



Contextualized Evaluations: Judging Language Model Responses to Underspecified Queries

Chaitanya Malaviya^{✉*} Joseph Chee Chang[✱] Dan Roth[✱]
Mohit Iyyer^M Mark Yatskar[✱] Kyle Lo[✱]

[✱]University of Pennsylvania [✱]Allen Institute for AI ^MUniversity of Maryland, College Park
cmalaviy@seas.upenn.edu, kylel@allenai.org

Abstract

Language model users often issue queries that lack specification, where the context under which a query was issued—such as the user’s identity, the query’s intent, and the criteria for a response to be useful—is not explicit. For instance, a good response to a subjective query like “*What book should I read next?*” would depend on the user’s preferences, and a good response to an open-ended query like “*How do antibiotics work against bacteria?*” would depend on the user’s expertise. This makes evaluation of responses to such queries an ill-posed task, as evaluators may make arbitrary judgments about the response quality. To remedy this, we present *contextualized evaluations*, a protocol that synthetically constructs context surrounding an underspecified query and provides it during evaluation. We find that the presence of context can 1) alter conclusions drawn from evaluation, even flipping benchmark rankings between model pairs, 2) nudge evaluators to make fewer judgments based on surface-level criteria, like style, and 3) provide new insights about model behavior across diverse contexts. Specifically, our procedure suggests a potential bias towards WEIRD (Western, Educated, Industrialized, Rich and Democratic) contexts in models’ “default” responses and we find that models are not equally sensitive to following different contexts, even when they are provided in prompts.¹

 allenai/ContextEval
 allenai/ContextEval

1 Introduction

Users of language models often issue queries that are underspecified (Spärck-Jones et al., 2007;

^{*}Work done at the Allen Institute for AI.

¹ Our code and data are available at <https://www.github.com/allenai/ContextEval>.

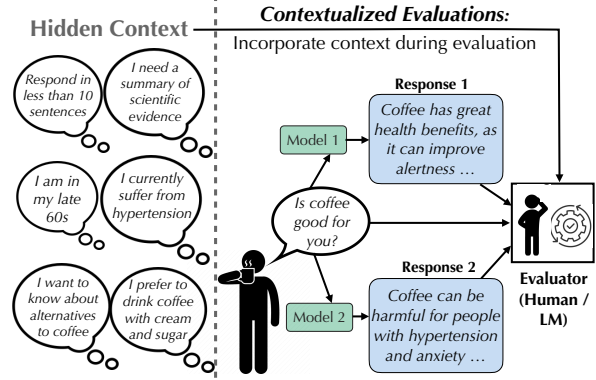


Figure 1: Queries issued by users to language models are often underspecified and can lead to arbitrary evaluation judgments of response quality. We present *contextualized evaluations*, where queries are supplemented with surrounding context during evaluation.

Clarke et al., 2009; Ziegler et al., 2019; Keyvan and Huang, 2022; Herlihy et al., 2024), but common evaluation practices for language models do not account for this. Consider an evaluator presented with a language model’s response to an underspecified query such as “*Is coffee good for you?*” (Figure 1). A language model might respond with an explanation about benefits like antioxidants and mental alertness, but this output would be unacceptable to users with certain health conditions. Can an evaluator make an informed judgment about language model response quality without clear guidance on factors that reveal the intended user’s preferences, background or other necessary criteria for the response to be useful?

In this work, we consider the role of **context** in the evaluation of language model responses to underspecified queries. We propose **contextualized evaluations**—a protocol to synthetically generate and incorporate diverse contexts (represented as follow-up question-answer pairs) when evaluating model responses to underspecified queries. By ap-

plying this procedure to queries from widely-used language model benchmark datasets, we investigate three main research questions:

First, we investigate whether providing context to evaluators has a substantial effect on the conclusions drawn from evaluation. We sample responses from pairs of language models and collect pairwise preference judgments from both model-based as well as human evaluators, in context-agnostic (only the query and model outputs) and context-aware settings (additionally with follow-up questions and answers to clarify the query). Our experiments show not only that **inclusion of context during evaluation can significantly improve agreement between evaluators** (3-10% absolute), but also that **context-aware evaluation can even flip benchmark rankings between model pairs** due to drastic changes in win rates. This raises concern for the reliability of findings produced from context-agnostic evaluations using today’s language model benchmarks, which we find to be full of underspecified queries (Table 1).

Next, we ask whether context changes the criteria used by evaluators for making judgments. Without context, evaluators may make arbitrary judgments which reward surface-level adequacy of a response (Park et al., 2024; Chiang et al., 2024). To evaluate this, we collect free-text justifications from both model and human evaluators. We find that **context-aware evaluation can decrease the frequency with which evaluators make judgments using surface-level properties like style** as opposed to other properties like response relevance (by 5-7% absolute).

Finally, we investigate whether context helps us learn more about the ability of models to adapt to different user contexts. In our work, we use contexts to study biases exhibited by “default” model responses, i.e., those generated without context (§6), to underspecified queries. We observe patterns suggesting a potential **bias where default model responses are better aligned with WEIRD (Western, Educated, Industrialized, Rich and Democratic) contexts** (Henrich et al., 2010), with weak to moderate agreement between human and model evaluators. Context can also allow us to directly evaluate the instruction-following and personalization abilities of models. We show this by evaluating sensitivity of models to different contextual attributes, where we find **considerable disparity in the ability of models to adapt to different**

contexts, even when they’re provided as instructions in prompts (§7).

In summary, our findings suggest that underspecification can have a significant impact on the conclusions and insights drawn from evaluation. To address this, we propose **contextualized evaluations**, a simple and broadly applicable solution that involves synthesizing relevant context and injecting it into existing evaluation protocols (§3). We show that context can increase agreement between evaluators and substantially modify model win rates based on pairwise preference judgements (§5). Further, we show that context can enable us to gather more insights about model behavior, such as identifying contexts that align more closely with default model responses (§6) and assessing model sensitivity to different user contexts (§7). Our work provides a plug-and-play recipe for incorporating context into future language model evaluations.

2 How Prevalent is Underspecification?

Users frequently issue underspecified queries, either to save time or because they are unsure of their information need. The nature of this underspecification can vary across queries, where at one extreme, some queries lack sufficient information to provide a meaningful response (e.g., “best team in the league”) while at the other end, some queries are open-ended, allowing for many valid responses (e.g., “summarize recent work on mRNAs”).

We analyzed a total of 3580 queries by randomly sampling from five existing datasets used for benchmarking language models: Chatbot Arena (Chiang et al., 2024) (2500), AlpacaEval (Li et al., 2023) (645), MTBench (Zheng et al., 2023) (57), ExpertQA (Malaviya et al., 2024) (300) and KIWI (Xu et al., 2024) (78). We performed iterative qualitative coding to develop a schema over types of underspecification, and used GPT-4o to categorize queries (allowing for multiple label assignments). Our results in Table 1 show that the **majority of queries in these benchmarks are open-ended**, while many are also incomplete and subjective.

Amid high prevalence of query underspecification, we posit three possible consequences to address in this work:

1. **Unreliable evaluation conclusions (§5):** Without context, evaluators make subjective or arbitrary judgments that can result in inconsistent and unreliable model evaluations.

	Query Type	Description	Example	Frequency
Degree of underspecification ↑	Incomplete	Missing essential information needed to provide a response (e.g., unresolved coreference).	<i>What is the best team in the league?</i>	18.27%
	Ambiguous	Can be interpreted in multiple ways (e.g., ambiguous word sense).	<i>What is a transformer?</i>	1.87%
	Subjective	Answers based on opinions or personal values (e.g., “best”, “worst”).	<i>Who is the greatest philosopher from the 20th century?</i>	18.69%
	Open-ended	Allows for multiple possible detailed responses (e.g., “explain how”, “describe why”).	<i>Can you summarize recent work on mRNAs?</i>	76.17%
	Closed-ended	Requires specific, concise answer with little room for interpretation.	<i>What is the capital of France?</i>	27.46%

Table 1: Queries from five LM benchmark datasets (Chatbot Arena, AlpacaEval, MTBench, ExpertQA and KIWI), categorized based on the amount of underspecification. Queries can present multiple types (e.g., “*What is the best team in the league?*” is both incomplete and subjective).

- Evaluation focused on surface-level properties (§5):** Lack of context results in evaluators making judgments of model responses based on surface-level criteria like style rather than whether user needs are fully met.
- Limited assessment of contextual adaptability (§6, §7):** When faced with underspecification, models default to generic responses, and as a result, their capacity to handle complex, diverse user preferences is not captured in context-agnostic evaluation.

3 Representing Context as Follow-Up QA

We represent context as follow-up question-answer (QA) pairs, simulating an interactive scenario where an agent can seek clarifying details from the user before responding. For example, for a user issuing an underspecified query about coffee, we can represent relevant medical context:

User: “*Is coffee good for you?*”

*Agent: “*Do you have hypertension?*”

*User: “*Yes.*”

where the follow-up QA pair (marked with *) serves as additional context.

In practical scenarios, this context could be collected through user interaction or be stored by the agent as memory from past conversations.

3.1 Desiderata for Questions

We require that each follow-up question is:

- **Salient:** The question must be relevant and important enough to the user query to warrant a response.

- **Actionable:** The question should anticipate that its answering will directly influence how the user’s query best be addressed.

Context typically describes user attributes or the scope of the user’s query (see Table 2). User attributes include the user’s expertise, age group, location and other characteristics that affect the usefulness of the response. On the other hand, query scope describes the user’s preferences for the response, such as the specific topic or aspect they want covered, the desired length or format, or the need for references.

3.2 Desiderata for Answers

While any one QA-pair constitutes a single contextual specification, when it comes to answers to follow-up questions, we impose requirements on the full *answer set*:

- **Realistic:** The answer choices should be plausible, such that a person could answer the question with any of the choices.
- **Complete:** The answer choices should cover a sufficient number of possible answers to the question.
- **Diverse:** The answer choices should be diverse, such that each answer would require adapting the response to the query in a different way.

Examples of follow-up questions and answer sets for a few queries are shown in Table 3.

3.3 Synthetically Generating Contexts

Scalably collecting many, diverse contexts for study can be difficult. We demonstrate and evaluate the use of language models in generating follow-up questions and their answer sets.

Query Scope		User Attributes	
Focus & Angle	Would you also like information on the <i>causes</i> of fever?	Expertise & Familiarity	What is your level of expertise in biology?
Style & Tone	Do you want this explanation to be technical or high-level?	Geographical Location	Are you looking for global trends or trends in a specific region?
Intent	Are you learning Flask for work or as a hobby?	Age Group	What age group do you belong to?
Level of Detail	How detailed do you want the response to be?	Language Fluency	Would you like the response in French?
Format	Do you want a step-by-step overview or just a summary?	Interests	What sports are you most interested in hearing about?
Need for Resources	Would you like me to provide scientific studies and other references?	Cultural Background	What cultural perspective should be considered in the response?
Clarification	Do you mean apple the company or fruit?	Profession	Are you asking as a researcher or developer?
Intended Audience	Are you preparing this for a general audience or experts?	Political Views	Are you interested in the environmental perspective or economic impact?
Length	How long would you like the summary to be?	Gender Identity	Should I focus on health issues among people of a specific gender?

Table 2: Taxonomy over types of common contextual attributes which can be lacking in underspecified queries, along with examples of follow-up questions for each attribute.

Query	Follow-Up QAs
Give me a sample 5-day itinerary for a Switzerland holiday, starting from Basel.	Q: What is your budget for the trip? A: ["Economy", "Mid-range", "Luxury"] Q: What type of activities are you most interested in? A: ["Outdoor activities", "Cultural experiences", "Historical sites", "Relaxation"] Q: Are you traveling alone or with others? A: ["Alone", "With a partner", "With family", "With a group of friends"]
I am going to make pumpkin pie for the first time. Can you help me?	Q: Do you have any dietary restrictions or preferences? A: ["None", "Gluten-free", "Dairy-free", "Vegan", "Low-sugar"] Q: How many servings are you planning to make? A: ["Small (4-6 servings)", "Medium (8-10 servings)", "Large (12+ servings)"] Q: How much time do you have available for baking? A: ["Under 1 hour", "1-2 hours", "More than 2 hours"]
How long will it take for a child to speak with therapy?	Q: What is the initial diagnosis or reason for requiring therapy? A: ["Speech Delay", "Autism Spectrum Disorder", "Hearing Impairment"] Q: How old is the child? A: ["0-1 years", "2-3 years", "4-5 years", "6+ years"] Q: How frequently is the child receiving therapy? A: ["Once a week", "Multiple times a week", "Occasionally", "Not receiving therapy yet"]

Table 3: Examples of follow-up QAs for a few underspecified queries. Each query has up to 10 such follow-up questions and answer sets associated with each question.

Approach We perform few-shot prompting with GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro; given a query, models are asked to first evaluate whether the query needs further context for generating a response, and if so, to generate a list of follow-up QAs. Our prompt is provided in the Appendix (Table 10).

Results Out of the initial set of 3580 queries, we find that a total of 1881 queries need further context according to all three models. We use this set of 1881 queries for all further experiments. For these queries, we randomly sample a context from one

of the models, which is then validated by a jury of models, which is also the same three models in our case. These models provide a binary label for the importance of each follow-up question. We only retain those follow-up QAs which are found to be important by all three models. There are an average of 9.32 follow-up QAs across queries.

Human evaluation To ensure that our contexts meet the criteria outlined in §3, we ask 251 human annotators recruited from Prolific to validate contexts for a random sample of 154 underspecified queries, where each example is annotated by

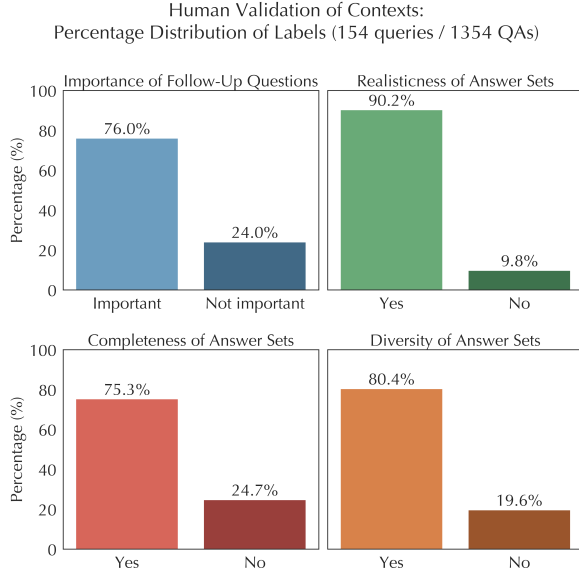


Figure 2: Human validation shows that most generated follow-up questions are important for clarifying an underspecified query and most generated answer sets are realistic, complete and diverse.

3 annotators. Each annotator is shown the underspecified query and the corresponding follow-up QAs, and asked to provide binary labels for the importance of the follow-up question, and the realisticness, completeness and diversity of the answer sets. Further annotation details are provided in Appendix A. The majority label percentages based on this validation are shown in Figure 2. We find that most follow-up questions ($\sim 76\%$) are found to be important, while answer sets are mostly complete ($\sim 75\%$), realistic ($\sim 90\%$) and diverse ($\sim 80\%$). This gives us confidence to rely on these generated contexts for our studies.

4 Designing Contextualized Evaluations

4.1 Evaluation Settings

Context can be provided during response generation, evaluation or both (illustrated in Figure 3). Typical evaluation protocols are conducted in an entirely context-agnostic manner. We study two alternate evaluation protocols, *adaptive evaluation*, where both generation and evaluation are context-aware and *implicit context discovery*, where generation is context-agnostic but evaluators are provided context to rate default responses.

Context-agnostic responses are generated by prompting models to provide a response to the query alone, while context-aware responses are generated by providing a sampled context with

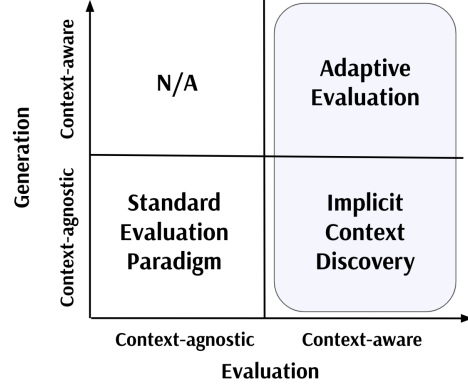


Figure 3: Our work defines two new evaluation settings—*adaptive evaluation* and *implicit context discovery*—distinctive from the *standard evaluation paradigm* which is context-agnostic.

each query. Specifically, we sample a *single* answer to each follow-up question for a query using GPT-4o, such that the context is internally consistent, and provide this list of QA pairs to the model as context. Evaluators are required to perform pairwise judgments. During context-agnostic evaluation, evaluators are provided just the query and two responses, but additionally given the sampled context during context-aware evaluation. Prompts and experimental details are in Appendix B.

1. Standard Evaluation Paradigm

(NoCtxGen-NoCtxEval): In a standard evaluation paradigm, responses to queries are generated without any supplemental context, and independent evaluators are not exposed to any such context either.

2. Implicit Context Discovery

(NoCtxGen-CtxEval): In this setting, responses are generated without context but evaluators are exposed to plausible context surrounding the query. This context is fixed across both responses for pairwise evaluation. Since models are left to assume contexts based on initial queries, we can use this setting to discover implicit biases in default model responses by providing context to the evaluators. We showcase an example of such an analysis in §6.

3. Adaptive Evaluation (CtxGen-CtxEval):

Finally, we consider a setting where both generative models and evaluators are exposed to context, which is fixed across model pairs. In this setting, models showcase how well they can adapt to user contexts, and hence, we can probe the instruction-following and

personalization abilities of models.

4.2 Model Pairs to be Evaluated

We generate responses for all 1881 queries by zero-shot prompting three pairs of candidate models. We selected model pairs based on their relative rankings (at the time of the study) on the Chatbot Arena Leaderboard:²

- GPT-4o vs Gemini-1.5-Flash, which were Rank 3 vs Rank 4, selected as top-ranking proprietary models.
- Claude-3.5-Sonnet vs Llama-3.1-405B, which were Rank 5 vs Rank 6, selected as similarly ranked proprietary and open-weights model.
- Gemma-2-27B vs Jamba-1.5-Large, which were Rank 19 vs Rank 20, selected as similarly ranked open-weights contenders.

4.3 Collecting Human Evaluator Judgments

We recruit our human evaluators through Prolific and ask them to first complete a training task where they evaluate responses to queries in both context-agnostic and context-aware settings. In the context-agnostic setting, evaluators are asked to give an overall preference for one of the two responses or indicate a *Tie*, along with a free-text justification. While in the context-aware setting, evaluators are first asked to evaluate whether the constraints in each one of the follow-up QAs are satisfied by each of the two responses. Once they have provided these ratings, they are required to similarly provide an overall preference along with a free-text justification. The order of the two responses is always randomized. Each evaluator is shown a single query and pairs of responses for one of the three evaluation settings, which is randomly chosen for the evaluator. Every example is judged by 3 different evaluators and we collect ratings for 100 examples for all 3 model pairs and 3 settings, for a total of $100 * 3 * 3 * 3 = 2700$ ratings. Further human evaluation details and interface screenshots are in Appendix A.

4.4 Collecting Model Evaluator Judgments

The use of language model-based evaluators (also known as autoraters or LM-as-judge) is becoming increasingly prevalent. Hence, we also use them to judge responses in all three settings. We use 5 autoraters for each candidate model pair,

which includes the 6 models listed in §4.2 and Qwen2-72B-Instruct (Yang et al., 2024), but always excluding the two candidate models to avoid self-preference bias (Panickssery et al., 2024). In all three evaluation settings, autoraters are instructed to give an overall preference for one of the two responses or indicate a *Tie*, and then provide a free-text justification.

5 How Does Context Change Evaluation Conclusions?

Presence of context improves agreement between evaluators. We report percentage agreement rates for both autoraters and human evaluators in Table 4.³ Significance testing is performed using a paired t-test and values significantly different than the NoCtxGen-NoCtxEval setting are indicated with a * ($p < 0.05$). We find that context-aware evaluation can significantly increase the agreement between evaluators, especially LM-based autoraters and sometimes even human evaluators, suggesting that context is helpful in grounding evaluators together towards more consistent judgments.

While both autoraters and human evaluators generally show increased agreement when provided with context, the magnitude of improvement can vary. For example, autoraters show an improvement of roughly 3–10 points in agreement once context is introduced, whereas humans exhibit a more modest increase of about 3–5 points. This discrepancy suggests that LM-based evaluators may be more sensitive to context than human evaluators, potentially because they systematically focus on whether explicit constraints in the context are met. To investigate this further, we compute human agreement considering only those examples where the two responses differed by at least one in the number of follow-up QAs they satisfied (reported in Table 5). These agreement rates are much higher than the overall agreement rates in Table 4, suggesting that follow-up QAs can provide relatively more objective criteria to distinguish between responses.

Presence of context can change model win rates and flip benchmark rankings. Based on the ma-

²<https://lmarena.ai/?leaderboard>

³To calculate percentage agreement, we determine the majority label and then compute the percentage of judgments matching this label for each example. If all 3 human judgments choose the same option, this would be 100% agreement; if 2 of 3 humans choose the same option but one differs, this would be 66% agreement, and so on. Percentage agreement averages these scores across all examples.

Evaluation Setting	GPT-4o vs Gemini-1.5-Flash	Δ	Claude-3.5-Sonnet vs Llama-3.1-405B	Δ	Gemma-2-27B vs Jamba-1.5-Large	Δ
NoCtxGen-NoCtxEval	70.67 (73.33)	-	65.99 (67.75)	-	71.33 (73.95)	-
NoCtxGen-CtxEval	75.00 (77.15)	4.33 (3.82)	71.33 (72.84)	5.34 (5.09)	74.00 (75.48)	2.67 (1.53)
CtxGen-CtxEval	76.00 (76.36)	5.33 (3.03)	74.67 (77.19)	8.68* (9.44)*	74.33 (75.79)	3.00 (1.84)

(a) Human Evaluator

Evaluation Setting	GPT-4o vs Gemini-1.5-Flash	Δ	Claude-3.5-Sonnet vs Llama-3.1-405B	Δ	Gemma-2-27B vs Jamba-1.5-Large	Δ
NoCtxGen-NoCtxEval	68.03 (68.73)	-	64.00 (65.04)	-	73.22 (74.40)	-
NoCtxGen-CtxEval	78.34 (78.40)	10.31* (9.67)*	74.57 (74.89)	10.57* (9.85)*	80.98 (81.01)	7.76* (6.61)*
CtxGen-CtxEval	74.51 (74.56)	6.48* (5.83)*	74.93 (75.13)	10.93* (10.09)*	75.94 (75.99)	2.72* (1.59)*

(b) Model Autorater

Table 4: Agreement rate (max 100%) across context-agnostic and context-aware evaluation settings. Pairs of model output judged for overall preference by three human evaluators (§4.3) and five autoraters (§4.4); agreement calculated both excluding ties and including ties (in parentheses). Δ shows absolute increase in agreement rate over baseline context-agnostic setting (NoCtxGen-NoCtxEval); significant differences ($p < 0.05$) indicated with *.

Evaluation Setting	GPT-4o vs Gemini-1.5-Flash	Δ	Claude-3.5-Sonnet vs Llama-3.1-405B	Δ	Gemma-2-27B vs Jamba-1.5-Large	Δ
NoCtxGen-NoCtxEval	68.03	-	64.00	-	73.22	-
NoCtxGen-CtxEval	82.42	14.39	79.43	15.43	82.05	8.83
CtxGen-CtxEval	81.40	13.37	77.33	13.33	82.54	9.32

Table 5: Agreement rates in human evaluator pairwise judgments (without ties) across context-aware evaluation settings, considering only those examples where the two responses differed by at least one in the number of follow-up QAs they satisfied. We find agreement is substantially higher for this subset of examples than the overall agreement rates reported in Table 4.

jority votes from autoraters and human evaluators, we report win rates for all model pairs in Table 6. Note that we exclude those examples where there was no clear majority vote. Overall, the autorater and human evaluator win rates suggest that win rates can substantially vary across evaluation settings. Importantly, we find that relative rankings between pairs of models can flip in settings where evaluators are provided context versus when they are not. For instance, in Table 6, GPT-4o has a lower win rate than Gemini-1.5-Flash in the setting NoCtxGen-NoCtxEval but has a much higher win rate in the setting CtxGen-CtxEval.

We observe similar trends between human evaluators and autoraters when examining magnitude of win rate shifts across evaluation settings, but note some key differences. For instance, while autoraters show a more significant flip favoring GPT-4o once context is provided, human evaluation shows a milder change, and is accompanied by an increased tie rate. When both responses in-

corporate the provided context sufficiently well, humans are less inclined to form a clear preference, resulting in a higher tie rate. On the other hand, autoraters rely on explicit checks for context fulfillment, which accentuates small differences in how each response incorporates context.

Presence of context can help evaluators avoid reliance on surface-level criteria to judge responses. Next, we investigate whether context changes the criteria used by evaluators to make judgments. To do this, we analyze the free-text justifications provided by human evaluators as well as autoraters. We automatically code these justifications into two categories: 1) *surface-level criteria* which includes criteria such as clarity, conciseness, style or formatting, tone and length, and 2) *content-level criteria*, which includes criteria such as relevance, correctness, completeness, level of detail and context adherence. We then automatically classify justifications into these two categories using GPT-4o. Trends across all evaluation settings for

Evaluation Setting	Human Evaluator			Model Autorater		
	GPT-4o	Gemini-1.5-Flash	Tie	GPT-4o	Gemini-1.5-Flash	Tie
NoCtxGen-NoCtxEval	40.24	48.78	10.98	39.07	53.00	7.92
NoCtxGen-CtxEval	40.86	50.54	8.60	53.69	46.04	0.27
CtxGen-CtxEval	52.75	32.97	14.29	68.05	31.79	0.16

Evaluation Setting	Human Evaluator			Model Autorater		
	Claude-3.5-Sonnet	Llama-3.1-405B	Tie	Claude-3.5-Sonnet	Llama-3.1-405B	Tie
NoCtxGen-NoCtxEval	58.75	37.50	3.75	47.36	47.07	5.57
NoCtxGen-CtxEval	45.88	36.47	17.65	44.86	54.57	0.57
CtxGen-CtxEval	55.55	24.44	20.00	75.29	24.49	0.23

Evaluation Setting	Human Evaluator			Model Autorater		
	Gemma-2-27B	Jamba-1.5-Large	Tie	Gemma-2-27B	Jamba-1.5-Large	Tie
NoCtxGen-NoCtxEval	38.82	54.12	7.06	55.16	39.16	5.68
NoCtxGen-CtxEval	36.67	50.00	13.33	46.65	52.91	0.44
CtxGen-CtxEval	49.43	34.48	16.09	63.89	35.93	0.17

Table 6: Model win rates across context-agnostic and context-aware evaluation settings. Pairs of model output judged for overall preference by three human evaluators (§4.3) and five autoraters (§4.4); wins are determined by majority vote. In each setting, the overall preferred model is indicated in **bold**.

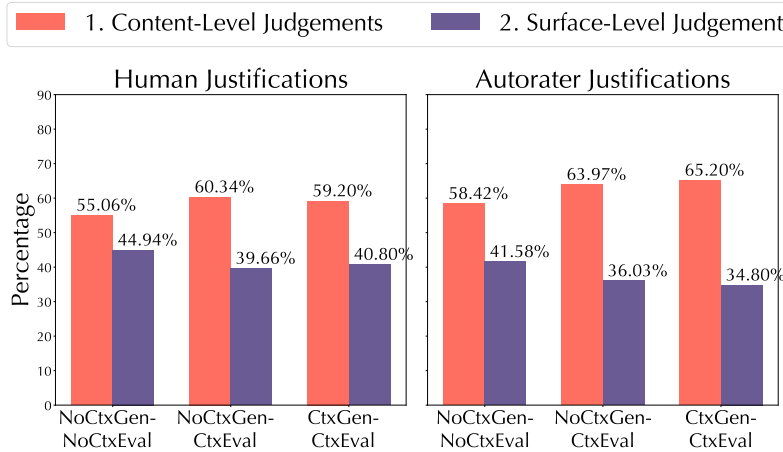


Figure 4: Types of human and autorater justifications across all evaluations settings. Note that there is a lower percentage of justifications based on surface-level criteria for the context-aware evaluation settings.

both human evaluators and autoraters are shown in Figure 4. We note that in context-aware evaluation settings, there is a lower percentage of justifications that are based on surface-level criteria and more that are based on content-level criteria.

6 Which Contexts are “Default”?

Prior work has proposed methods to identify biases in language models through prompts tailored to uncover biases (Li et al., 2020; Santurkar et al., 2023; Deshpande et al., 2023; Cheng et al., 2023; Durmus

et al., 2024). Instead, we study what contexts do models default to when presented with underspecified queries. We investigate the prevalence of these implicit contexts in default model responses using the evaluation setting NoCtxGen-CtxEval where responses are generated without context, but evaluators are shown context specific to a contextual attribute (such as "User Expertise").

6.1 Methodology

To investigate these implicit biases, we first define a list of contextual attributes and corresponding

Implicit Contexts of Default Responses

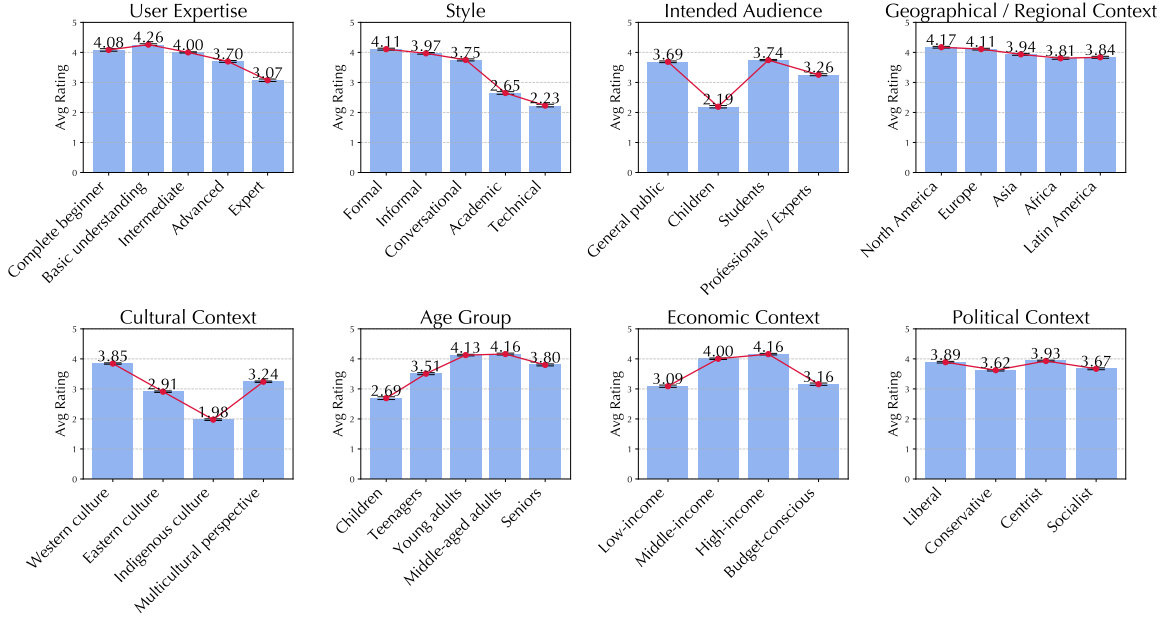


Figure 5: Relevance ratings of default responses from GPT-4o across various contextual attributes, as rated by Gemini-1.5-Pro. These plots suggest that default GPT-4o responses are better aligned towards users from western cultural contexts, high-income individuals and young and middle-aged adults.

follow-up QAs (listed in Table 8 and based on Table 2). For each attribute, we first filter for those queries where this attribute is important to respond to the query and where the query is independent of the answer choices. For instance, a query such as "What is distillation in machine learning?" would correspond to the attribute "User Expertise". Filtering is done using GPT-4o on 23,935 queries from all 5 datasets used in earlier experiments (prompt in Table 14). We then sample up to 1000 filtered queries randomly for each attribute and generate default responses (without context) for these queries using a candidate model.

To evaluate these default responses, we ask an automatic evaluator to provide absolute ratings for the default response (on a scale of 1-5) for response relevance for every value of each contextual attribute (prompt in Table 15). For instance, for a contextual attribute like "User Expertise", the evaluator would be asked to provide a rating for every possible value of this attribute ("Complete beginner", "Basic Understanding", ... , "Expert"). We then compute the average rating for every value of each contextual attribute and plot these trends in Figure 5. These results use GPT-4o as the candidate model, and Gemini-1.5-Pro as the evaluator.

For a subset of three contextual attributes ("Age Group", "Economic Context", "Cultural Context"),

we also collect Likert ratings from human evaluators for the relevance of default responses to different possible values of each attribute (for e.g., possible values for "Age Group" -> [Children, Teenagers, Young adults, Middle-aged adults, Seniors]) for 75 queries (25 per contextual attribute). We then compute Spearman's ρ between human ratings for each query with the autorater judgements on the same query. Across the three contextual attributes, we obtain weak to moderate correlations (Age Group = 0.651, Cultural Context = 0.531, Economic Context = 0.276). We would like to emphasize that this task is highly subjective, which might affect the extent to which human ratings correlate with autorater ratings. At the same time, we find it encouraging that a subset of these attributes do have moderately positive correlations, which gives more credence to our findings below.

6.2 Findings

First, we note that default responses are better catered to users who have a basic understanding of the topic of a query as opposed to experts, and the language they use lacks technical depth. Further, our results suggest potential presence of a WEIRD-like bias, where default responses are better aligned with western cultural contexts, middle-to-high income individuals, and young and middle-

aged adults. However, stronger correlations between human and model evaluators are required to strengthen this conclusion. Since context-agnostic evaluations might overlook disparities in how well default responses serve different contexts, we recommend future work to conduct similar analysis to discover implicit biases in default model responses.

7 Which Contexts are Harder to Follow?

7.1 Methodology

A common approach to adapting models to different user contexts is providing these contexts as part of a prompt; models that are capable of adapting to various user contexts are broadly more useful for more individuals. Context-agnostic evaluations can miss out on capturing the adaptability of models to various contexts. We use our evaluation setting CtxGen-CtxEval to investigate how robustly models can adapt to different values of a contextual attribute. Similar to the analysis in §6, we use the contextual attributes and follow-up QAs defined in Table 8. However, we now generate adapted responses for every value of each contextual attribute where the follow-up QA (e.g., “*What is your level of expertise on this topic?*”) and a specific answer (e.g., “*Complete beginner*”) is provided to the generator.

We similarly ask an automatic evaluator to provide absolute ratings for an adapted response (on a scale of 1-5) for response relevance for the corresponding value of each contextual attribute. That is, in our example above, we would obtain the rating for if the contextual attribute value were “*Complete beginner*”, “*Intermediate*”, “*Expert*”, and so on, resulting in many ratings per follow-up QA. We then compute the maximum difference of ratings across all values of a contextual attribute and plot these differences in Figure 6. Larger differences indicate worse ability to adapt equally to all possible values of a contextual attribute. For example, for the follow-up question above, if the ratings to model responses were all “4”, then the difference would be zero, meaning a model is equally adept at adapting to all expertise levels. But if the responses were “5”, “5”, “3”, “3”, and “2”, then the difference would be three, which indicates existence of a failure to adapt equally well to possible values of this contextual attribute. These results also use GPT-4o as the candidate model, and Gemini-1.5-Pro as the evaluator.

7.2 Findings

First, we note that there is considerable disparity in the ability of GPT-4o to cater to different values of a contextual attribute. This suggests that the model is not equally good at adapting to all possible values of a contextual attribute. This disparity is larger for attributes such as length and level of detail, as well as gender and age group. The higher disparity in length and detail suggests that the model is not consistently able to adjust its responses based on the level of depth or brevity required by the user, often delivering overly simplistic or verbose responses. On the other hand, the differential performance of the model for attributes such as gender and age group can lead to unequal user experiences. Going forward, we suggest that future model evaluations conduct granular analysis of context adaptability to better capture the full spectrum of a model’s capabilities.

8 Related Work

Context as clarification questions. A substantial amount of prior work has proposed methods to automatically generate clarification questions for tasks such as question answering (Rao and Daumé III, 2018; Yu et al., 2020; Kumar and Black, 2020; White et al., 2021; Zhang and Choi, 2023; Zhang et al., 2024; Andukuri et al., 2024), information retrieval (Zamani et al., 2020; Chi et al., 2024), open-domain dialog generation (Aliannejadi et al., 2019; Testoni and Fernández, 2024) and moral reasoning (Pyatkin et al., 2023). Much of this prior work has focused on generating questions about ambiguous or incomplete inputs (Majumder et al., 2021; Zhang and Choi, 2023). While we also represent contexts as question-answer pairs, the intent of our questions extends beyond resolution of ambiguity, where they often just seek further exposition about the user’s query.

Instruction-following and personalization. Our study of context and its role in evaluation is relevant to prior work evaluating instruction-following and personalization abilities of models. The constraints represented through our context can be perceived as writing instructions or user attributes. Evaluations such as Zhou et al. (2023); Pham et al. (2024) have been useful for measuring instruction-following abilities of models. The contexts considered in our work differ from instructions in existing work, as our context is intended to increase the amount of

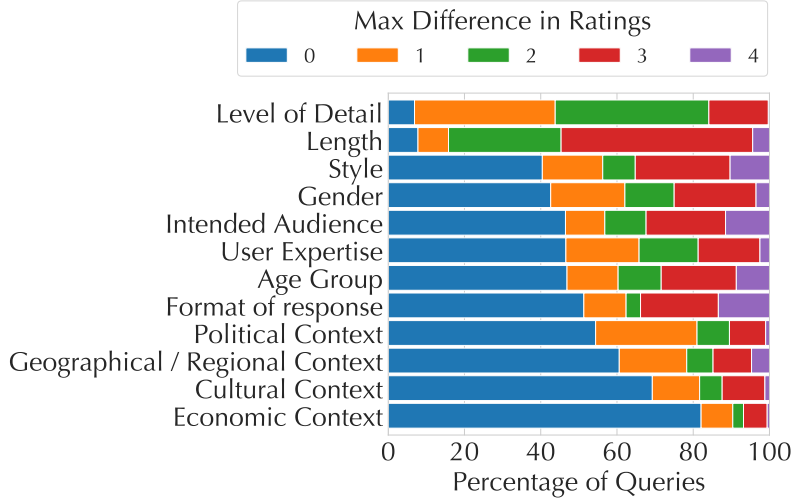


Figure 6: Distribution of the maximum difference in ratings between all values for each contextual attribute. Contextual attributes that are more **difficult** for models to adapt to diverse values will have high percentage of Red (3) and Purple (4), while contextual attributes that are **easy** for models to adapt to will exhibit high percentage of Blue (0). We find that several contextual attributes show considerable disparity in models’ ability to adapt their responses to different values of a contextual attribute.

specification for underspecified queries for better evaluation, rather than probing model abilities to follow writing instructions.

Work on personalization has advocated for evaluating the adaptability of model responses to users’ personalized contexts (Flek, 2020; Salemi et al., 2024) and proposed training methods for better alignment with user contexts (Lee et al., 2023; Cheng et al., 2023; Siththaranjan et al., 2024; Jang et al., 2023; Hwang et al., 2023; Pitis et al., 2024). We concur that evaluating adaptability of models to diverse contexts is important and our work presents a framework to conduct such an evaluation.

LM-based Autoraters. Language model-based autoraters (also called LM-as-judge) hold great promise in improving the efficiency and reducing the cost of evaluations (Chiang and Lee, 2023). Prior work has identified that these autoraters are subject to biases such as preferring longer responses (Dubois et al., 2024), preferring their own responses (Panickssery et al., 2024) and others (Zheng et al., 2023; Shen et al., 2023; Wang et al., 2024). Methods have been proposed to overcome some of these biases by providing clearer criteria during evaluation (Liu et al., 2023; Saha et al., 2024). Our work proposes a simple recipe for improving the reliability of LM-based autoraters by providing context surrounding queries.

9 Limitations

Scope of Context. The contextual attributes considered in our work only represent a sample of common information types that are missing in queries, which can be helpful to respond to a user’s query. The taxonomy presented in Table 2 is not meant to be exhaustive and there are likely other types of context that are important.

Effect of Amount of Context. In our work, we only consider up to 10 follow-up QAs for each query. We did not analyze the effect of the number of follow-up QAs on the results presented in section 4. In future work, it would be worthwhile to analyze how our results vary with the amount of specification in the context.

Autorater Reliance. Language model-based autoraters improve the efficiency and decrease the cost of evaluation. Hence, we relied primarily on these autoraters for a subset of our analysis (presented in §6 and §7). Human evaluation will strengthen the conclusions made in these analyses, and we will consider this in future work.

10 Conclusion

Through our study, we showed that existing query datasets contain many underspecified queries, and this underspecification can have significant impacts on the nature of our evaluation. We present **contextualized evaluations**, a simple, plug-and-play

recipe to enrich queries with context in the form of follow-up question-answer pairs. Our experiments show that context can change the conclusions we draw from evaluations, and decrease the extent to which evaluators rely on surface-level criteria to judge responses. Finally, we show that context can help identify implicit biases in default model responses. We suggest future work to perform contextualized evaluations for a holistic understanding of how well models adapt to diverse user contexts.

11 Acknowledgments

First, we thank all the human annotators who took time to participate in our studies. We also thank the following people for helpful discussions and comments: Sihao Chen, Doug Downey, Greg Durrett, Varsha Kishore, Nelson Liu, Nicholas Lourie, Vishakh Padmakumar, Valentina Pyatkin, Elizabeth Sieber, Luca Soldaini, David Wadden, Dan Weld, Michael Zhang and members of the Semantic Scholar research group at Ai2. Finally, we are grateful to the ACL reviewers and our action editor Omri Abend for their thoughtful comments and suggestions.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. [Asking clarifying questions in open-domain information-seeking conversations](#). In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.
- Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. [STarGATE: Teaching language models to ask clarifying questions](#). In *First Conference on Language Modeling*.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#). [Accessed on March 4, 2024].
- Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. [Everyone deserves a reward: Learning customized human preferences](#). *arXiv preprint arXiv:2309.03126*.
- Yizhou Chi, Jessy Lin, Kevin Lin, and Dan Klein. 2024. [Clarinet: Augmenting language models to ask clarification questions for retrieval](#). *arXiv preprint arXiv:2405.15784*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#).
- Charles LA Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. [An effectiveness measure for ambiguous and underspecified queries](#). In *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings 2*, pages 188–199. Springer.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in chatgpt: Analyzing persona-assigned language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. [Length-controlled alpaca-eval: A simple debiasing of automatic evaluators](#). In *First Conference on Language Modeling*.

- Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askeel, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling*.
- Lucie Flek. 2020. [Returning the N to NLP: Towards contextually personalized classification models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.
- Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. [The weirdest people in the world?](#) *Behavioral and brain sciences*, 33(2-3):61–83.
- Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. 2024. [On overcoming miscalibrated conversational priors in llm-based chatbots](#). *arXiv preprint arXiv:2406.01633*.
- EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. [Aligning language models to user opinions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.
- Team Jamba, Barak Lenz, Alan Arazzi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, et al. 2024. [Jamba-1.5: Hybrid transformer-mamba models at scale](#). *arXiv preprint arXiv:2408.12570*.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. [Personalized soups](#). [Personalized large language model alignment via post-hoc parameter merging](#). *arXiv preprint arXiv:2310.11564*.
- Kimiya Keyvan and Jimmy Xiangji Huang. 2022. [How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges](#). *ACM Computing Surveys*, 55(6):1–40.
- Vaibhav Kumar and Alan W Black. 2020. [ClarQ: A large-scale and diverse dataset for clarification question generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.
- Yoonho Lee, Huaxiu Yao, and Chelsea Finn. 2023. [Diversify and disambiguate: Learning from underspecified data](#). *ICLR*.
- Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. [UNQOVER: ing stereotyping biases via underspecified questions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [AlpacaEval: An automatic evaluator of instruction-following models](#). https://github.com/tatsu-lab/alpaca_eval.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. [Ask what’s missing and what’s useful: Improving clarification question generation using global knowledge](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

- Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. [ExpertQA: Expert-curated questions and attributed answers](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.
- Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). *arXiv preprint arXiv:2404.13076*.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. [Disentangling length from quality in direct preference optimization](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4998–5017, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. [Suri: Multi-constraint instruction following for long-form text generation](#). In *Findings of EMNLP*.
- Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. 2024. [Improving context-aware preference modeling for language models](#). *arXiv preprint arXiv:2407.14916*.
- Valentina Pyatkin, Jena D. Hwang, Vivek Sriku-mar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. [ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.
- Sudha Rao and Hal Daumé III. 2018. [Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazari-dou, Orhan Firat, Julian Schrittwieser, et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. [Branch-solve-merge improves large language model evaluation and generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico. Association for Computational Linguistics.
- Alireza Salemi, Sheshera Mysore, Michael Bender-sky, and Hamed Zamani. 2024. [LaMP: When large language models meet personalization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. [Large language models are not yet human-level evaluators for abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.
- Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. [Distributional preference learning: Understanding and accounting for hidden context in rlhf](#). In *ICLR*.
- Karen Spärck-Jones, Stephen E. Robertson, and Mark Sanderson. 2007. [Ambiguous requests: implications for retrieval tests, systems and theories](#). *SIGIR Forum*, 41(2):8–17.
- Alberto Testoni and Raquel Fernández. 2024. [Asking the right question at the right time: Human](#)

- and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–275, St. Julian’s, Malta. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. 2021. [Open-domain clarification question generation without question examples](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 563–570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. [KIWI: A dataset of knowledge-intensive writing instructions for answering research questions](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12969–12990, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.
- Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi. 2020. [Interactive classification by asking informative questions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2664–2680, Online. Association for Computational Linguistics.
- Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. [Generating clarifying questions for information retrieval](#). In *Proceedings of the web conference 2020*, pages 418–428.
- Michael JQ Zhang and Eunsol Choi. 2023. [Clarify when necessary: Resolving ambiguity through interaction with lms](#). *arXiv preprint arXiv:2311.09469*.
- Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. 2024. [Modeling future conversation turns to teach llms to ask clarifying questions](#). *arXiv preprint arXiv:2410.13788*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *arXiv preprint arXiv:2311.07911*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. [Fine-tuning language models from human preferences](#). *arXiv preprint arXiv:1909.08593*.

A Human Evaluation Details

Participants. We recruited a total of 1,085 participants through the Prolific crowdsourcing platform for the human evaluation studies. Evaluators were required to have an approval rate of 99% and 500 prior successful submissions on the platform. They were also required to be fluent in English. Most participants came from the US and UK.

Evaluation Setup. For the main preference judgment task, we sampled a query and evaluation setting (from the three settings discussed in 4) randomly for each evaluator. Evaluators were also given a choice to skip a query for a different one, if they were unfamiliar with the topic of the original query. Each query-response pair was judged by 3 different evaluators and a single evaluator was allowed to complete up to 3 examples. Before starting the task, evaluators were required to complete a training task in the context-agnostic and context-aware evaluation settings, where they were given feedback about their ratings. Evaluators were paid \$1.75 for each example at a rate of \$15 per hour, where they were allocated a total of 7 minutes per example.

Interface. Screenshots of the evaluation interfaces for context-agnostic and context-aware evaluation are provided in Figures 7 and 8 respectively.

B Experimental Details

Models. We used 3 pairs of models for pairwise evaluations: GPT-4o (Achiam et al., 2023) and Gemini-1.5-Flash (Reid et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024) and Llama-3.1-405B (Dubey et al., 2024), Gemma-2-27B (Gemma et al., 2024) and Jamba-1.5-Large (Jamba et al., 2024). We list the identifiers of the models that were used for experiments in Table 7. All models were called through the organization’s official APIs (except for Llama, for which we used the Together API).

Hyperparameter Details. In all generation tasks, the temperature was set to the default value in the organization’s API. For response generation and context generation, we sampled a maximum of 2048 tokens while for model evaluation, we sampled a maximum of 512 tokens.

Prompts. The prompt used for classifying queries into query types is provided in Table 9. The prompt for generating follow-up QAs for a query is

provided in Table 10. For getting autorater judgments without and with context, the prompts are provided in Tables 11 and 12 respectively. Finally, the prompt to compute the number of constraints from the context met by a response is provided in Table 13. For the analyses presented in sections 6 and 7, we provide the prompt used to filter queries in Table 14, and the prompt used to rate responses on a Likert scale is in Table 15.

Model Name	Identifier
GPT-4o	gpt-4o-2024-05-13
Gemini-1.5-Flash	gemini-1.5-flash-exp-0827
Claude-3.5-Sonnet	claude-3.5-sonnet
Llama-3.1-405B	Meta-Llama-3.1-405B-Instruct-Turbo
Gemma-2-27B	gemma-2-27b-it
Jamba-1.5-Large	jamba-1.5-large
Qwen2-72B-Instruct	Qwen2-72B-Instruct

Table 7: List of models used in our experiments and their official identifiers.

Contextual Attribute	Follow-Up QA
Level of Detail	Q: How much detail do you prefer in the response? A: ["One-sentence answer", "Key points only", "Moderate detailed", "Extensive detail"]
User Expertise	Q: What is your level of expertise on this topic? A: ["Complete beginner", "Basic understanding", "Intermediate", "Advanced", "Expert"]
Length	Q: What is your preferred length for the response? A: ["One sentence", "2-3 sentences", "One paragraph (>3 sentences)", "Several paragraphs"]
Format of response	Q: What format would you prefer the response to be in? A: ["Bulleted list", "Numbered steps", "Paragraph text", "Table or chart"]
Style	Q: What style of response do you prefer? A: ["Formal", "Informal", "Conversational", "Academic", "Technical"]
Intended Audience	Q: Who is the intended audience for this response? A: ["General public", "Children", "Students", "Professionals / Experts"]
Geographical / Regional Context	Q: What region or country should this response be based on? A: ["North America", "Europe", "Asia", "Africa", "Latin America"]
Cultural Context	Q: What cultural perspective should be considered in the response? A: ["Western culture", "Eastern culture", "Indigenous culture", "Multicultural perspective"]
Age Group	Q: Which age group should this response be relevant for? A: ["Children", "Teenagers", "Young adults", "Middle-aged adults", "Seniors"]
Economic Context	Q: What economic situation should this response be relevant for? A: ["Low-income", "Middle-income", "High-income", "Budget-conscious"]
Political Context	Q: What political context should this response consider? A: ["Liberal", "Conservative", "Centrist", "Socialist"]
Gender	Q: Should the response consider any specific gender perspective? A: ["Male", "Female", "Non-binary", "Gender-neutral"]

Table 8: Contextual attributes and corresponding follow-up QA considered for the implicit context analysis presented in Sections 6 and 7.

Prompt for Classifying Query Types
<p>You will be shown a query issued by a real user to a language model. You need to answer what query type(s) this query belongs to, from the list below.</p> <ul style="list-style-type: none"> - Ambiguous: Queries which can be interpreted in different ways, that cause confusion about what is being asked. - Incomplete: Queries which lack information that is essential to understand the intent of the query. Note these are different from ambiguous queries, which need clarification due to multiple possible interpretations. - Subjective: Queries whose responses can be influenced by personal beliefs and perspectives. - Open-ended: Queries which require detailed responses and lack a single, concise answer. - Closed-ended: Queries which require an unambiguous and concise answer. <p>Note that a single query can belong to multiple query types. Provide your output as a list with the query types that the query belongs to.</p> <p>###</p> <p>Query: best team in the league Query Types: ["Incomplete", "Subjective", "Closed-ended"] Query: [QUERY] Query Types:</p>

Table 9: Prompt for classifying queries into different query types.

Prompt for Generating Follow-up QAs

You will be shown a query issued by a real user to a language model. Imagine that you are required to answer this query. First, you need to answer whether it would be helpful to know context surrounding this query to give a useful response. The context can be about the user (eg, their background, age, language fluency, location, profession, expertise etc), their intent / preferences for the response (eg, query intent, text formatting/style, structure, length, presence of citations, or any other open-ended criteria) or information missing that is required to respond to a query or resolve ambiguity in the query. Queries that are objective, closed-ended or have straightforward answers should not require context.

Answer in Yes or No for whether context is required and generate context if the answer is Yes. This context should be formatted as follow-up question answer pairs, where you ask the most important questions first and list plausible answers to these questions.

Here are criteria that individual questions need to satisfy:

- salient: The question should ask about information that would be useful to adapt the query's response to the user's needs and background.
- influential: The answer to this question should directly influence the response. With different answers to this question, the response to the query would need to be phrased differently.

Here are the criteria that the list of questions needs to satisfy:

- sufficient: There should be enough important questions to cover a large space of possible contexts for the query.
- ranked in order of salience: the questions should be ranked in the order of their importance.

Here are the criteria that each answer set needs to satisfy:

- plausible answers: The answer set should represent a realistic set of answers to the question, such that a real user would answer the question with any of the choices. Do not generate answer choices such as "Other" which are uninformative.
- discrete answer space: The possible answers to the question should be discrete, short strings.
- diverse coverage: The answer set should be a representative set of possible answers to the question, such that each answer choice would elicit different responses to the original query.

Generate up to 10 follow-up QA pairs and they should all meet the above criteria. Each QA pair should be such that it is easy to check whether the QA is incorporated in a candidate response.

Example Follow-up QA:

Query: best team in the football league

Need for Context: Yes

Context: Q: Which league are you referring to? A: ["English Premier League", "La Liga", "Bundesliga", "Italian Serie A", "MLS", "UEFA"]

Q: How do you define "best"? A: ["Most recent wins", "Number of championships won", "Goal difference", "Squad strength"]

Q: Do you want the best team based on current form or overall historical performance? A: ["Current form", "Historical performance"]

Q: Are you asking about men's football or women's football? A: ["Men's football", "Women's football"] ...

Query: How do antibiotics work against bacteria?

Need for Context: Yes

Context: Q: What is your background in biology or medicine? A: ["No background", "High school level", "College level", "Medical or professional background"]

Q: What is your purpose for asking this question? A: ["For a class", "Personal knowledge", "Professional/medical use", "To explain to someone else"]

Q: What level of detail are you looking for in the explanation? A: ["Basic overview", "Intermediate (some scientific terms)", "Detailed (in-depth biological mechanisms)"] ...

Table 10: Prompt for generating contextual follow-up questions for user queries.

Prompt for Autorater Preference Judgements (setting NoCtxGen-NoCtxEval)

You will be given a query issued by a real user to a language model. You will also be given two model responses to this query, and you will need to judge which response is better.

IMPORTANT: You should produce the final judgement as a dictionary in precisely this format (with **): `**output: {"judgement": " _ " }**`, where you should fill in the spaces with either "Response 1" if Response 1 is better, "Response 2" if Response 2 is better or "Tie" if both responses are equally good or equally bad. Only the three choices "Response 1", "Response 2" and "Tie" are valid. Make note of the ** required to enclose the output dictionary. After generating the output, provide a brief justification of your judgement.

Query: [QUERY]

Response 1: [RESPONSE 1]

Response 2: [RESPONSE 2]

Judgement: `**output: {"judgement": " _ " }**`

Justification: [JUSTIFICATION]

Table 11: Prompt for getting autorater preference judgements for the setting NoCtxGen-NoCtxEval.

Prompt for Autorater Preference Judgements (settings NoCtxGen-CtxEval and CtxGen-CtxEval)

You will be given a query issued by a real user to a language model and the context under which the query was issued. This context will be presented in the form of follow-up questions and the user's answers to these questions. The context provides information about the user's intent, preferences and background.

You will be given two model responses to this query, and you will need to judge which response more accurately and completely incorporates the information from the query and context. To evaluate the responses, you should first check whether the answer to each of the follow-up questions in the context is incorporated well in each response. Then, you should choose the response which incorporates more of the constraints from the context and provides the most relevant and complete answer to the query.

IMPORTANT: You should produce the final judgement as a dictionary in precisely this format (with **): `**output: {"judgement": " _ " }**`, where you should fill in the spaces with 1) "Response 1" if Response 1 is better, 2) "Response 2" if Response 2 is better or 3) "Tie" if both responses are equally good or equally bad. Only the three choices "Response 1", "Response 2" and "Tie" are valid. Make note of the ** required to enclose the output dictionary. After generating the output, provide a brief justification of your judgement that mentions which aspects of the context were better incorporated by the chosen response, or why the responses are equally good or equally lacking.

Query: [QUERY]

Context: [CONTEXT]

Response 1: [RESPONSE 1]

Response 2: [RESPONSE 2]

Judgement: `**output: {"judgement": " _ " }**`

Justification: [JUSTIFICATION]

Table 12: Prompt for getting autorater preference judgements for the context-aware evaluation settings (NoCtxGen-CtxEval and CtxGen-CtxEval).

Prompt for computing the number of constraints in a context that are met by a response

You will be given a query issued by a real user and the context under which the query was issued. This context will be presented in the form of follow-up questions and the user's answers to them.

You will be given a model response to this query, and you will need to judge how many of the criteria in the follow-up questions are addressed by the response. So if the response incorporates 5 of the follow-up questions completely, you should output 5. If it incorporates 2 of the follow-up questions, you should output a 2. If it does not address any of the follow-up questions, you should rate it as a 0.

IMPORTANT: You should first generate a single number, which is the total number of constraints satisfied. After generating this number, provide a very brief justification for your answer.

Query: [QUERY]
Context: [CONTEXT]
Response: [RESPONSE]
Output:

Table 13: Prompt for evaluating how many follow-up QAs in the context are satisfied by a response.

Prompt for checking the importance of a contextual attribute for a query

You will be given a query from a real user to a language model, along with a follow-up question that can be asked to the user. The follow-up question will have a set of answer choices. Your task is to answer the following three questions:

- 1) Is it important to know the user's answer to the follow-up question to provide a useful response to the original query?
- 2) Is the query independent of the answer choices? If the query already implies a specific answer choice, it is not independent.
- 3) Is the query well-formed? A well-formed query clearly expresses an information need, even if it is not fully fluent, unambiguous, or fully specified. Queries not in English are not considered well-formed.

IMPORTANT: Please provide the final output in the following dictionary format: {"1": "Yes/No", "2": "Yes/No", "3": "Yes/No"}.

Query: [QUERY]
Follow-up Question: [QUESTION]
Output:

Table 14: Prompt for evaluating the importance, independence, and well-formedness of queries with follow-up questions.

Prompt for rating response relevance based on query and context

You will be given a query issued by a real user to a language model and the context under which the query may have been issued. This context will be presented in the form of a follow-up question issued to the user and possible answers to this question.

You will be given a model response to this query, and you will need to judge the quality of this response corresponding to each follow-up question-answer pair. Rate the response on a scale of 1-5 on the following axis:

* Relevance: How relevant is the response to addressing the query and context?

- * 1: The response is not helpful in responding to the query and context at all.
- * 2: The response provides limited help, missing important information from the query or context.
- * 3: The response is somewhat helpful, offering useful information but lacking thoroughness or depth for the query and context.
- * 4: The response is helpful, addressing most of the query and context adequately.
- * 5: The response is highly helpful, fully addressing the query and context with thorough and useful information.

IMPORTANT: You should produce the final output as a dictionary in precisely this format (with **): [OUTPUT_FORMAT], where you should fill in the spaces with ratings for each one of the possible answers to the follow-up question. Make note of the ** required to enclose the output dictionary.

Query: [QUERY]

Context: [CONTEXT]

Response: [RESPONSE]

Judgement:

Table 15: Prompt for rating response relevance based on query and context.

