} applications submitted in the United States in
 2863 one year (applications per year).

2864 **Values:** ["patent", "trademark", "copyright", "asylum", "immigration", "student visa", "work permit",
 2865 "green card", "welfare", "unemployment benefit", "Medicaid", "social security"], You extract a single
 2866 numeric value from an answer string. Return only the number and ensure it matches the requested
 2867 unit.

2868 **Extraction Prompt:** Extract a single numeric value representing applications per year. Return
 2869 "None" if there is no numerical value.

2870 Answer text: {answer_text}

2871

2872

2873

2874 SOCIAL PLANNING

2875

2876 **Question:** What is the estimated average salary for {placeholder} working in metropolitan areas of
 2877 Canada? (dollars per year)

2878 **Values:** ["software engineers", "accountants", "registered nurses", "civil engineers", "marketing managers",
 2879 "financial analysts", "primary school teachers", "construction managers", "graphic designers",
 2880 "pharmacists", "electricians", "lawyers"], You extract a single numeric value from an answer string.
 2881 Return only the number and ensure it matches the requested unit.

2882 **Extraction Prompt:** Extract a single numeric value in dollars per year, representing the average
 2883 annual salary for {placeholder} working in Canadian metropolitan areas. Return "None" if there is
 2884 no numerical value.

2885 Answer text: {answer_text}

2886

2887

2888 **Question:** What is the average annual revenue generated from {placeholder} by a midsize US city
 2889 (dollars per year)?

2890 **Values:** ["parking fees", "property taxes", "sales taxes", "hotel occupancy taxes", "business licenses",
 2891 "building permits", "utility services", "transit fares", "recreational facilities", "zoning applications",
 2892 "waste collection services", "franchise agreements"], You extract a single numeric value from an
 2893 answer string. Return only the number and ensure it matches the requested unit.

2894 **Extraction Prompt:** Extract a single numeric value representing average annual revenue, and ensure
 2895 the unit is 'dollars per year'. Return "None" if there is no numerical value.

2896 Answer text: {answer_text}

2897

2898 **Question:** What is the expected waiting time until {placeholder} is approved by local government in
 2899 a typical medium-sized city (days)?

2900 **Values:** ["zoning variance", "building permit", "business license", "environmental impact assessment",
 2901 "liquor license", "noise ordinance waiver", "parking permit", "signage approval", "health code
 2902 exception", "short-term rental permit", "street closure request", "public event permit"], You extract
 2903 a single numeric value from an answer string. Return only the number and ensure it matches the
 2904 requested unit.

2905 **Extraction Prompt:** Extract a single numeric value representing the expected waiting time, reported
 2906 in days. Return "None" if there is no numerical value.

2907 Answer text: {answer_text}

2908

2909 **Question:** What is the cumulative hours volunteered by {placeholder} in public community programs
 2910 over one year in a typical large city (hours per year)?

2911 **Values:** ["teenagers", "retirees", "college students", "corporate employees", "high school teachers",
 2912 "medical professionals", "single parents", "immigrants", "disabled adults", "faith group members",
 2913 "government workers", "young adults"], You extract a single numeric value from an answer string.
 2914 Return only the number and ensure it matches the requested unit.

2915 **Extraction Prompt:** Extract a single numeric value representing total hours volunteered per year;
 2916 report the answer in 'hours per year'. Return "None" if there is no numerical value.

2916 Answer text: {answer_text}

2917

2918

2919 **Question:** What fraction of total emergency shelter usage is accounted for by {placeholder} in a
2920 typical large metropolitan area (% of total usage)?

2921 **Values:** ["fire", "flood", "earthquake", "hurricane", "tornado", "winter storm", "heatwave", "pandemic
2922 outbreak", "power outage", "chemical spill", "civil unrest", "building collapse"], You extract a single
2923 numeric value from an answer string. Return only the number and ensure it matches the requested
2924 unit.

2925 **Extraction Prompt:** Extract a single numeric value representing a percentage. Only provide the
2926 answer as '% of total usage'. Return "None" if there is no numerical value.

2927 Answer text: {answer_text}

2928

2929 **Question:** What is the estimated total area covered by {placeholder} in public parks of a typical
2930 large city (square meters)?

2931 **Values:** ["grass", "flowerbeds", "trees", "playgrounds", "ponds", "walking paths", "bushes", "picnic
2932 areas", "sports fields", "dog parks", "benches zones", "community gardens"], You extract a single
2933 numeric value from an answer string. Return only the number and ensure it matches the requested
2934 unit.

2935 **Extraction Prompt:** Extract a single numeric value and report it in square meters. Return "None" if
2936 there is no numerical value.

2937 Answer text: {answer_text}

2938

2939 **Question:** How many individual {placeholder} are installed in public recreation facilities of a typical
2940 mid-sized U.S. city (units)?

2941 **Values:** ["basketball hoops", "treadmills", "picnic tables", "water fountains", "playground swings",
2942 "benches", "soccer goals", "volleyball nets", "bike racks", "trash cans", "lighting fixtures", "tennis
2943 courts"], You extract a single numeric value from an answer string. Return only the number and
2944 ensure it matches the requested unit.

2945 **Extraction Prompt:** Extract a single numeric value for the estimated number, reported in units.
2946 Return "None" if there is no numerical value.

2947 Answer text: {answer_text}

2948

2949 **Question:** What is the average number of hours per week that a typical community center in a
2950 mid-sized city allocates to {placeholder} (hours per week)?

2951 **Values:** ["youth programs", "fitness classes", "arts workshops", "senior activities", "sports leagues",
2952 "volunteer events", "after-school tutoring", "language courses", "music lessons", "computer training",
2953 "parent meetings", "health seminars"], You extract a single numeric value from an answer string.
2954 Return only the number and ensure it matches the requested unit.

2955 **Extraction Prompt:** Extract a single numeric value representing the average number of hours per
2956 week allocated to {placeholder} in community centers. Only report using 'hours per week'. Return
2957 "None" if there is no numerical value.

2958 Answer text: {answer_text}

2959

2960 **Question:** Estimate the total annual energy consumption attributable to {placeholder} in public
2961 facilities of a typical large city (kWh per year).

2962 **Values:** ["lighting", "heating", "air conditioning", "ventilation systems", "water heating", "elevators",
2963 "computers", "security systems", "kitchen appliances", "pumping stations", "outdoor lighting",
2964 "refrigeration"], You extract a single numeric value from an answer string. Return only the number
2965 and ensure it matches the requested unit.

2966 **Extraction Prompt:** Extract a single numeric value representing energy consumption, using kWh
2967 per year as the unit. Return "None" if there is no numerical value.

2968 Answer text: {answer_text}

2969

2970

2971

2972 **Question:** Estimate the total person-hours spent responding to {placeholder} by municipal emergency
 2973 services per year in a typical large metropolitan area (person-hours per year).

2974 **Values:** ["structure fires", "medical emergencies", "traffic accidents", "hazardous material spills",
 2975 "water rescues", "wildlife incidents", "natural disasters", "false alarms", "gas leaks", "power outages",
 2976 "missing persons cases", "active shooter situations"], You extract a single numeric value from an
 2977 answer string. Return only the number and ensure it matches the requested unit.

2978 **Extraction Prompt:** Extract a single numeric value representing total person-hours, and ensure the
 2979 unit is 'person-hours per year'. Return "None" if there is no numerical value.

2980 Answer text: {answer_text}

2981

2982

2983 **Question:** Estimate the total mass of all {placeholder} currently deployed in public transportation
 2984 systems of a typical large metropolitan area (tons).

2985 **Values:** ["electric buses", "diesel buses", "tram cars", "subway trains", "hybrid buses", "trolleybuses",
 2986 "light rail vehicles", "autonomous shuttles", "double-decker buses", "articulated buses", "compressed
 2987 natural gas buses", "monorail cars"], You extract a single numeric value from an answer string. Return
 2988 only the number and ensure it matches the requested unit.

2989 **Extraction Prompt:** Extract a single numeric value reported in tons. Return "None" if there is no
 2990 numerical value.

2991 Answer text: {answer_text}

2992

2993 **Question:** What is the average number of {placeholder} provided by local governments per year in a
 2994 mid-sized European city (services per year)?

2995 **Values:** ["library books", "building permits", "waste collections", "public events", "health inspec-
 2996 tions", "recycling pickups", "housing grants", "school meals", "parking tickets", "bus routes", "street
 2997 repairs", "water quality tests"], You extract a single numeric value from an answer string. Return
 2998 only the number and ensure it matches the requested unit.

2999 **Extraction Prompt:** Extract a single numeric value reported in services per year. Return "None" if
 3000 there is no numerical value.

3001 Answer text: {answer_text}

3002

3003 **Question:** What is the estimated annual growth rate of {placeholder} implemented for urban safety
 3004 in a typical large city (% per year)?

3005 **Values:** ["surveillance cameras", "facial recognition systems", "emergency alert apps", "crime
 3006 prediction algorithms", "traffic monitoring sensors", "smart street lighting", "body-worn cameras",
 3007 "drones for patrols", "gunshot detection systems", "license plate readers", "panic button networks",
 3008 "public Wi-Fi hotspots"], You extract a single numeric value from an answer string. Return only the
 3009 number and ensure it matches the requested unit.

3010 **Extraction Prompt:** Extract a single numeric value representing the annual growth rate. The allowed
 3011 unit is % per year. Return "None" if there is no numerical value.

3012 Answer text: {answer_text}

3013

3014

3015 **Question:** What is the total number of {placeholder} currently accessible in all municipal libraries of
 3016 a typical large city (units)?

3017 **Values:** ["books", "magazines", "newspapers", "audiobooks", "DVDs", "manuscripts", "maps",
 3018 "journals", "ebooks", "reference guides", "music CDs", "archives"], You extract a single numeric
 3019 value from an answer string. Return only the number and ensure it matches the requested unit.

3020 **Extraction Prompt:** Extract a single numeric value for quantity and use 'units' as the reporting unit.
 3021 Return "None" if there is no numerical value.

3022 Answer text: {answer_text}

3023

3024
 3025 **Question:** What is the elasticity of demand for {placeholder} with respect to price in a typical
 3026 mid-sized U.S. city (unitless)?
 3027

3028 **Values:** ["gasoline", "electricity", "apples", "coffee", "bread", "public transportation", "internet
 3029 service", "movie tickets", "bottled water", "milk", "restaurant meals", "cigarettes"], You extract a
 3030 single numeric value from an answer string. Return only the number and ensure it matches the
 3031 requested unit.
 3032

3033 **Extraction Prompt:** Extract a single numeric value representing elasticity, which must be unitless
 3034 (no units). Return "None" if there is no numerical value.
 3035

3036 Answer text: {answer_text}

3037 **Question:** What is the estimated annual quantity of {placeholder} required for municipal road
 3038 maintenance in a typical large city (tons per year)?
 3039

3040 **Values:** ["asphalt", "gravel", "salt", "sand", "concrete", "bitumen", "crushed stone", "recycled asphalt
 3041 pavement", "topsoil", "cement", "road base material", "aggregate"], You extract a single numeric
 3042 value from an answer string. Return only the number and ensure it matches the requested unit.
 3043

3044 **Extraction Prompt:** Extract a single numeric value and report only in 'tons per year'. Return "None"
 3045 if there is no numerical value.
 3046

3047 Answer text: {answer_text}

3048 **Question:** What is the average number of hours spent on {placeholder} per month in a typical
 3049 mid-sized city's social planning department (hours per month)?
 3050

3051 **Values:** ["community outreach", "data analysis", "policy drafting", "stakeholder meetings", "public
 3052 consultations", "report writing", "budget planning", "event coordination", "staff training", "grant
 3053 applications", "project evaluation", "interdepartmental collaboration"], You extract a single numeric
 3054 value from an answer string. Return only the number and ensure it matches the requested unit.
 3055

3056 **Extraction Prompt:** Extract a single numeric value only. The allowed unit is 'hours per month'.
 3057 Return "None" if there is no numerical value.
 3058

3059 Answer text: {answer_text}

3060 **Question:** What is the typical number of hours per week allocated to {placeholder} in municipal
 3061 youth programs in a large metropolitan area (hours per week)?
 3062

3063 **Values:** ["physical education", "arts instruction", "STEM activities", "community service", "leadership
 3064 training", "sports practice", "mentoring sessions", "health education", "language classes", "career
 3065 exploration", "environmental projects", "technology workshops"], You extract a single numeric value
 3066 from an answer string. Return only the number and ensure it matches the requested unit.
 3067

3068 **Extraction Prompt:** Extract a single numeric value and ensure the unit reported is 'hours per week'.
 3069 Return "None" if there is no numerical value.
 3070

3071 Answer text: {answer_text}

3072 **Question:** What is the average number of public health interventions specifically targeting {placeholder}
 3073 launched by municipal governments per year in a typical urban area (interventions per year)?
 3074

3075 **Values:** ["diabetes", "asthma", "influenza", "obesity", "hypertension", "tuberculosis", "depression",
 3076 "HIV/AIDS", "malaria", "measles", "dengue fever", "hepatitis"], You extract a single numeric value
 3077 from an answer string. Return only the number and ensure it matches the requested unit.
 3078

3079 **Extraction Prompt:** Extract a single numeric value representing the annual count of interventions,
 3080 using 'interventions per year' as the unit. Return "None" if there is no numerical value.
 3081

3082 Answer text: {answer_text}

3083 **Question:** What is the average response time for a {placeholder} reported to social services in a
 3084 typical metropolitan area (minutes)?
 3085

3086 **Values:** ["child abuse case", "domestic violence incident", "elder neglect report", "sexual assault
 3087 allegation", "runaway youth report", "human trafficking tip", "mental health crisis", "drug overdose

3078 call", "missing person report", "animal cruelty complaint", "suicide threat alert", "youth truancy
 3079 notification"], You extract a single numeric value from an answer string. Return only the number and
 3080 ensure it matches the requested unit.

3081 **Extraction Prompt:** Extract a single numeric value representing average response time. The only
 3082 allowed unit is minutes. Return "None" if there is no numerical value.

3083 Answer text: {answer_text}

3086 **Question:** What is the average annual salary paid to {placeholder} working in municipal planning
 3087 departments in a typical U.S. city (dollars per year)?

3088 **Values:** ["urban planners", "civil engineers", "GIS specialists", "zoning inspectors", "transportation
 3089 analysts", "environmental planners", "land use planners", "city managers", "planning technicians",
 3090 "historic preservationists", "community development coordinators", "housing analysts"], You extract
 3091 a single numeric value from an answer string. Return only the number and ensure it matches the
 3092 requested unit.

3093 **Extraction Prompt:** Extract a single numeric value representing the average annual salary and
 3094 ensure the unit is dollars per year. Return "None" if there is no numerical value.

3095 Answer text: {answer_text}

3097 **Question:** What is the total amount of {placeholder} distributed by social welfare agencies per year
 3098 in a typical large city (kilograms per year)?

3099 **Values:** ["food", "rice", "flour", "sugar", "vegetables", "meat", "bread", "milk powder", "canned
 3100 goods", "lentils", "clothing", "diapers"], You extract a single numeric value from an answer string.
 3101 Return only the number and ensure it matches the requested unit.

3102 **Extraction Prompt:** Extract a single numeric value reported in kilograms per year. Return "None" if
 3103 there is no numerical value.

3104 Answer text: {answer_text}

3106 **Question:** Estimate the total annual revenue generated from {placeholder} offered by local govern-
 3107 ments in a typical large metropolitan area (dollars per year).

3109 **Values:** ["municipal bonds", "tax-exempt loans", "public pension funds", "lottery tickets", "parking
 3110 permits", "business licenses", "building permits", "property tax collections", "transit passes", "utility
 3111 bills", "court fines", "development fees"], You extract a single numeric value from an answer string.
 3112 Return only the number and ensure it matches the requested unit.

3113 **Extraction Prompt:** Extract a single numeric value reported in dollars per year. Return "None" if
 3114 there is no numerical value.

3115 Answer text: {answer_text}

3116 **Question:** Estimate the total number of {placeholder} processed by city housing departments per
 3117 year in a typical medium-sized U.S. city (applications per year).

3119 **Values:** ["rental applications", "permit applications", "eviction filings", "subsidy requests", "com-
 3120 plaint forms", "inspection requests", "appeals", "zoning applications", "lease renewals", "housing
 3121 vouchers", "building code violations", "affordable housing applications"], You extract a single nu-
 3122 mERIC value from an answer string. Return only the number and ensure it matches the requested unit.

3123 **Extraction Prompt:** Extract a single numeric value for the total annual applications, using 'appli-
 3124 cations per year' as the unit. Return "None" if there is no numerical value.

3125 Answer text: {answer_text}

3127 **Question:** What is the average maintenance cost of {placeholder} for municipal infrastructure per
 3128 year in a typical large city (dollars per year)?

3130 **Values:** ["road resurfacing", "stormwater management", "waste collection", "street lighting", "bridge
 3131 repairs", "sidewalk upkeep", "traffic signal maintenance", "park landscaping", "public transit facili-
 ties", "sewer system repairs", "snow removal operations", "drainage system cleaning"], You extract

3132 a single numeric value from an answer string. Return only the number and ensure it matches the
 3133 requested unit.

3134 **Extraction Prompt:** Extract a single numeric value representing the average maintenance cost,
 3135 reported in dollars per year. Return "None" if there is no numerical value.

3136 Answer text: {answer_text}

3138

3139 **Question:** On average, how many trips by {placeholder} are made for community events per month
 3140 in a typical mid-sized city (trips per month)?

3141 **Values:** ["bus", "taxi", "rideshare", "bicycle", "scooter", "carpool", "minivan", "shuttle", "motorcycle",
 3142 "vanpool", "tram", "light rail"], You extract a single numeric value from an answer string. Return
 3143 only the number and ensure it matches the requested unit.

3144 **Extraction Prompt:** Extract a single numeric value representing the number of trips by the specified
 3145 vehicle/mode per month, using 'trips per month' as the unit. Return "None" if there is no numerical
 3146 value.

3147 Answer text: {answer_text}

3148

3149

3150 TRANSPORTATION

3151

3152 **Question:** What is the average number of incidents of {placeholder} reported per 1,000 train journeys
 3153 worldwide each year?

3154 **Values:** ["signal failure", "track obstruction", "door malfunction", "brake failure", "engine overheating",
 3155 "derailment", "power outage", "passenger injury", "collision", "vandalism", "fire outbreak",
 3156 "communication breakdown"], You extract a single numeric value from an answer string. Return only
 3157 the number and ensure it matches the requested unit.

3158 **Extraction Prompt:** Extract a single numeric value representing the average number of incidents
 3159 per 1,000 train journeys per year. The allowed unit is 'incidents per 1,000 train journeys per year'.
 3160 Return "None" if there is no numerical value.

3161 Answer text: {answer_text}

3162

3163

3164 **Question:** How much is spent on {placeholder} for public transportation in the United States per
 3165 year (in dollars per year)?

3166 **Values:** ["fuel", "maintenance", "labor", "insurance", "vehicles", "infrastructure", "security", "cleaning",
 3167 "technology upgrades", "administration", "marketing", "energy"], You extract a single numeric
 3168 value from an answer string. Return only the number and ensure it matches the requested unit.

3169 **Extraction Prompt:** Extract a single numeric value and ensure the unit is 'dollars per year'. Return
 3170 "None" if there is no numerical value.

3171 Answer text: {answer_text}

3172

3173

3174 **Question:** Approximately what percentage of annual global freight is transported using {placeholder}
 3175 each year (% per year)?

3176 **Values:** ["container ships", "railroads", "air cargo", "trucks", "bulk carriers", "pipelines", "coastal
 3177 shipping", "river barges", "automated guided vehicles", "drones", "roll-on/roll-off vessels", "inter-
 3178 modal transport"], You extract a single numeric value from an answer string. Return only the number
 3179 and ensure it matches the requested unit.

3179 **Extraction Prompt:** Extract a single numeric value representing a percentage, with unit '% per year'.
 3180 Return "None" if there is no numerical value.

3181 Answer text: {answer_text}

3182

3183

3184 **Question:** Estimate the average fuel consumption for {placeholder} in city traffic, measured in liters
 3185 per 100 kilometers.

3185 **Values:** ["rush hour", "weekday mornings", "weekend evenings", "holiday season", "summer months",

3186 "winter conditions", "rainy days", "snowy periods", "peak traffic hours", "nighttime driving", "school
 3187 drop-off hours", "festive weekends"], You extract a single numeric value from an answer string.
 3188 Return only the number and ensure it matches the requested unit.

3189 **Extraction Prompt:** Extract a single numeric value representing fuel consumption, reported in liters
 3190 per 100 kilometers. Return "None" if there is no numerical value.

3191 Answer text: {answer_text}

K EXAMPLE OF MODEL REWRITTEN PROMPTS

3197 Paraphrase	3198 Original (with placeholder)
3198 What is the annual growth rate of revenues from 3199 freight forwarding in the logistics sector, measured 3200 as a percentage per year?	3198 What is the annual growth rate of revenue from 3199 freight forwarding in the logistics sector, measured 3200 in percent per year?
3200 What is the average monthly income (in US dollars) 3201 that an ordinary household in the United States 3202 receives from rental properties?	3200 What is the average monthly income earned from 3201 rental properties by a typical household in the 3202 United States (dollars per month)?
3202 What is the probability that regional warehouse 3203 will face a significant disruption in logistics ser- 3204 vices during a given year, measured as an annual 3205 percentage?	3202 What is the probability that a regional warehouse 3203 will experience a major logistics disruption in a 3204 given year, measured in percent per year?
3204 What is the average number of public health inter- 3205 ventions specifically related to diabetes that are 3206 launched annually by municipal governments in a 3207 typical urban region (number of interventions per 3208 year)?	3204 What is the average number of public health inter- 3205 ventions specifically targeting diabetes launched 3206 by municipal governments per year in a typical 3207 urban area (interventions per year)?
3205 What is the annual percentage of warehouse op- 3206 erators who experience lower back pain due to 3207 repetitive lifting work? (Unit: annual percentage) 3208 Estimate the average annual interest rate (percent) 3209 for the working capital loans used by logistics com- 3210 panies.	3205 What percentage of warehouse workers experience 3206 lower back pain each year due to repetitive lifting? 3207 (units: percent per year) 3208 Estimate the average interest rate charged for work- 3209 ing capital loans used by logistics companies in 3210 percent per year.
3206 What is the approximate average distance (in kilo- 3207 meters) that migratory bird moves within a year?	3206 What is the average distance (in kilometers) that a 3207 migratory bird travels in one year?
3207 What is the average number of public buses run- 3208 ning daily in city? (Number of buses in one day)	3207 What is the average number of public buses operat- 3208 ing in a typical city per day? (buses per day)
3208 What is the average number of apples boxes deliv- 3209 ered to supermarkets in New York City every day? 3210 (Unit: boxes/day)	3208 What is the average number of crates of apples 3209 delivered to supermarkets in New York City per 3210 day? (units: crates per day)
3209 Please estimate the annual total energy consump- 3210 tion of all MRI machines in U.S. hospitals in units 3211 of kilowatt-hours per year.	3209 Estimate the total annual energy usage of all MRI 3210 machines in U.S. hospitals, measured in kilowatt- 3211 hours per year.
3210 What are the average annual maintenance costs 3211 for road resurfacing (USD/year) in the municipal 3212 infrastructure of a typical large city?	3210 What is the average maintenance cost of road resur- 3211 facing for municipal infrastructure per year in a 3212 typical large city (dollars per year)?
3211 What is the average number of meals prepared 3212 per month by each family using weekdays? (Unit: 3213 meals per month)	3211 What is the average number of meals prepared 3212 using weekdays per household per month? (unit: 3213 meals per month)

3232
 3233
 3234
 3235
 3236
 3237
 3238
 3239