

Eliciting Human Preferences with Language Models

Anonymous ACL submission

Abstract

Language models (LMs) can be directed to perform target tasks by using labeled examples or natural language prompts. But selecting examples or writing prompts can be challenging—especially in tasks that involve unusual edge cases, demand precise articulation of nebulous preferences, or require an accurate mental model of LM behavior. We propose to use *LMs themselves* to guide the task specification process. In this paper, we introduce **generative active task elicitation** (GATE): a learning framework in which models elicit and infer intended behavior through free-form, language-based interaction with users. We study GATE in three domains: email validation, content recommendation, and moral reasoning, corresponding to three common scenarios where task specification is hard. In preregistered experiments, we show that LMs prompted to perform GATE (e.g., by generating open-ended questions or synthesizing informative edge cases) elicit responses that are often more informative than user-written prompts or labels. Users report that interactive task elicitation requires less effort than prompting or example labeling and surfaces considerations that they did not anticipate on their own. Our findings suggest that LM-driven elicitation can be a powerful tool for aligning models to complex human preferences and values.

1 Introduction

The complexity of human preferences makes them challenging to encode in machine learning systems. Consider the problem of designing a recommendation system for songs or websites: first, system builders must develop a formal model of the potential factors influencing user preferences; second, users must describe their preferences in a format that a learning algorithm can use to make

future recommendations. Each of these steps requires mental effort and continual refinement by users and system builders. Until recently, the dominant approach in machine learning has specified preferences using *examples*: users first label a dataset with examples of the desired model behavior, then train a machine learning model on this dataset. This strategy has seen widespread use across diverse tasks, including image classification and question answering (Krizhevsky et al., 2012; Devlin et al., 2019). In more recent years, this paradigm has changed with the advent of *instruction following* methods (Brown et al., 2020a): by pre-training language models (LMs) on large-scale text corpora, it is possible to induce desired behaviors by conditioning only on natural language task specifications, in tasks as diverse as code generation and text summarization.

However, this progress has also accentuated the challenges described above: complex behaviors require an increasing amount of *prompt engineering* or *dataset design* to overcome the imprecision of natural language and prevent models from misunderstanding or misgeneralizing from spurious features of prompts or examples. For example, a user who says they enjoy reading tennis articles could either be interested in the competitive tennis circuit or in improving their own serve. A few user-provided examples of tennis-related articles might fail to specify whether the user is interested in broader tennis content, such as tennis-themed satire. These challenges of *task ambiguity* (Finn et al., 2018; Tamkin et al., 2022a) loom large as models continue to be applied to more open-ended tasks and higher-stakes domains.

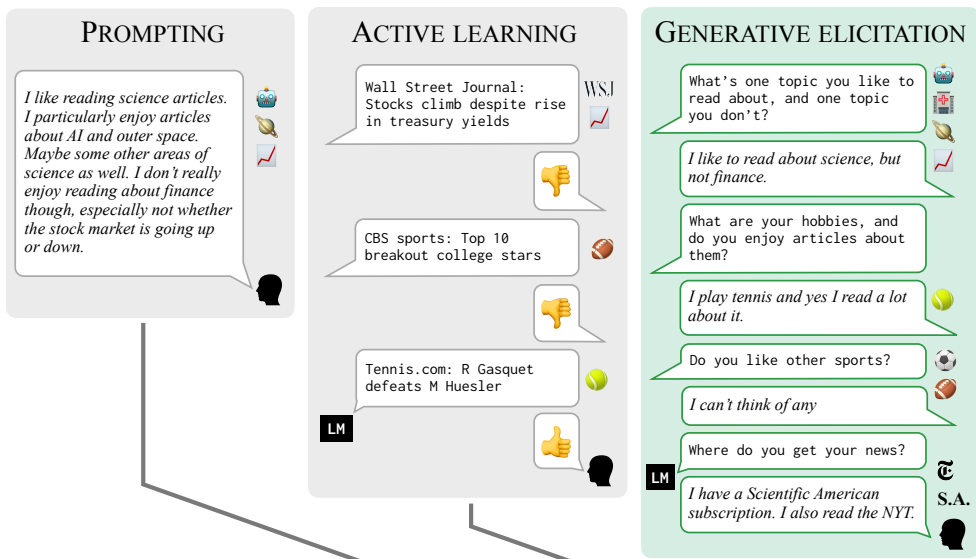
To address these challenges, we propose to use *models themselves* to help convert human preferences into automated decision-making systems. In this paper, we introduce **generative active task elicitation** (GATE), a learning framework in which models elicit and infer user preferences through

041
042
043
044
045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081

(A) Unknown user preferences



(B) Task elicitation



(C) Evaluation on test cases

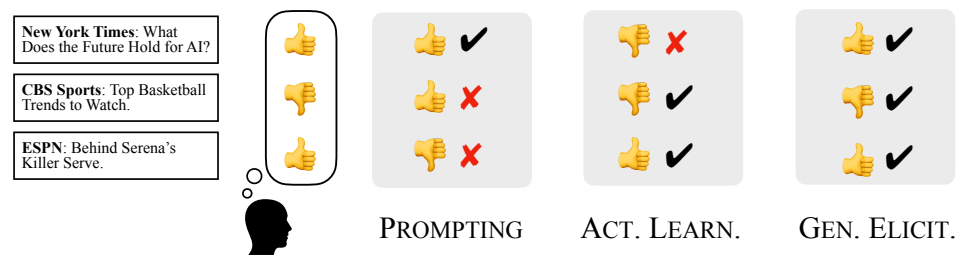


Figure 1: **Generative Active Task Elicitation (GATE)** elicits user preferences through interactive, free-form questions, which can then be used in downstream decision-making. Unlike non-interactive elicitation approaches (e.g., prompting), which rely entirely on the human to elucidate their preferences, generative elicitation is better able to probe nuances of human preferences. Unlike active learning approaches, generative elicitation can ask more generic, free-form questions. The three parts of this figure illustrate: **(A) Fuzzy user preferences:** A user wishes to translate their fuzzy preferences for how a task should be performed into a specification for a machine learning model. This is challenging because users lack perfect introspection, preferences can be difficult to specify in language, the specification needs to anticipate tricky real-world edge cases, and models may misgeneralize from provided examples or instructions. **(B) Task elicitation:** We consider various ways of eliciting these fuzzy preferences from users, including non-interactive prompting, active learning, and generative elicitation (GATE). **(C) Evaluation:** We evaluate methods on a held-out test set, scoring how well a language model predicted the true decisions made by the user.

open-ended interaction. We describe several techniques for leveraging LMs to perform GATE—for example, by asking informative open-ended questions or generating edge cases for users to label. We then evaluate these methods in three domains where task specification is hard. People can often fail to articulate their personal preferences, their moral values, or struggle to formulate task that requires expert knowledge. Thus, we study the problems of content recommendation, moral reasoning, and email validation respectively.¹ In pre-registered experiments, we find that LM-based task elicitation often yields more accurate models than existing prompting or active learning techniques while requiring comparable (or less) mental effort from users and surfacing novel considerations.

In summary, this paper introduces a new learning framework (GATE), a family of methods that perform GATE using pre-trained language models, and experimental evidence showing that these methods outperform existing prompting and labeling methods. Our results show that interactive, language-based task elicitation is a flexible and powerful tool for building personalized models, capable of overcoming many challenges inherent in prompt- and example-based methods.

2 Learning as Task Elicitation

2.1 The Task Elicitation Framework

We study the problem of efficiently training a machine learning model to perform a task of interest. Throughout this paper, we use **task** to refer generically to any function $f : x \mapsto y$ that maps inputs x to outputs y . When building a personalized website recommendation system, for example, x are websites and y are user preference scores for that website. Because different users may prefer different content, each user’s individual preferences specify a distinct task: *content recommendation for Pat* and *content recommendation for Avery* are different tasks within the **domain** of content recommendation (Ziegler et al., 2020). To build such a model, we must collect some **task specification** from a human user (e.g., revealing what websites they are interested in). As noted above, current learning approaches admit a wide variety of speci-

¹While this paper focuses on language-based elicitation procedures, we note that generative active task elicitation is modality-agnostic and could be applied to other settings (e.g., speech-based or multimodal models).

fication types, including collections of labeled examples, natural language instructions, or combinations of the two. What makes one type of specification preferable to another? Ideally, we would like specifications that are both (1) easy for humans to create and (2) informative to learners, enabling them to model human preferences accurately. Abstractly, we seek a framework for gathering and learning from specifications that optimizes an objective:

$$\alpha \cdot \text{specification cost} + \beta \cdot \text{human-predictor alignment} \quad (1)$$

where **specification cost** measures human time and mental effort, **human-predictor alignment** measures the extent to which model choices agree with choices the human would have made, and α and β tradeoff between the two. To formalize this, let \mathcal{H}_f denote a human user whose preferences are represented by a function f . We wish to design an **elicitation policy** \mathcal{E} that interacts with \mathcal{H}_f to produce a **task specification** s . This specification may then be input to a learning algorithm to produce a model $\hat{f}(s)$. Then, letting $C(\cdot)$ denote a scalar measure of specification cost, and $A(\cdot, \cdot)$ denote a measure of alignment between two predictors, we wish to minimize (in expectation over the population of human users):

$$\mathbb{E}_{\mathcal{H}_f} \mathbb{E}_{s \sim \mathcal{E}(\mathcal{H}_f)} [\alpha \cdot C(s) + \beta \cdot A(f, \hat{f}(s))] \quad (2)$$

Here, C might measure the number of words the user typed to produce the specification s , while A might measure model-predictor agreement at the level of individual predictions from some population: $A(f, \hat{f}) = \mathbb{E}_x \|f(x) - \hat{f}(x)\|$. In general, appropriate definitions of C and A are domain-dependent; in this paper, our experiments compare the alignment of different predictors at a fixed cost. Evaluation of cost, alignment, and tradeoffs between them are discussed more in Section 5.

2.2 Existing Learning Paradigms in the Task Elicitation Framework

Several existing frameworks for learning and task specification can be described within the framework given above. Understood as task elicitation procedures, existing frameworks differ along two key axes (visualized in Figure 2): their level of *interactivity* and their level of *flexibility*. In interactive elicitation methods, queries can change depending on user responses (e.g., querying for the

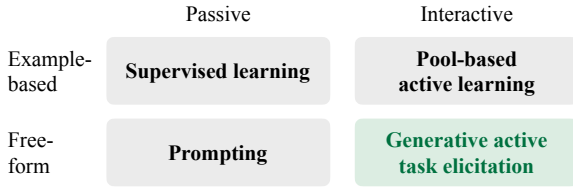


Figure 2: Axes of variation in task elicitation.

most useful information based on what is known thus far) while passive elicitation methods expect the user to provide specifications in a single shot. Example-based specification methods ask users to label a set of examples, while free-form elicitation approaches are less restrictive, allowing the user to provide a much wider range of inputs, including natural language instructions and explanations.

Supervised learning: passive, example-based

In the most common supervised learning setup, the elicitation policy \mathcal{E} simply instructs the human user \mathcal{H}_f to generate a collection of labeled (input, output) pairs, after which $\hat{f}(s)$ is produced by fitting or fine-tuning a learned model using standard algorithms. This is an *example-based* process because the specification is provided via labeled examples and is *passive*, as the model does not interactively query the user to label additional data.

Active learning: interactive, example-based

In active learning, the elicitation policy is interactive. Users first assemble a fixed pool of unlabeled inputs x . Next, \mathcal{E} , selects from this pool an example whose label would be most informative. The user \mathcal{H}_f provides a label for this example, then \mathcal{E} selects the next-most-informative example, and so on (Cohn et al., 1994; Dagan and Engelson, 1995; Lewis and Gale, 1994; Settles, 2009). Finally, $\hat{f}(s)$ is trained as in supervised methods. Optimal experiment design methods (Emery and Naranokomov, 1998) may be viewed as generalizations of this paradigm in which inputs x are generated rather than selected. *Interactive* processes enable the model to query for examples that may resolve uncertainty or ambiguity in the task specification (Tamkin et al., 2022b).

Prompting: passive, free-form Modern pre-trained models allow for specifying tasks in more flexible ways than simply labeling examples. For example, models can be conditioned with a *prompt* describing the user’s intended task in natural language (Brown et al., 2020b), or even a mix of

language and image inputs (Alayrac et al., 2022). As with supervised learning, the labeling policy \mathcal{E} here is simply an instruction to write a natural language task description (s), but the final predictor $\hat{f}(s)$ is produced by passing s to a pre-trained language model.

3 Generative Active Task Elicitation

All of the methods above have important drawbacks: the burden typically falls upon the user to ensure that prompts or example sets are truly comprehensive specifications of the task, as any lack of clarity in the prompt could lead to task ambiguity (Tamkin et al., 2022a), resulting in undesired behavior during deployment. Resolving task ambiguity by crafting better prompts is challenging and time-consuming due to the difficulties of articulating nebulous personal preferences and anticipating edge cases that will emerge during deployment time.

However, one quadrant of Fig. 2 is not occupied by any of the aforementioned approaches: there is currently no method that leverages the flexibility of a free-form specification while using interaction to resolve uncertainty. We explore whether it is possible to combine the flexibility and richness of prompting-based specifications with the advantages of interactive methods such as active learning, by having a model interactively query users for these rich specifications. We term this family of methods **generative active task elicitation (GATE)**.

3.1 Methods for GATE

The effectiveness of language models (LMs) for understanding and producing free-form text suggests that they may be capable of eliciting and understanding user preferences. In this paper, we thus experiment with a family of GATE methods in which LMs serve as the backbone for both the elicitation policy \mathcal{E} and the predictor $\hat{f}(s)$. See Figure 1 for examples. In particular, we implement the elicitation policy \mathcal{E} by prompting an LM to ask the user questions while conditioning on the history of previous questions and answers. To make predictions $\hat{f}(s)$, an LM is prompted to predict a label conditioned on an input x and a complete elicitation transcript s provided as input. We experiment with several different information gathering policies, realized by simply prompting an LM to ask different kinds of questions:

Generative active learning The LM generates examples for the user to label. This approach has the advantage of providing concrete scenarios to the user, including some they may not have considered a priori. For example, for content recommendation, the LM might generate an article such as: *Are you interested in the following article? The Art of Fusion Cuisine: Mixing Cultures and Flavors [...]*.

Generative yes-or-no questions We restrict the LM to generating binary yes-or-no questions. This approach enables the model to elicit more abstract preferences while still being easy for the user to answer. For example, the model might probe a user’s preferences by asking: *Do you enjoy reading articles about health and wellness?*

Generative open-ended questions The LM generates arbitrary questions requiring free-form natural language responses. This enables the LM to elicit the broadest and most abstract pieces of knowledge at the potential cost of being overly broad or challenging for the user to answer. For example, the LM might generate the question: *What hobbies or activities do you enjoy in your free time[...] and why do these hobbies or activities captivate you?*

The user is not constrained in their response in any of the above settings; they are free to provide as much detail as they want. We present example elicitation transcripts for each policy in Figure 8.

4 Experiment Setup

We consider tasks in three different domains to evaluate our generative active task elicitation methods. A common feature of these domains is that they do not feature a single correct behavior that could be learned during LM pre-training; instead, models must elicit an individual human’s preferences in order to make accurate predictions. We allow each human user to interact open-endedly with an elicitation policy \mathcal{E} for five minutes. Next, humans and learned models $\hat{f}(s)$ independently label a set of held-out examples. Finally, we measure agreement between humans and learned predictors. See Figure 8 for examples of environments and dialogues.²

²A link to the preregistration of experiments and analyses will be made available upon publication.

4.1 Domains and datasets

We consider three different domains corresponding to roughly three common scenarios where task specification is difficult.

Content Recommendation Users often fail to fully specify *personal preferences*. Thus, we consider the domain of online article recommendations, where user preferences vary widely. Models are evaluated on their ability to predict whether a user would like to read a given held-out article. These test cases are taken from popular online newspaper and magazine articles collected by the authors, from which we take the website name, article title, and a short description.

Moral Reasoning Moral preferences can be deeply personal and hard to articulate. Thus, as a test-bed for eliciting moral values, we consider the question of when (if ever) it is ethical to steal a loaf of bread. During evaluation, models are presented with textual descriptions of scenarios and asked to predict whether users will judge it appropriate to steal a loaf of bread. These test cases are constructed manually by the authors.

Email Verification Finally, specification is difficult in situations requiring specialized knowledge. For example, in software engineering, developers need to anticipate and account for many edge cases. In particular, we focus on specifying requirements for email address validation, where people have varied preferences over how long emails can be, how many subdomains they may possess, and which special characters are allowed, among other factors. Models are evaluated on their agreement with users about the validity of a set of held-out emails; this test set is again manually constructed by the authors.

4.2 Human interaction

Human participants in these experiments were recruited from English-speaking users of Prolific. For the email validation task, we additionally recruited participants from several computer science programs at US universities. We recruited 20–30 participants for each domain-method pair (6 elicitation methods across 3 domains), for a total of 388 participants. Participants were paid an average of \$12/hr. Our experiments received IRB approval, and all participants consented to having their data used for our experiments. The breakdown of the number of participants allocated to

each scenario and method can be found in Appendix B.1. Details of the user interface used in experiments may be found in Appendix B.2.

4.3 Modeling details

We use the GPT-4 model (gpt-4-0613 snapshot; OpenAI, 2023) to both elicit user preferences (as an elicitation policy \mathcal{E}) and make predictions based on the elicited preferences (as a predictor $\hat{f}(s)$). We additionally run experiments on Mixtral, an open-source LM, in Appendix C.3. To elicit user preferences, we prompt GPT-4 with a domain description and the current interaction history, and ask it to generate an informative but easy-to-answer edge case (for generative active learning) or question (for generative yes-or-no questions and generative open-ended questions). To make predictions, we prompt GPT-4 with the task specification s and a test sample x and ask it to generate a prediction for the test sample. The full text of the prompts can be found in Appendix A.

4.4 Baseline methods

We compare GATE with several baseline approaches for specifying tasks. Here, the elicitation policy \mathcal{E} is not parameterized by an LM, but constructed by the user or based on a pool of real examples.

Supervised learning We consider supervised learning as a baseline, as described in Section 2.2. We randomly present participants with questions from a large pool of examples and ask them to annotate up to the five-minute time limit. We study this approach exclusively in the content recommendation domain because pools of examples are not readily available in the other two domains. We use the Microsoft News Dataset (Wu et al., 2020) as our pool for this domain, a dataset of 160k news articles with descriptions. The license terms for research use of this dataset can be found at https://github.com/msnews/MIND/blob/master/MSRLicense_Data.pdf. We use the data consistent with the terms in “(1) Use Rights”.

Pool-based active learning As a baseline active learning approach, we consider a pool-based active learning approach, as described in Section 2.2. For the elicitation policy, we use the diversity-based sampling approach of Margatina et al. (2023); we first cluster the examples using a Sentence-BERT embedding model (Reimers and

Gurevych, 2019) into 15 different clusters, then iteratively ask questions from each cluster in a round-robin fashion, up until the five-minute time limit.³ This baseline is intended to capture the difficulty of selecting informative examples from a pool of unlabeled examples relative to generating informative examples from scratch. As with supervised learning, we study this approach exclusively in content recommendation.

User-written prompts As a baseline that does not use interactive elicitation, we ask participants to write a short paragraph describing their preferences for the task. We then use the text of this paragraph to prompt a LM to make decisions. This baseline is intended to capture the difficulty of specifying preferences in writing, both in terms of the effort it takes to write the paragraph and difficulty of writing a paragraph that fully specifies one’s preferences.

4.5 Evaluation and metrics

We measure how well models can predict the probability that users will answer questions a certain way, which we call $p(\text{correct})$. Specifically, we prompt the model with the interaction history s as a single test case, then ask the model to output a real-valued probability that a user would answer “yes” to the test case (e.g. the probability the user likes an article for content recommendation), which we call p_{LM} . This probability is outputted in token space as a number between 0.0 and 1.0, similar to past work (Branwen, 2020; Lin et al., 2022). The exact prompts we use for predicting probabilities can be found in Appendix A.2.

We define $p(\text{correct})$ as the probability the model assigns to the user-preferred answer. For example, if $p_{LM} = 0.8$ for a given question, then $p(\text{correct})$ would be 0.8 if the user’s answer were “yes” to the same question, and 0.2 if the user’s answer was “no”.

We use this metric instead of accuracy because we found modeling the uncertainty in (our estimate of) user’s preferences was a more informative metric than predicting exact user decisions. In pilot experiments prompting the LM to predict bi-

³Margatina et al. (2023) explored several different popular active learning sampling approaches for in-context learning (including random, uncertainty, and diversity sampling) and found little difference in empirical performance between them. We also ran exploratory model-model experiments in our domains and found no significant difference between these three sampling strategies. See details in Appendix D.

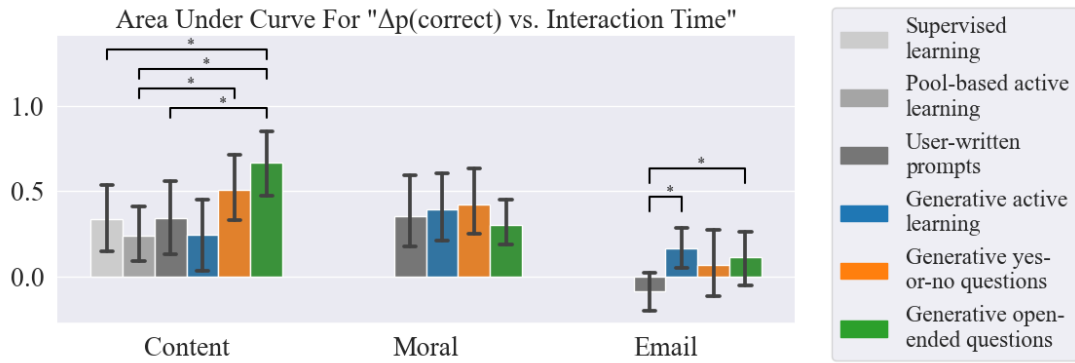


Figure 3: Across three domains, our LM-prompting implementations of GATE are generally able to elicit human preferences beyond baseline supervised learning, active learning, or human-written prompts. We measure the Area Under the “ $\Delta p(\text{correct})$ vs. Interaction time” Curve, which gives us a time-normalized metric for how well (and how quickly) each elicitation method is at aligning with human preferences. While GATE methods generally outperform the baseline methods as well as no interaction (represented by a $\Delta p(\text{correct})$ of 0), we are only able to establish statistical significance between GATE and baselines in the content recommendation and email verification domains. * indicates a statistically significant difference ($p < 0.05$).

nary yes/no decisions, we found this resulted in skewed predictions where the LM would predict one of ‘yes’ or ‘no’ for the entire test set, perhaps due to miscalibration of the model’s implicit decision threshold. Furthermore, at the time of writing, token probabilities for GPT-4 were not available via the OpenAI API. That said, we also discuss and report a classification-based metric in Appendix C.2.

Given $p(\text{correct})$, we compute:

Area under the $p(\text{correct})$ -time curve We do not just care about the total information elicited, but about *how quickly* good information is elicited. That is to say, if two methods arrived at the same $p(\text{correct})$ at the end of five minutes, we want to reward the method that arrived a higher $p(\text{correct})$ faster. To do this, we compute the average change in $p(\text{correct})$ after *every minute* of human elicitation time (conditioning on the state of the transcript at that time). This produces a curve where the x -axis is time, and the y -axis is the average change in $p(\text{correct})$. By taking the total area beneath this curve (AUC), we reward methods that arrive at higher $p(\text{correct})$ faster.

Rating of perceived effort across elicitation policies In addition to these performance-based metrics, we also ask users to rate how difficult they found the elicitation process to be. Specifically, we asked users “How mentally demanding was writing your answer?” in the non-interactive-elicitation setting, and “How mentally demanding

was interacting with the chatbot?” in all elicitation settings (which include all other settings from Section 2.2). The “mentally demanding” wording was taken from the NASA TLX (Hart and Staveland, 1988). The question was assessed via a Likert scale from 1 (Very Little) to 7 (Very High). We also consider several additional questions to assess other usability tradeoffs. See Appendix E for the full list.

5 Results

Evaluation results are shown in Figures 3 and 4. Additional results, including sample conversations, can be found in Appendix C. Additional analyses can be found in Appendix D.3. Our results show that GATE methods...

...are successfully able to elicit human preferences. GATE improves over no elicitation, where the model is prompted to make decisions before any user interaction. This is the case across all domains studied (a positive score in Figure 3), with significance at the 0.05 level for all but the email domain, where only generative active learning was significant.

...are comparable to or better than other elicitation methods. In the majority of settings (7 out of 9), GATE elicitation methods improve over user-written prompts. In particular, generative yes/no questions improve over user-written prompts in every setting studied (although we lack enough power to assess significance in the moral reasoning domain). Furthermore, in the content recom-

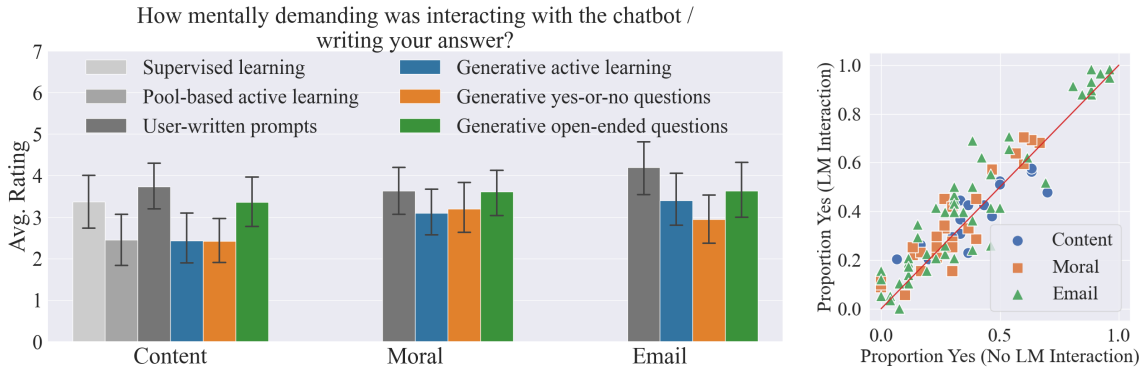


Figure 4: **Left: GATE methods are equally or less mentally demanding than other methods.** We plot the perceived mental demand across methods and domains (higher = greater mental demand). **Right: Language model elicitation does not shift human preferences.** We plot the proportion of participants who answered "yes" to each test question, comparing no LM interaction (user-written prompts) to LM interaction (GATE) elicitation. The red line is the $y = x$ curve, which serves as a guideline to see how well humans' no-LM interaction preferences align with their preferences post-LM interaction (if they align perfectly, the points should fall along this curve). We see that the points generally hover around this curve.

514 mentation setting, GATE elicitation methods (par- 546
 515 ticularly generative open-ended questions) signifi- 547
 516 cantly improve over supervised learning and pool- 548
 517 based active learning. 549

518 **...are equally or less mentally demand-** 550
 519 **ing than user-written prompts.** As shown 551
 520 in Figure 4 (left), users generally find interactive 552
 521 elicitation methods to be less mentally demand- 553
 522 ing, especially ones that involve labeling samples or 554
 523 answering yes/no questions, than non-interactive 555
 524 prompting. 556

525
 526 We additionally run some of the settings using 557
 527 an open-source LM, Mixtral, which can be found 558
 528 in Appendix C.3. Our results show that Mix- 559
 529 tral performs comparably to GPT-4, indicating that 560
 530 open-source models can be used in place of GPT-4 561
 531 for GATE. 562

532 6 Other Related Work

533 A fundamental challenge across many fields is 563
 534 how to obtain information about people's nebu- 564
 535 lous thoughts, preferences, and goals (Ericsson 565
 536 and Simon, 1980; Henderson et al., 1995; Chris- 566
 537 tel and Kang, 1992; Zowghi and Coulin, 2005; 567
 538 Pacheco et al., 2018). Many works attempt to 568
 539 computationally describe or query human prefer- 569
 540 ences, through bandits, Bayesian methods, inverse 570
 541 reinforcement learning, generative modeling, and 571
 542 more (Robbins, 1952; Yue et al., 2012; Chajewska 572
 543 et al., 2000; Emery and Nenarokomov, 1998; Ng 573
 544 et al., 2000; Hadfield-Menell et al., 2016; Mulla 574
 545 and Gharpure, 2023; Zhu and Bento, 2017). Most 575

546 relevant to our work is active learning, which cen- 547
 548 ters on how models can choose useful data points 548
 549 to learn from (Lewis and Catlett, 1994; Settles and 549
 550 Craven, 2008; Settles, 2009; Houlby et al., 2011; 550
 551 Tamkin et al., 2022b). We extend this line of inves- 551
 552 tigation to the generative setting, clarifying user 552
 553 intent by querying a user with *generated* examples 553
 554 and questions. 555

554 7 Discussion and Conclusion

555 We introduced the GATE framework to interac- 555
 556 tively elicit preferences from human users with 556
 557 free-form queries and answers. We presented ini- 557
 558 tial evidence that LMs can successfully imple- 558
 559 ment GATE to elicit human preferences (some- 559
 560 times) more accurately and with less effort than 560
 561 supervised learning, active learning, or prompting- 561
 562 based approaches. There are many ways to ex- 562
 563 pand on our implementation of GATE: Future work 563
 564 may explore more principled methods for elicita- 564
 565 tion, for example, integrating explicit notions of 565
 566 uncertainty. Second, larger models may be more 566
 567 capable elicitors: future work can explore scal- 567
 568 ing laws for elicitation. Finally, many real-world 568
 569 tasks such as software design and legal/medical 569
 570 decision-making present a richer set of constraints 570
 571 and edge cases. These applications thus offer a 571
 572 rich space of possible extensions of GATE. 572

573 Ethical Considerations

574 Our work presents several potential ethical bene- 574
 575 fits and risks. 575

576 There are many potential benefits of machines 576
 577 that can better elicit and understand human prefer- 577

ences. For example, by making it easier for software designers to incorporate nuanced user preferences, GATE may empower people with rare preferences or preferences that have historically not been considered when building software systems. In addition, improving the effort-performance ratio, especially by requiring less user typing, may help make language models more accessible to users with less time, familiarity with language models, or physical ability to use such systems.

However, this direction carries risks as well. In particular, work on *thin slicing* (Ambady and Rosenthal, 1992) has demonstrated that small amounts of information about a user can sometimes be used to predict a broader range of personal characteristics, raising potential privacy considerations. The interactive nature of GATE also risks increasing *automation bias* (Goddard et al., 2012), where users place undue weight on a model’s predictions. However, further work is necessary to establish if or when these risks are more significant for GATE than for prompting-based approaches to steering language models.

Limitations

In this work, our exploration of GATE methods has been limited prompt-based approaches, and no explicit optimization of the objective in Equation (2). Future work can examine different ways of implementing free-form interactive querying, including approaches that might combine explicit optimization with the flexibility of language models.

In our human experiments (Section 5), we did not have the budget to survey a massive number of humans for human experiments. Thus, we were unable to establish statistical significance of GATE above baselines in certain domains. Furthermore, our sample of humans may be biased, as all of them speak English and are from the United States. This means that we have likely not captured the full spectrum of human preferences.

Finally, we would like note that our moral reasoning domain is very simplistic, and may be unable to capture all the nuances of human moral preference. This paper also does not endorse aligning to every potential human preference, understanding there are ethical risks to doing so. Overall, designers of public-facing systems that make decisions may wish to implement safeguards against allowing anyone to specify moral judgments. (While this paper is not an endorsement of any particular moral preference, it provides a

framework for understanding the nuances of a particular set of preferences. Once a particular standard, or set of standards, has been decided upon, we would like the systems to ideally *fully comprehend* the nuances of the standard, to be in full alignment with that standard.)

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256.

Gwern Branwen. 2020. *GPT-3 nonfiction — calibration*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. *Language Models are Few-Shot Learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Urszula Chajewska, Daphne Koller, and Ronald Parr. 2000. Making rational decisions using adaptive utility elicitation. In *Aaai/Iaai*, pages 363–369.

Michael G Christel and Kyo C Kang. 1992. Issues in requirements elicitation.

David Cohn, Les Atlas, and Richard Ladner. 1994. *Improving generalization with active learning*. *Mach. Learn.*, 15(2):201–221.

Ido Dagan and Sean P. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning, ICML’95*, page 150–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

682	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.	735
683		736
684		737
685		738
686		739
687		740
688		
689		741
690		742
		743
691	Ashley F Emery and Aleksey V Nenarokomov. 1998. Optimal experiment design. <i>Measurement Science and Technology</i> , 9(6):864.	744
692		745
693		746
694	K Anders Ericsson and Herbert A Simon. 1980. Verbal reports as data. <i>Psychological review</i> , 87(3):215.	747
695		
696	Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. <i>Advances in neural information processing systems</i> , 31.	748
697		749
698		750
699	Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. <i>Journal of the American Medical Informatics Association</i> , 19(1):121–127.	751
700		752
701		753
702		754
703		
704	Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. <i>Advances in neural information processing systems</i> , 29.	755
705		756
706		757
707		758
708	Sandra G. Hart and Lowell E. Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research . In Peter A. Hancock and Najmedin Meshkati, editors, <i>Human Mental Workload</i> , volume 52 of <i>Advances in Psychology</i> , pages 139–183. North-Holland.	759
709		760
710		761
711		762
712		763
713		764
714	Ron D Henderson, Mike C Smith, John Podd, and Hugo Varela-Alvarez. 1995. A comparison of the four prominent user-based methods for evaluating the usability of computer software. <i>Ergonomics</i> , 38(10):2030–2044.	765
715		766
716		767
717		768
718		769
719	Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. <i>arXiv preprint arXiv:1112.5745</i> .	770
720		771
721		772
722		773
723	Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks . In <i>Advances in Neural Information Processing Systems</i> , volume 25. Curran Associates, Inc.	774
724		775
725		776
726		777
727		778
728	David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. <i>arXiv preprint arXiv:2009.09153</i> .	779
729		780
730		781
731	David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In <i>Machine learning proceedings 1994</i> , pages 148–156. Elsevier.	782
732		783
733		784
734		785
		786
		787
		788
	David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In <i>Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94</i> , page 3–12, Berlin, Heidelberg. Springer-Verlag.	
	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. <i>arXiv preprint arXiv:2205.14334</i> .	
	Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models .	
	Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. <i>Progress in Artificial Intelligence</i> , 12(1):1–32.	
	Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning. In <i>Icml</i> , volume 1, page 2.	
	OpenAI. 2023. Gpt-4 technical report .	
	Carla Pacheco, Ivan García, and Miryam Reyes. 2018. Requirements elicitation techniques: a systematic literature review based on the maturity of the techniques. <i>IET Software</i> , 12(4):365–378.	
	Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing</i> . Association for Computational Linguistics.	
	Herbert Robbins. 1952. Some aspects of the sequential design of experiments. <i>Bulletin of the American Mathematical Society</i> , 58(5):527 – 535.	
	Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.	
	Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In <i>Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing</i> , pages 1070–1079.	
	Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022a. Task ambiguity in humans and language models. <i>arXiv preprint arXiv:2212.10711</i> .	
	Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022b. Active learning helps pretrained models learn the intended task. <i>Advances in Neural Information Processing Systems</i> , 35:28140–28153.	
	Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation . In <i>Proceedings of the 58th Annual Meeting</i>	

789 *of the Association for Computational Linguistics*,
790 pages 3597–3606, Online. Association for Compu-
791 tational Linguistics.

792 Yisong Yue, Josef Broder, Robert Kleinberg, and
793 Thorsten Joachims. 2012. The k-armed dueling ban-
794 dits problem. *Journal of Computer and System Sci-*
795 *ences*, 78(5):1538–1556.

796 Jia-Jie Zhu and José Bento. 2017. Genera-
797 tive adversarial active learning. *arXiv preprint*
798 *arXiv:1702.07956*.

799 Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B.
800 Brown, Alec Radford, Dario Amodei, Paul Chris-
801 tiano, and Geoffrey Irving. 2020. *Fine-tuning lan-*
802 *guage models from human preferences*.

803 Didar Zowghi and Chad Coulin. 2005. Requirements
804 elicitation: A survey of techniques, approaches, and
805 tools. *Engineering and managing software require-*
806 *ments*, pages 19–46.

A GATE Prompts 807

A.1 Elicitation Prompts 808

809 Below, we list the prompts we use for generat-
810 ing elicitation queries (to instantiate the elicitation
811 policy \mathcal{E}) for the various GATE methods.

Generative active learning The prompt for
812 generative active learning can be found below.
813 Bracketed, italicized text are placeholders, e.g. for
814 the elicitation transcript (so far). Color-coded text
815 is domain-specific, e.g. the blue text represents
816 text for the content recommendation domain, or-
817 ange text represents text for the moral reasoning
818 domain, and the green text represents text for the
819 email validation domain. 820

Your task is to [*learn what topics a user is interested in reading online article about. People's interests are broad, so you should seek to understand their interests across many topics; in other words, go for breadth rather than depth. Do not assume a user has given a complete answer to any question, so make sure to keep probing different types of interests | learn under what conditions the user would believe it is or is not ethical to steal a loaf of bread | learn what rules a user believes a valid email address format must adhere to (e.g. for developing a regex format checker)*].

Come up with a potential edge case to learn as much information as you can about what their desired behavior should be under different circumstances. Make sure the edge case addresses different aspects of the system than the edge cases that have already been considered.

An example edge case is: [*Are you interested in the following article? Website Name: TechTalkDaily Title: 10 Must-Have Gadgets for 2023: A Comprehensive Guide Description: Discover the top 10 gadgets that will revolutionize your daily life in 2023. From smart home devices to the latest wearables, we've got you covered with our comprehensive guide. | Situation: Is it ethical to steal a loaf of bread if you are starving to death and the bread would otherwise go to waste? | Should the following email be accepted? username@example.com*] 821

Current cases:
[*Elicitation transcript*]

Generate the most informative edge case that, when answered, will reveal the most about the desired behavior beyond what has already been queried for above. Generate the edge case in the following format, and nothing else: "[*Are you interested in the following article? [edge case] | Situation: [edge case] | Should the following be accepted? [edge case]*]"

Generating Questions The prompt for generat-
822 ing both yes-or-no and open-ended questions can
823 be found below. Once again, bracketed, italicized
824 text are placeholders and color-coding indicates
825 text for specific domains. 826

Your task is to [*learn what topics a user is interested in reading online article about. People's interests are broad, so you should seek to understand their interests across many topics; in other words, go for breadth rather than depth. Do not assume a user has given a complete answer to any question, so make sure to keep probing different types of interests* | *learn under what conditions the user would believe it is or is not ethical to steal a loaf of bread* | *learn what rules a user believes a valid email address format must adhere to (e.g. for developing a regex format checker)*].

Previous questions:
[*Elicitation transcript*]

Generate the most informative [*yes/no question* | *open-ended question*] that, when answered, will reveal the most about the desired behavior beyond what has already been queried for above. Make sure your question addresses different aspects of the implementation than the questions that have already been asked. At the same time however, the question should be bite-sized, and not ask for too much at once. Phrase your question in a way that is understandable to non-expert humans; do not use any jargon without explanation. Generate the [*yes/no question* | *open-ended question*] and nothing else:

A.2 Decision-making Prompts

Given an elicitation transcript, we prompt the LM as follows to serve as a predictor $\hat{f}(s)$ to make decisions about each test case.

Content Recommendation The prompt for content recommendation can be found below. Bracketed, italicized text are placeholders, for the elicitation transcript and the test case sample.

A user has a particular set of preferences over what articles they would like to read. They have specified their preferences below:
[*Elicitation transcript*]

Based on these preferences, would the user be interested in reading the following article? Answer with a probability between 0 and 1, where 0 means "definitely not interested" and 1 means "definitely interested". Only output the probability and nothing else. If uncertain, make your best guess.
[*Test case sample*]

Moral Reasoning The prompt for moral reasoning can be found below. Once again, bracketed, italicized text are placeholders.

A user has a particular ethical code they follow. The following response(s) represent when this user would believe it is ethical to steal a loaf of bread.
[*Elicitation transcript*]

"Based on these response(s), would the user believe it is ethical to steal a loaf of bread in the following situation? Answer with a probability between 0 and 1, where 0 means "definitely not" and 1 means "definitely yes". Only output the probability and nothing else. If uncertain, make your best guess.
[*Test case sample*]

Email Validation The prompt for email validation can be found below. Once again, bracketed, italicized text are placeholders.

A user has a particular format of emails that they believe to be valid. The following answer(s) represent this user's preferences of whether these emails adhere to their desired format.

[*Elicitation transcript*]

Based on the user's preferences, does the following email adhere to the user's desired format? Answer with a probability between 0 and 1, where 0 means "definitely not" and 1 means "definitely yes". Only output the probability and nothing else. If uncertain, make your best guess.

[*Test case sample*]

B Experimental Details

B.1 Number of Participants

The number of participants we recruited for our study, for each elicitation method and domain, can be found in the table below.

B.2 User Interface Details

Details about the UI we built for our experiments can be found below. Recall that the human studies proceeded in two parts: elicitation, followed by decision-making.

B.2.1 Elicitation

For supervised learning, pool-based active learning, and the GATE methods, we had participants respond to a series of queries using the chatbot interface (Figure 5). For prompting, we had participants input a task description using the text-input interface (Figure 6).

The instructions for this phase can be found below.

Supervised Learning / Pool-based Active Learning

We present users with the following instructions for both supervised learning and pool-based active learning. Bracketed, italicized text represent placeholders for domain-specific text. [*Domain instructions*] is a placeholder for the top-level instructions for each domain (see Table 2). Otherwise, blue text represents text for the content recommendation domain, orange text represents text for the moral reasoning domain, and green text represents text for the email validation domain.

	Content Recommendation	Moral Reasoning	Email Validation	Total
Supervised learning	30	-	-	30
Pool-based active learning	31	-	-	31
Prompting	30	30	26	86
Generative active learning	30	30	20	80
Generative yes-or-no questions	31	30	19	80
Generative open-ended questions	31	31	19	81
Total	183	121	84	388

Table 1: Breakdown of how many participants we recruit for each domain and elicitation method.

How do you perceive the balance between personal survival and societal laws when it comes to stealing a loaf of bread?

Timer: 3m 28s

personal survival takes precedence

How do you view the impact of stealing a loaf of bread on the person or business from whom it is stolen, and how does this influence your ethical stance on the matter?

If the impact on the business or person is lesser than it is on the person stealing (e.g. big business that can absorb the cost), then it may be ethical

How do you consider the potential long-term societal consequences of stealing a loaf of bread, such as the potential for increased prices or stricter security measures, in your ethical evaluation?

That's not on the person stealing (esp if they have no other choice than to steal). That's a reflection of social structures.

Send

Figure 5: Chatbot UI built for elicitation phases of GATE methods, supervised learning, and pool-based active learning.

In the final part of the study, you will give feedback on a test so we can enable us to see how well a chatbot reading your responses handles.

Timer: 2m 27s

I like to read about science, politics, and lifestyle. In particular I think AI is really cool and I want to read more articles about AI. I also enjoy reading articles about the latest developments in astronomy but I don't want the articles to be too dense because I am not an expert in the field. ...

Figure 6: Text-input UI built for elicitation phase for prompting.

Content	<p>We are testing a system for understanding people's interest in reading different kinds of online articles.</p> <p>For example, you might be interested in articles about some topics, but not about others.</p>
Moral	<p>We are testing a system for understanding people's fuzzy intuitions and preferences.</p> <p>In this experiment, we'll be capturing your moral intuitions about the act of stealing a loaf of bread, and whether there are certain cases where stealing may be morally permissible.</p>
Email	<p>We are testing a system for understanding people's fuzzy intuitions and preferences.</p> <p>In this activity, we're going to be looking at different strings of text and you'll be deciding if they look like they could be an email address or not. For example, most people would agree that "username@domain.com" looks like an email address, while "n12z5IFEN4" does not. However, the rules for what can be an email address can be very unusual, so what we're really interested in is your intuition on what an email address could look like.</p> <p>Important: We are not asking you to determine the rules for a *good* email address, or a *real (non-spam)* email address. We are simply asking about your intuition as to why certain strings look like email addresses and certain strings do not.</p> <p>Tip: in an email such as username@cs.stanford.edu, "username" is called the local-part of the email, while "cs.stanford.edu" is the domain. Furthermore, "cs" is a subdomain, and "edu" is a top-level domain.</p>

Table 2: Domain-specific instructions presented to users for the elicitation phases.

<p><i>[Domain instructions]</i></p> <p>Try to answer in a way that accurately and comprehensively conveys your preferences, such that someone reading your responses can understand and make judgments as close to your own as possible. Feel free to respond naturally (you can use commas, short phrases, etc), and press [enter] to send your response. Note that the chatbot technology is imperfect, and you are free to avoid answering any questions that are overly broad or uncomfortable. When interacting with the chatbot, please avoid asking follow-up questions or engaging in open-ended dialogue as the chatbot is unable to respond to you.</p> <p>Note: The chatbot will stop asking questions after 5 minutes, after which you can send your last response and you will be taken to the final part of the study.</p> <p>In the final part of the study, you will give feedback on a test set of <i>[article headline and descriptions moral situations email addresses]</i>, which will enable us to see how well a chatbot reading your responses has learned <i>[what you like and dislike your moral preferences your email preferences]</i>.</p>	<p><i>[Domain instructions]</i></p> <p>To the best of your ability, please explain all details about <i>[your preferences of what kinds of online articles you would like to read your belief of when it is moral to steal a loaf of bread your intuition of what makes email addresses look like email addresses]</i>, such that someone reading your responses can understand and make judgments as close to your own as possible. Try to be as detailed as possible. For example, if you were writing a regex that accepts only email-address-like strings, what might that regex look like? What are permissible / non-permissible symbols and characters, and in what positions?</p> <p>Note: You will have up to 5 minutes to articulate your preferences. Please try to submit your response within that time. After you submit, you will be taken to the final part of the study.</p> <p>In the final part of the study, you will give feedback on a test set of <i>[article headline and descriptions moral situations email addresses]</i>, which will enable us to see how well a chatbot reading your responses has learned <i>[what you like and dislike your moral preferences your email preferences]</i>.</p>
---	--

Prompting We present users with the following instructions for prompting. Similar to above, bracketed, italicized text represent places where we insert domain-specific text.

GATE methods We present users with the following instructions for the three GATE methods (generative active learning, generative yes-or-no questions, generative open-ended questions). Once again, bracketed italicized text represent domain-specific text.

876

881

877

882

878

883

879

884

880

885

886

887

[Domain instructions]

This chatbot will ask you a series of questions about [*your preferences of what kinds of online articles you would like to read* | *your belief of when it is moral to steal a loaf of bread* | *your intuition of what makes email addresses look like email addresses*]. Try to answer in a way that accurately and comprehensively conveys your preferences, such that someone reading your responses can understand and make judgments as close to your own as possible. Feel free to respond naturally (you can use commas, short phrases, etc), and press [enter] to send your response. Note that the chatbot technology is imperfect, and you are free to avoid answering any questions that are overly broad or uncomfortable. When interacting with the chatbot, please avoid asking follow-up questions or engaging in open-ended dialogue as the chatbot is unable to respond to you.

Note: The chatbot will stop asking questions after 5 minutes, after which you can send your last response and you will be taken to the final part of the study.

In the final part of the study, you will give feedback on a test set of [*article headline and descriptions* | *moral situations* | *email addresses*], which will enable us to see how well a chatbot reading your responses has learned [*what you like and dislike* | *your moral preferences* | *your email preferences*].

B.2.2 Decision-making

For the decision-making phase, we presented users with the following instruction:

Content Recommendation

Please indicate whether you would like to read the following articles: “yes” if you would, “no” if you would not.

Moral Reasoning

Please indicate whether you think the following situations are morally permissible or not: “yes” if they are, “no” if they aren’t.

Email Validation

Please indicate whether you think the following strings look like reasonably well-formatted email addresses or not: “yes” if they do, “no” if they don’t.

Users are then presented with a list of test samples, and can use radio buttons to select whether each test-case sample is acceptable. See Figure 7.

C Additional Results

C.1 Sample Transcripts

Sample transcripts of users interacting with the various generative active task elicitation methods can be found in Figure 8.

C.2 AUROC results

We measure AUROC over model-generated probabilities in addition to $\Delta p(\text{correct})$. Figure 9 is the analogous plot to Figure 3, but we measure the improvement in AUROC instead of $p(\text{correct})$, over

interaction time, rewarding methods that achieve higher improvements in AUROC sooner.

The general trends hold from Section 5: language models can elicit human preferences (beyond no interaction), and language model elicitation is comparable or better than other elicitation baselines. However, unlike the $p(\text{correct})$ metric, the AUROC metric is a simple classification-based metric. Due to potential miscalibration in LMs, making it difficult for them to output well-calibrated probabilities with the same threshold across questions, the overall improvements in this metric are lower (particularly for generative open-ended questions) and the variances are much higher. Thus, we see that it is harder to establish statistical significance using this metric.

C.3 Mixtral Results

To test the robustness of GATE to the choice of underlying language model (for elicitation and prediction), and particularly to test out how well open-source models perform relative to GPT-4, we try to replicate some of our experiments on Mixtral (Mixtral-8x7B-Instruct-v0.1), one of the state-of-the-art open-source models as of February 2024. We focus on the content recommendation setting as the one where GPT-4 models performed the best. Using the same prompts and overall setup as for GPT-4, we test Mixtral’s ability to generate *open-ended* questions. We separately test out how effective Mixtral is as an elicitor and as a predictor. Thus we have six total settings, with three elicitation modes:

1. Mixtral yes-or-no questions
2. GPT-4 yes-or-no questions
3. User-written prompts

each of which are evaluated by two prediction models:

1. Mixtral predictor
2. GPT-4 predictor

Results are shown in Figure 10. We find that the Mixtral model performs comparably to GPT-4 as both an elicitor, and as a predictor, indicating that our results are robust across different choices of underlying language models. This is also a promising indication that open-source models can be used instead of GPT-4 for preference elicitation.

Task Elicitation Study

Please indicate whether you think the following situations are morally permissible or not: "yes" if they are, "no" if they aren't. Optionally, you may provide an explanation for your decision for each example.

Test Case	Yes	No	Explanation (Optional)	Your Response(s)
Is it ok to steal the loaf of bread if you are starving to death but have the option of begging for food instead?	<input type="radio"/>	<input type="radio"/>	<input type="text"/>	<p>Your Response(s) Your submitted response(s) are provided for reference, but please make decisions based on your present intuition, not strictly based on these responses.</p> <ol style="list-style-type: none">How do you perceive the balance between personal survival and societal laws when it comes to stealing a loaf of bread? <i>personal survival takes precedence</i>How do you view the impact of stealing a loaf of bread on the person or business from whom it is stolen, and how does this influence your ethical stance on the matter? <i>If the impact on the business or person is lesser than it is on the person stealing (e.g. big business that can absorb the cost), then it may be ethical</i>How do you consider the potential long-term societal consequences of stealing a loaf of bread, such as the potential for increased prices or stricter security measures, in your ethical evaluation? <i>That's not on the person stealing (esp if they have no other choice than to steal). That's a reflection of social structures.</i>How do you weigh the ethical implications of stealing a loaf of bread in a situation where there are other potential, albeit more difficult, legal alternatives to obtain food? <i>If the person is in a survival situation, it's still fine. Otherwise, depends on how difficult the alternative is.</i>
Would it be acceptable to steal the loaf of bread to feed an animal starving to death?	<input type="radio"/>	<input type="radio"/>	<input type="text"/>	
Is it ok to steal the loaf of bread if you are starving to death, but the store owner is also struggling financially?	<input type="radio"/>	<input type="radio"/>	<input type="text"/>	
Would it be acceptable to steal the loaf of bread if you are not starving to death, but you know someone who is, however they are known to be proud and refuse charity?	<input type="radio"/>	<input type="radio"/>	<input type="text"/>	
Is it morally permissible to steal the loaf of bread if you are starving to death, and the store owner is known to be an evil, cruel person?	<input type="radio"/>	<input type="radio"/>	<input type="text"/>	

Figure 7: UI for the decision-making phase.

958 Qualitatively, we find that the open model likes
959 to quickly hyperfixate on a topic once a user raises
960 it as a topic of interest. An example can be found
961 in Appendix C.3, where Mixtral, upon learning the
962 user likes post-apocalyptic fiction on the first turn,
963 pivots to only asking questions related to post-
964 apocalyptic fiction.

965 D Model-model experiments

966 D.1 Methods

967 We explore whether LMs can stand-in for human
968 participants, enabling faster iteration loops and
969 more research in this area. We generate various
970 personas (specified in natural language) for each
971 domain, and prompt LMs to respond to elicitation
972 queries as their persona would.

973 For each domain, we construct a set of personas
974 as follows:

975 **Content Recommendation** The personas are
976 constructed by providing a brief biographical
977 sketch of a hypothetical person, and were also con-
978 structed by the authors. A sample persona prompt
979 is “*Education: Medical Doctorate. Occupation:*
980 *Junior Surgeon at a regional hospital. Hobbies:*
981 *Running marathons, traveling, and learning new*
982 *languages.*”

983 **Moral Reasoning** We construct a variety of per-
984 sonas with a diverse array of moral perspectives,
985 including Kantianism, Utilitarianism, and ethical
986 egoism. A sample persona prompt is “*You sub-*
987 *scribe to a Kantian code of ethics.*”

Email Validation Personas are instantiated by
988 providing a regex to the model. The test cases
989 are constructed by the authors. A sample persona
990 prompt is “*You are validating that an email ad-*
991 *dress adheres to a specific format (e.g. for de-*
992 *signing a Python regex). The gold regex is ...*
993 *user@domain.co.co.co.co*”
994

995 We prompt as the LM as follows to answer
996 questions according to their personas:

[Persona] Answer the question in the shortest way with
minimal additional explanation.
[Question]

997
998 Furthermore, in the content recommendation
999 domain, we implement three different selection
1000 strategies for pool-based active learning and ex-
1001 plore their trade-offs, including random sampling
1002 (randomly selecting the next example to query),
1003 uncertainty-based sampling (selecting the exam-
1004 ple whose answer the LM is most uncertain about,
1005 i.e. the example with the highest-entropy),⁴ and
1006 diversity sampling (described in Section 4.5).⁵

1007 D.2 Results

1008 Figures 11 and 12 shows results in each domain
1009 when we use a LM to simulate humans. Because
1010 human interaction times are unavailable for these
1011 experiments, we run interactive elicitation up to 5

⁴Note that because GPT-4 does not return logits, we use a smaller GPT-3 text-davinci-003 model to compute entropy over the answer distribution

⁵To avoid massive costs in uncertainty sampling, the pool was pre-filtered to a sensible size of a few hundred samples using diversity metrics. For comparability across methods, the same pre-filtered pool was used for all three sampling methods.

Content Recommendation	<p>System: Are you interested in the following article? Website Name: HealthMindBody Title: The Intersection of Yoga and Mental Health: A Deep Dive Description: Explore the profound effects of yoga on mental well-being. This comprehensive analysis brings together research, personal stories, and tips for integrating yoga into your mental health routine. User: No</p> <p>System: Are you interested in the following article? Website Name: ClimateFutures Title: The Interplay of Modern</p>	<p>System: Do you enjoy reading articles about advancements in technology and science? User: Sometimes</p> <p>System: Are you interested in reading articles about politics and current events? User: I try to avoid politics, but I do like to read political articles sometimes</p> <p>System: Do you enjoy reading articles related to health and wellness? User: not really</p> <p>System: Do you like reading articles about travel and exploring different cultures?</p>	<p>System: What types of online communities or forums do you actively participate in, and what draws you to these specific spaces? User: Well, I currently use Reddit for the extreme fine-tuning of my preferences. If a feed is too generic, I won't like it. I go to a lot of specific subreddits.</p> <p>System: Can you share some of the specific subreddits you frequent and what topics or discussions within those subreddits you find most engaging? User: Let me try using the first five that come to mind:</p>
Moral Reasoning	<p>System: Situation: Is it ethical to steal a loaf of bread if it is from a large corporation that wouldn't notice a small loss, but you are not in immediate physical need, but rather stealing it to feed your family due to being unable to find work? User: If I am stealing it to feed my family then I probably wouldn't have a problem with it. If I was stealing it to get back at the corporation that makes it then it would be a problem.</p> <p>System: Situation: Is it ethical to steal a loaf of bread if you're in</p>	<p>System: Do you believe it is ethical to steal a loaf of bread if it is to feed a starving child? User: Yes bc its to help a child in need</p> <p>System: Do you believe it is ethical to steal a loaf of bread if you can afford to buy it but the store is closed? User: No bc you are causing damage to the store in order to steal it</p> <p>System: Do you believe it is ethical to steal a loaf of bread if it is the only way to feed a large group of people in need? User: Yes</p>	<p>System: How do you perceive the impact of personal circumstances, such as extreme hunger or poverty, on the ethical considerations of stealing a loaf of bread? User: I think it's a terrible tragedy when people don't have enough money to eat or feed their families. While stealing is wrong, I don't think stealing bread for survival is immoral, it's survival.</p> <p>System: How do you view the role of societal systems and structures in creating situations where someone might feel compelled to steal a loaf of bread, and does this</p>
Email Verification	<p>System: Should the following be accepted? user.name+alias@example.co.uk User: no</p> <p>System: Should the following be accepted? user_name-123@example.org User: yes</p> <p>System: Should the following be accepted? user@subdomain.example.com User: sure, but the "example.com" is a bit weird</p> <p>System: Should the following be accepted? user.name@123-example.com User: sure</p> <p>System: Should the following be accepted?</p>	<p>System: Do you want to allow email addresses with subdomains and/or special characters (e.g., "+" or "_") in the username part? User: Yes, I want to allow email addresses with subdomains, but not with '+' sign but '_' is okay</p> <p>System: Do you want to enforce a specific length limit for the username and domain parts of the email address, or allow any length as long as it is a valid email format? User: Yeah, we need a 128 characters limit for username part and same for domain parts</p>	<p>System: Can you describe what kind of special characters, if any, you believe should be allowed in an email address and where they should be placed? User: A dot separating name</p> <p>System: What are your thoughts on the use of numbers and capital letters in an email address, and do you believe there should be any restrictions on where they can be placed? User: No use of capital letters, email should be insensitive to that</p> <p>System: What are your thoughts on the length of an email address? Do</p>
	Generative Active Learning	Generative Yes / No Questions	Generative Open-Ended Questions

Figure 8: Excerpts of real transcripts across the different domains and elicitation methods we investigate. The System messages are generated by the language model, while the User messages are produced by human participants. Overall, the model is able to generate diverse and contextually-appropriate questions in each setting. See Sections 3.1 and 4.1 for more details on the domains and methods respectively.

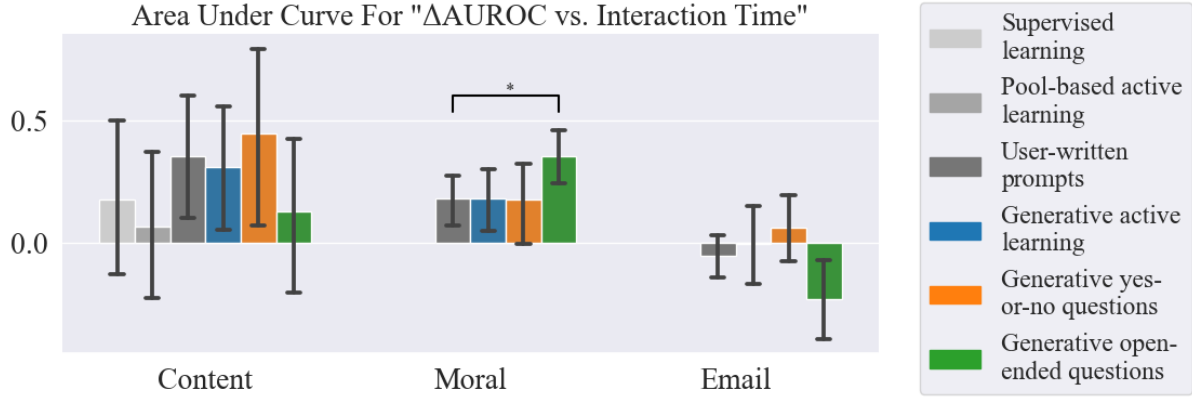


Figure 9: We plot the **Area Under the “ Δ AUROC vs. Interaction time” Curve**, which gives us a metric for how well (and how quickly) each elicitation method is at aligning with human preferences. This plot is analogous to Figure 3, only we are using AUROC instead of $p(\text{correct})$ for the alignment metric, which means that we are not measuring uncertainty. We see the same trends hold of GATE methods (generally) beating supervised learning, pool-based learning, and prompting approaches, while also beating no interaction (Δ AUROC = 0) using this metric. However, we see generally smaller Δ s over non-interaction using this metric, and higher variances, which make it harder to establish statistical significance.

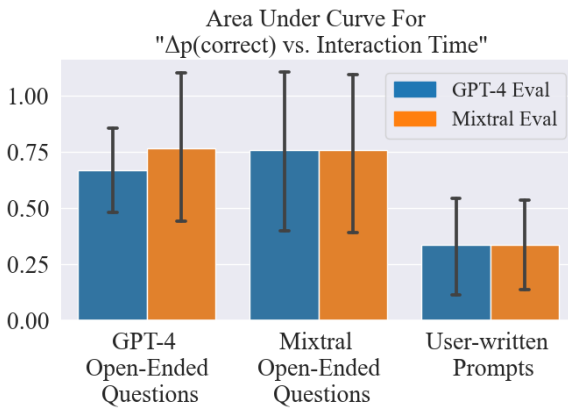


Figure 10: Effectiveness of Mixtral as an *elicitor* and a *predictor* in GATE.

turns, where we use number of turns as a proxy for human effort. Note that instead of measuring AUC of the “ $\Delta p(\text{correct})$ vs. interaction time” curve, we instead measure AUC of the “ $\Delta p(\text{correct})$ vs. number of turns” curve.

Can models be used to simulate human participants? In Figure 13, we plot the correlation between human experiment results and model-model experiment results for various elicitation methods. For both the human experiments and the model-model experiments, we compute the area under the “ $\Delta p(\text{correct})$ vs. number of turns” curve, in addition to the average change in $p(\text{correct})$ after 5

turns.⁶

We find that on both metrics we evaluate, the model-model results generally correlate with human results in the content recommendation and email validation domains (methods that perform better in the model-model experiments generally also perform better in the human experiments), but not the moral reasoning domain. This could be for various reasons, including that the subtleties in human moral reasoning may be difficult to capture in a single persona prompt, and difficult to simulate even with our biggest LMs.

Which sampling strategy is the best for pool-based active learning? As seen in Figure 11, we experiment with three different pool-based active learning strategies (random, diversity-based, and uncertainty-based sampling), which perform comparably, with diversity sampling perhaps performing slightly better than the rest. This is in line with the findings from (Margatina et al., 2023). Thus, we use diversity sampling in our main human experiments.

⁶Note that these metrics differ from we use to evaluate the human experiments in Section 4.5 – in particular by being turn-based instead of time-based – meaning we had to additionally compute these metrics on the human transcripts. This is necessary here because we must ensure that the model-model results and human results are measured along the same metric(s).

Assistant: What hobbies or activities do you enjoy doing in your free time, and have you ever read any articles related to those interests?
User: I really love post apocalyptic fiction and I can't recall reading any articles in particular
Assistant: That's interesting! Have you ever explored any non-fiction articles or research related to post-apocalyptic scenarios, societal collapse, or emergency preparedness?
User: no, not really
Assistant: Have you ever come across any online articles or discussions about how post-apocalyptic fiction might reflect or influence real-world emergency planning or disaster response strategies?
User: no
Assistant: Have you ever looked into how post-apocalyptic fiction might inspire or impact emergency shelters, food storage, or water purification methods in real-world scenarios?
User: no, but it sound interesting
 ...[10 turns later]...
Assistant: Have you ever encountered any post-apocalyptic fiction that highlights unique or innovative approaches to education, healthcare, or governance in reimagined societies, and what intriguing ideas have you discovered that could potentially be applied to enhance resilience and adaptability in real-world communities?
User: can we talk about sometthing else? Let's chat about running. I love to run!
Assistant: That's great to hear about your love for running! Have you ever read any articles about how running or other physical activities can play a role in building resilience and adaptability during challenging times, either in a post-apocalyptic scenario or in real-world emergency preparedness?
User: Interesting, but no

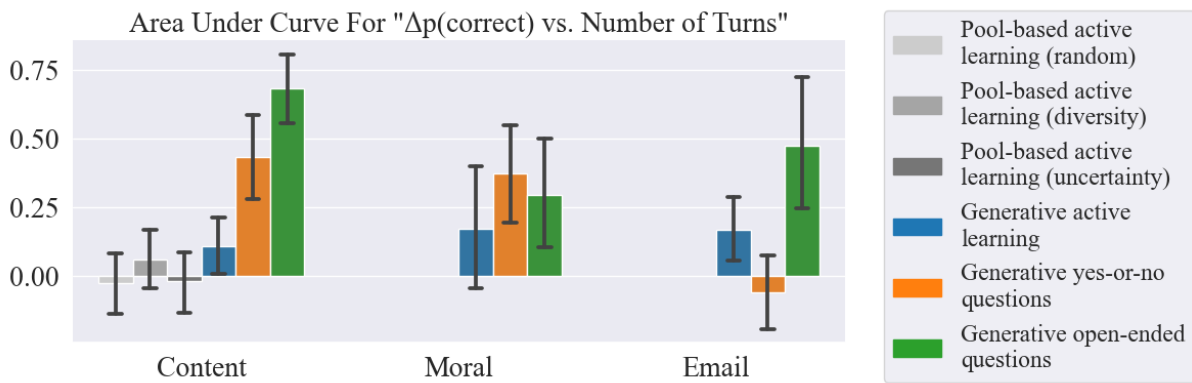


Figure 11: We plot the **Area Under the “ $\Delta p(\text{correct})$ vs. Number of Turns” Curve** for model-model experiments. This plot is analogous to Figure 3, only we are using LMs to simulate human users, and we are using number of turns as a proxy for interaction time. We see the same general trends as in Figure 3: GATE methods beat both no elicitation and pool-based active learning.

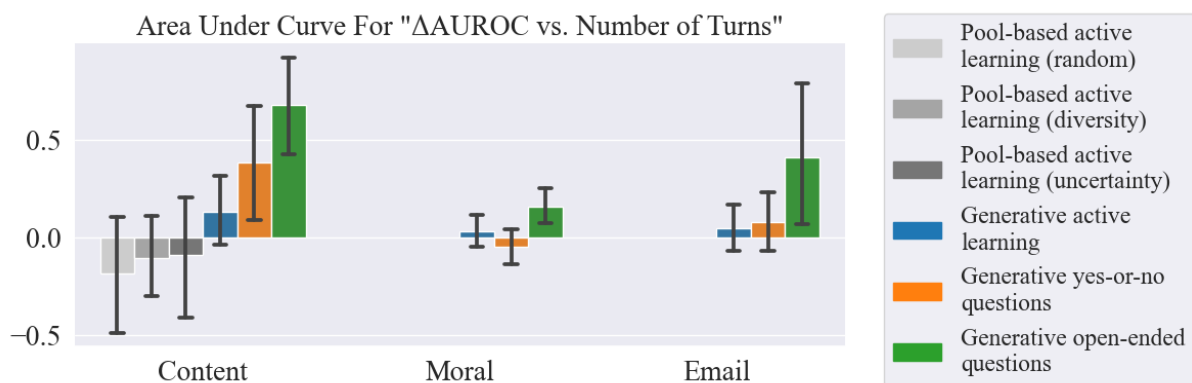


Figure 12: We plot the **Area Under the “ ΔAUROC vs. Number of Turns” Curve** for model-model experiments. This plot is analogous to Figure 9, only we are using LMs to simulate human users, and we are using number of turns as a proxy for interaction time. We see the same general trends as in Figure 9: GATE methods beat both no elicitation and pool-based active learning.

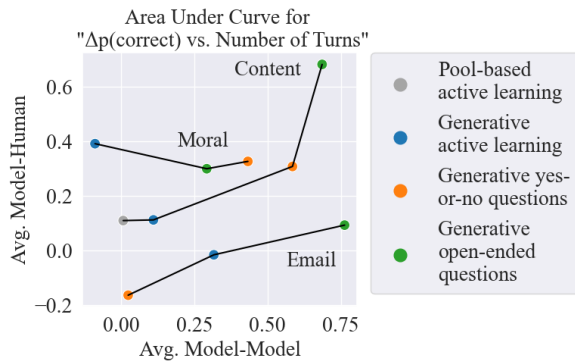


Figure 13: **Predictivity of model-model for model-human results.** We match up the *Area Under “ $\Delta p(\text{correct})$ vs. Number of Turns”* Curve metric for each elicitation method in each domain. We see that using the model to simulate human users is predictive of actual human results in the content and email domains, but not the moral domain.

D.3 Analysis

Here, we present some additional analyses to better characterize the experiments.

How much variation there is in people’s preferences? Elicitation is only helpful if there is variation in people’s preferences; otherwise, a model could simply attain maximum performance by relying on its prior and ignoring the elicited information. To quantify how much variation there is in people’s preferences, we compute the entropy in $p(\text{yes})$ for each question across participants. We find that many questions have high entropy while many others have little entropy, for an average entropy of 0.77 bits. Broadly, the results validate that our settings have significant variation in human preferences, enabling models to personalize themselves based on human preferences.

Does language model elicitation influence user preferences? Human preferences may shift when interacting with language models for a variety of reasons. For example, past work has studied *auto-induced distributional shift*, where machine learning models shift human behavior to be easier to predict (Krueger et al., 2020). To investigate whether this occurs in our experiments (or indeed if different elicitation methods induce different human preferences for any other reason), we compare the distribution of human labels on test samples from the three GATE methods with those from the user-written prompt experiments to see whether interacting with language models influ-

ences users’ subsequent judgments. As seen in Figure 4 (right), we see no such effect.

What kinds of questions did the language models ask? We show a few examples of the language model questions in Figure 8. As the figure shows, these questions are complex and subtle, often building on the previous questions, representing a broad-based knowledge of the domain as well as possible nuances therein.

Why does prompting make things worse in the emails domain? In the emails domain in Figure 3, we observe that user-written preferences slightly decrease performance relative to a no-elicitation baseline. While it is possible this is an effect of noise, we also observe that some participants articulated preferences that were actually different from those they experienced when viewing email addresses. For example, one user wrote “*an email address should finish with .com or co.uk*” yet later decided that “*user@domain.edu*” was an acceptable email address. This indicates that users may not have a clear and comprehensive understanding of their own preferences, especially in more technical domains.

Can we automate evaluation? To probe whether evaluation could be automated, we conducted experiments where we *simulated* different human preferences using language models prompted with a diverse set of (automatically-generated) *personas*. These personas varied by domain, but generally contained information about a hypothetical person’s preferences within that the domain. For example, in the content recommendation domain, we generated brief biographical sketches of hypothetical people, including their hobbies, interests, and careers, and conditioned GPT-4 on these biographical sketches to generate answers to queries. We found that model could simulate humans well in the content recommendation and email validation domains, but not in the moral reasoning domain. This suggests that while such personas may be a useful guide in some cases, they are not yet sophisticated enough to stand in for real human participants. See Appendix D for more details.

E Human ratings of usability across elicitation policies

E.1 Methods

We ask users several questions to assess usability tradeoffs across elicitation policies. The following are the full list of questions, which we ask at different points in the experiment.

After elicitation but before seeing the test-cases:

1. How mentally demanding was interacting with the chatbot? (See discussion in Section 5)
2. To what extent did the chatbot raise issues or aspects about your preferences that you hadn't previously considered?
3. How comprehensively do you feel the chatbot's questions characterized your preferences about the task?

After seeing and labelling the test cases:

4. After seeing the examples in the second part of the task, how well do you feel the answer you wrote (in the first part of the task) covered the important issues or aspects of these examples?
5. When performing the second part of the task, to what extent did you refer back to your conversation history from the first part of the task?
6. How much experience have you had (if any) with interacting with language models (e.g. ChatGPT, GPT4, etc.)?
7. Do you have any other feedback about the task?

The last question was free response. All other questions were assessed via a Likert scale from 1 (Very Little/Poorly) to 7 (Very High/Well) with radio buttons.

E.2 Results

The average ratings for the first question across each elicitation method and domain can be found in Figure 4. The average ratings for questions 2 – 5 are plotted in Figures 14 to 16.

From Fig. 14, we see that humans were on average overconfident on their ability to cover their preferences in prompts, particularly in the content

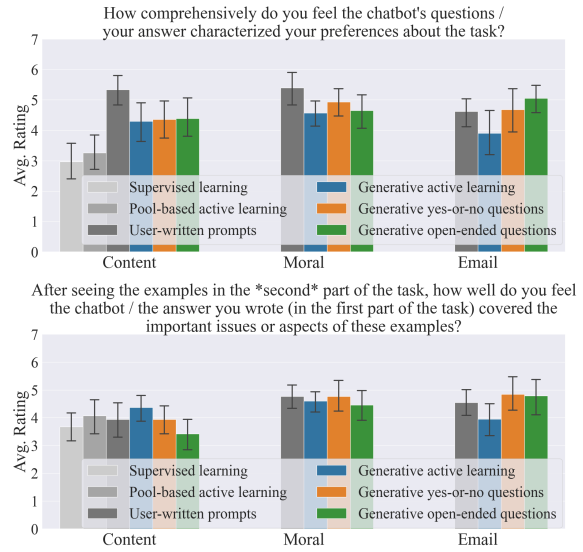


Figure 14: Average perceived coverage of each elicitation method, before (above) and after (below) seeing the test cases. Higher indicates greater coverage.

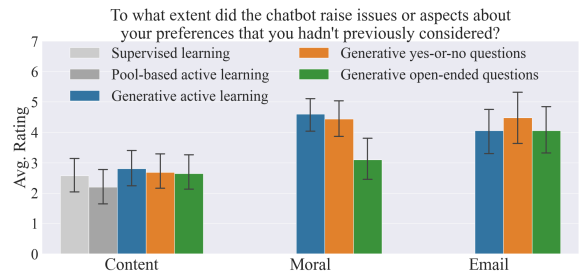


Figure 15: Extent participants perceived that each elicitation method drew out novel aspects of a domain that the user had not previously considered, averaged over each elicitation method. Higher indicates greater perceived novelty.

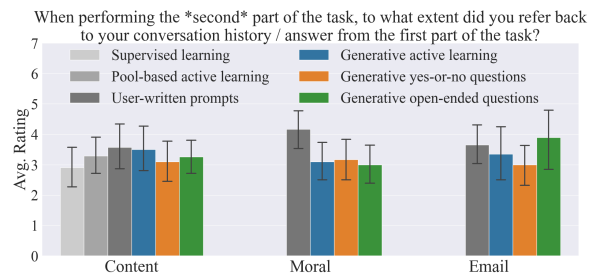


Figure 16: Extent participants referred back to the elicitation transcript when labelling test cases, averaged over each elicitation method. Higher indicates the user more heavily relied on the elicitation transcript.

1167 recommendation and moral reasoning domains,
1168 reflected in the average rating of their perceived
1169 coverage dropping from an average of 5.3 to 3.9
1170 (in the content recommendation domain) and an
1171 average of 5.4 to 4.8 (in the moral reasoning do-
1172 main) after seeing the test cases. This indicates
1173 that humans are usually not aware of their mental
1174 limitations when writing prompts.

1175 From Figure 15, we see that the generative elic-
1176 itation methods were on average able to surface
1177 more novel considerations in the moral reason-
1178 ing and email validation domains than in the con-
1179 tent recommendation domain, as they tend to have
1180 trickier and less intuitive edge cases.

1181 Finally, from Figure 16, we see the extent to
1182 which users explicitly referred back to the elic-
1183 itation history when making decisions on the test
1184 cases. This may influence how well-aligned the
1185 test case decisions are with the answers from the
1186 elicitation phase. When annotating test cases, we
1187 explicitly instruct participants *not* to follow the
1188 elicitation transcript if it does not align their in-
1189 tuition on a test sample (e.g. if the test sample
1190 surfaced a novel consideration not accounted for
1191 in the elicitation phase), though we were unable
1192 to validate how well participants followed this in-
1193 struction.

1194 **F Reproducibility**

1195 We will open-source all code used in creating
1196 GATE methods, constructing the user interface,
1197 and conducting the results and analysis. We will
1198 also release the pre-registration for our experi-
1199 ments. All prompts we used for querying GPT-
1200 4 (and Mixtral) in the decision-making and elic-
1201 itation phases, and all instructions we presented
1202 to the user, can be found in the Appendix. In all
1203 cases, we queried GPT-4 (or Mixtral) with temper-
1204 ature 0 for replicability of experiments.

1205 We also note that the model we mainly use is a
1206 closed-source model whose versions are periodi-
1207 cally deprecated, which may hinder reproducibil-
1208 ity. However, preliminary results with Mixtral
1209 indicate that open-source models are compatible
1210 with GATE and a promising avenue for future ex-
1211 ploration.