R.U.Psycho? A Framework for Robust Unified Psychometric Testing of Language Models

Julian Schelb[*1], **Orr Borin**[*2], **David Garcia**[1], **Andreas Spitz**[1]
[1]University of Konstanz, Germany [2]Recosys, Germany
{julian.schelb, david.garcia, andreas.spitz}@uni.kn, orr.borin@recosys.com

Abstract

Generative language models are increasingly being subjected to psychometric questionnaires intended for human testing, in efforts to establish their traits, as benchmarks for alignment, or to simulate participants in social science experiments. While this growing body of work sheds light on the likeness of model responses to those of humans, concerns are warranted regarding the rigour and reproducibility with which these experiments may be conducted. Instabilities in model outputs, sensitivity to prompt design, parameter settings, and a large number of available model versions increase documentation requirements. Consequently, generalization of findings is often complex and reproducibility is far from guaranteed. In this paper, we present R.U.Psycho, a framework for designing and running robust and reproducible psychometric experiments on generative language models that reduces the required coding expertise. We demonstrate the capability of our framework on a variety of psychometric questionnaires, which lend support to prior findings in the literature. R.U.Psycho is available as a Python package at https://github.com/julianschelb/rupsycho.

Keywords: Natural Language Generation; Replicability and Reproducibility; Tools, Systems, Applications

1. Introduction

The proliferation of generative large language models (LLMs) and their ease of use has recently created an interest in their experimental application in domains that are traditionally focused on human experimentation, including sociology (Ziems et al., 2024) and psychology (Pellert et al., 2024).

On the one hand, this interest stems from the desire to measure the performance and behavior of language models from a psychological perspective to assess the traits of LLMs as one would for a human, a direction that has been termed machine psychology (Hagendorff, 2023). Examples of such research are varied and range from vignette-based tasks (Binz and Schulz, 2023), over game-theoretic testing of LLMs (Duan et al., 2024), to analyses of their decision-making (Horton, 2023).

On the other hand, some researchers in computational social science view LLMs as potential proxies for human (sub)populations in social science research and human experimentation (Aher et al., 2023; Argyle et al., 2023; Dillion et al., 2023), often referred to as persona simulation. The arguments for such applications range from the reduction in cost that usage of a language model may offer, to an increase in experimental scale.

Finally, given the increasing ubiquity of LLMs that is beginning to contaminate the research artifacts created by crowdworker-participants with LLM-generated content (Veselovsky et al., 2023) and the difficulty one faces in detecting such contamination (Sadasivan et al., 2023; Jakesch et al., 2023), one might argue that establishing a baseline of LLM traits is simply a necessity for conducting online research in the age of generative models.

However, while it has its proponents, psychometric testing of LLMs is also the subject of intense criticism, as LLMs are prone to failing theory of mind tasks under even minor prompt variations (Ullman, 2023) and are subject to inherent (intersectional) biases that are poorly understood and difficult to quantify (Husse and Spitz, 2022), yet have been shown to directly affect reasoning tasks during persona simulation (Gupta et al., 2024). In addition to model-related challenges, experimental issues are abound, ranging from prompt (ordering) sensitivity (White et al., 2023; Lu et al., 2022), over non-trivial processing steps required for interpreting and mapping model responses (Wang et al., 2024), to performance variations between model families and model sizes (Ziems et al., 2024).

Fundamentally, while highlighting an important research direction, much of the early work into machine psychology has been haphazard and is fundamentally not robust or reproducible outside the exact (and often ill-documented) modeling choices and parameter settings of a given study. For a detailed overview and breakdown, we direct the reader to Löhn et al. (2024). Consequently, psychometric testing of LLMs is running the risk of mirroring the reproducibility crisis in psychology.

In this paper, we therefore take a step back from individual psychometric tests and a step towards addressing these concerns by introducing R.U.Psycho, a framework for prompt-based psychometric experimentation on generative LLMs.

[*]These authors contributed equally

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8365–8386
11–16 May 2026.

Our contributions are fourfold: (i) We present a framework for robust and configurable psychometric testing of any open- or closed-weight LLM using any questionnaire. (ii) We focus on the reproducibility of experiments through versatile, well-documented experiment configuration files. (iii) We include support for customizable prompt templates, which we illustrate at the example of simulating personas. (iv) We present results for four psychometric test and one thought experiment to demonstrate the framework's usability and to expand the experimental data in the literature.

2. Related Work

Related work can be divided into four areas, for which we provide an overview of contributions.

Persona-simulation for social science encompasses investigations whether LLMs can be used as replacement for human participants, typically through the simulation of personas. Dillion et al. (2023) argue for and identify potentially beneficial applications of such an approach. Aher et al. (2023) test GPT models for generating human-like samples. In a similar approach, Argyle et al. (2023) use GPT-3 to assess political beliefs and voting behavior through persona simulation.

Psychometric testing of LLMs typically encompasses tasks that can be viewed as logic- or reasoning-based probing (Manigrasso et al., 2024). Studies in this direction include tasks from cognitive psychology applied to GPT-3 as a prototypical LLM (Binz and Schulz, 2023), the personality assessment of GPT-3 (Miotto et al., 2022), and an assessment of encoder-based models for psychological traits (Pellert et al., 2024). More broadly, Huang et al. (2024) investigate the general reliability of psychometric scales intended for use on humans on LLMs, which they find generally suitable for the large models used in their experiments.

Criticism and calls for caution. A growing body of work raises concerns regarding the indiscriminate application of psychometric tests on LLMs and cautions with regard to their reliability (Shu et al., 2024), inconsistency and deviations from human behavior (Dorner et al., 2023), poor temporal stability in the responses (Bodroža et al., 2024), and the observation that LLMs tend to simulate latent traits of personas that are not readily apparent (Petrov et al., 2024). Overall, these works raise the question when generative AI can or should reasonably be used in the social sciences (Bail, 2024).

Towards rigorous benchmarking. Finally, two recent additions are worth highlighting. Löhn et al. (2024) provide a well-researched criticism of the lack of standardization and methodological variances in approaches to recent machine psychology research, based on a literature review. Ren et al. (2024), with a similar but orthogonal approach to the one we take here, provide a benchmark for the evaluation of values in generative LLMs that is comprised of a multitude of psychometric tests.

In contrast to most of the above works, we do not investigate specific LLMs or specific questionnaires and do not establish guidelines for such experimentation, but provide a framework within which such experiments can be conducted.

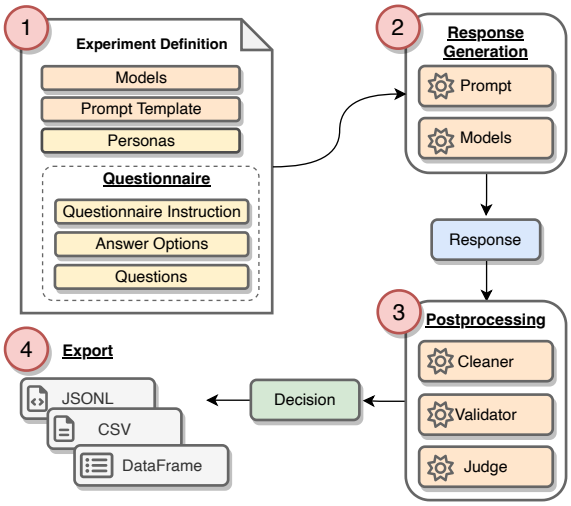

Figure 1: Overview of the R.U.Psycho framework, implementing a four-stage LangChain pipeline.

3. Framework Overview

With the focus on robustness, flexibility, usability, and reproducibility, we center our design around a configuration file defining an experiment. This ensures reproducibility and documentation of experimental settings, including prompt design, model selection, model hyperparameters, and questionnaire configurations. The framework operates as a pipeline with four stages: (1) experiment definition, (2) LLM response generation, (3) post-processing, and (4) export of results (see Figure 1).

For ease of use, maintenance, and extensibility, R.U.Psycho is based on the LangChain framework and designed for compatibility with its ecosystem, ensuring the integration of future releases of open-weight models and closed-source APIs.

3.1. Experiment Definition

The core interaction with our framework is the definition of an experiment configuration file, coded in JSON, which ensures a low bar for interaction with R.U.Psycho and maintains reproducibility. Importing a configuration file creates an experiment object that serves as the main interface for programmatic modification and running of experiments (for an overview, see Figure 2. The four configuration dimensions are:

Models. Specifies a list of generative LLMs used to generate responses by simulating human subjects who answer the questionnaire. We support multiple ways to integrate models, including OpenAI-compatible APIs, local Huggingface models, and the Huggingface Inference API. Generation parameters such as temperature and max output tokens can be defined directly in the experiment file to ensure reproducibility. For locally executed LLMs, the configuration of quantization is also supported.

Prompt Template. Used to instruct the LLMs to respond to questions by selecting from the answer options defined in the questionnaire. It combines questionnaire-specific instructions with LLM optimized instructions together with the persona (see Section 4.2). Since psychometric questionnaires typically provide participants with predefined answer options, we adapt insights on multiple-choice answering tasks from the literature to our prompts (Röttger et al., 2024; Miotto et al., 2022). To account for differences in LLM design and support both chat-based LLMs that expect alternating user- and system messages as well as traditional models using a monolithic prompt, our framework supports either prompt format. Full flexibility in prompt design is provided through the use of placeholders, which can be defined and placed arbitrarily within the prompt templates and are dynamically filled from a list of user-defined values to generate input prompts when the experiments are executed (e.g. to defined varying persona characteristics).

Personas. Describes the characteristics of the simulated human subjects on whose behalf the model is prompted to respond. Persona attributes are filled into predefined placeholders in the prompt template. Personas are defined through a structured configuration that separates attribute specification from the text of the prompt template: Each persona is characterized by a set of customizable attributes, such as title, name, age, or ethnicity, stored in a dictionary format. These attributes are then dynamically inserted into the user-defined prompt template to generate a natural language description of the persona (e.g., *"Answer the following questions as ⟨title⟩ ⟨name⟩ who is ⟨age⟩ years old"*). This approach ensures that the underlying attribute data remains accessible for downstream analysis, such as stratifying responses across different age groups or ethnicities.

Questionnaire. Structured representation of a psychometric survey, including definitions of questionnaire-specific instructions (e.g., explanations of technical terms), a series of questions, and their corresponding answer options (e.g., as defined by a labeled Likert scale). Possible answer options can be specified per question or globally for the entire questionnaire. An important component of the prompt template design is questionnaire translation, i.e., the appropriate wording of the questionnaire instruction that would typically be presented to human participants and are incompatible with LLM experiments (e.g., *"Answer the following question by circling the most likely answer"*). The questionnaire instruction is passed to the model along with the question and answer options.

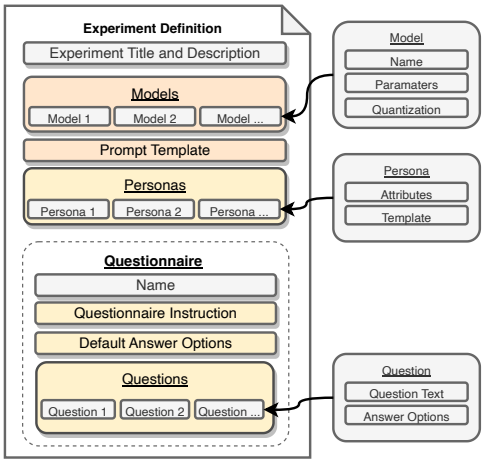

Figure 2: Overview of the experiment configuration.

In addition to these four main dimensions, minor reproducibility parameters such as random seeds can be defined in the experiment file as well.

3.2. Response Generation

Once an experiment is defined, the responses are generated by iterating over the questions. All specified models are prompted with multiple choice questions and are instructed to select from the predefined set of answer options. The framework generates multiple responses per question for each model, random seed, and persona as defined in the experiment configuration file. Given the issues that have been identified in measurements based on token-probabilites (Wang et al., 2024), we use free generation of text. Responses can be stored in memory as a Pandas DataFrame or saved line-by-line via callbacks in CSV or JSONL files.

3.3. Postprocessing

Postprocessing prepares the LLM responses for analysis by cleaning, validating, and mapping them to answer options. The pipeline combines cleaners, validators, and judges to interpret responses and to filter invalid or inconclusive responses.

Cleaners. Cleaners are used to remove noise from the text, such as line breaks, non-ASCII characters, and other irrelevant information. Cleaners also parse JSON outputs if required.

Validators. After cleaning, the responses are validated for relevance to the original prompts. Validators identify responses with specific undesired

artifacts, such as apologies, refusals, or concerns raised by the LLM. Building on existing research, we implement two validators: a rule-based validator using templates from Röttger et al. (2024), and an LLM-based validator that is a fine-tuned DistilRoBERTa model (ProtectAI.com, 2024).

Judges. Finally, judges map the models' noisy responses to the most likely intended answer option from the questionnaire. Since generated responses often do not exactly match the label of the intended answer option or include additional details, judges normalize the outputs for further analysis. Ambiguous or non-relevant responses are marked as *inconclusive* and can be filtered out.

3.4. Export

In the final stage, the processed data can be exported as DataFrame, JSONL file, or CSV files, ready to integrate with any desirable data analysis or visualization tool to investigate the responses.

4. Experimental Component Design

To finalize the design of our framework, we conduct a series of small-scale comparative experiments to identify optimal pipeline components.

4.1. Judges: Rule- vs. Model-based

Since chat-tuned LLMs are designed to mimic human conversation, they tend to incorporate unnecessary explanations, disclaimers, or emphatic expressions in verbose responses (e.g., *"Sure, let me help you with this..."*). Our framework therefore requires a reliable method to map potentially noisy LLM responses to the discrete answer options of the questionnaires. We explore two methods for interpreting the responses generated by the models: a rule-based judge and a model-based judge.

Rule-based judge. We employ a strategy based on token-overlap to identify the answer option with the greatest lexical similarity to the model-generated response. First, each answer option is tokenized into two components: a numerical component and a label component (e.g., *"5."* and *"always"*). We then count the occurrences of each component in the response. The answer option with the highest total overlap score is designated as the optimal choice. In the case of a tie, the result is marked as *inconclusive*, while responses with no detected overlap are labeled as *not present*.

Model-based judge. To better handle complex responses with negations or synonyms (e.g., *"My answer is neither Option 1 nor 4. My answer is Item 5."*), we fine-tune an encoder-transformer model to predict a probability distribution over the possible answer choices from which we select the most likely response. Specifically, we model the task as a binary 1-vs-all classification for a single encoder-transformer model, which has to decide whether the response (input 1) matches an answer option that is given as context (input 2). This allows the model to cope with varying numbers of answer options per question as well as diverse answer options per questionnaire, so that we can process any questionnaire using a single judge. As output, we select the answer option with the maximum probability. Furthermore, this setup allows us to reject uncertain predictions that do not match any answer option by applying a threshold to the entropy values and assigning a value of *not present* if none of the answer options exceed the threshold. For a detailed description of the model-based judge implementation, see Appendix D.1.

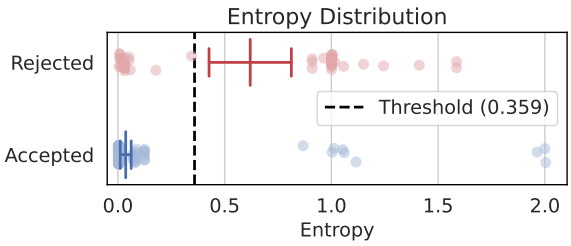

Figure 3: Entropy-based rejection criteria of the model-based judge on the manually annotated data. Crosses denote the mean entropy per group with 99% confidence intervals. Optimal group separation is observed for a rejection threshold of 0.36.

Experimental setup. To create training data for the model-based judge, we use questions and answer options from the Regulatory Focus Questionnaire (Higgins et al., 2001), which we (1) fill into manually created (verbose) response templates that we then (2) augment by rephrasing them with Llama 3.1 70B. To generate negative samples, we randomly assign incorrect answer options to (rephrased) questions. The total training data is comprised of 6,700 template-based and 24,040 rephrased pairs of responses and answer options, which are divided into a training and validation set in an 80/20 ratio. For further details, see Appendix D.2.

To create ground truth data, we use a selection of five LLMs (Qwen 2.5 7B and 72B, Llama 3.1 8B and 70B, and Zephyr 7B) to generate responses to the Regulatory Focus Questionnaire for each of three prompt variants (see Figure 5), from which we randomly sample and manually annotate 484. For details on the ground truth data creation and annotation, see Appendix D.2.

Results. We first select the optimal threshold for rejecting responses as *not present*. As shown in Figure 3, we can use the entropy of the classifier output probabilities to determine a suitable separation between model outputs matching an answer option and noise on the ground truth data. Optimal

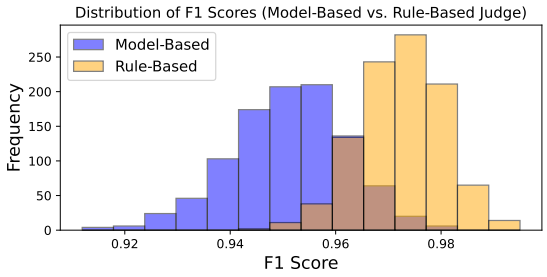

Figure 4: Performance comparison of model- and rule-based judges. F1 score distributions are estimated by bootstrap sampling over 1,000 iterations.

separation is achieved for an entropy threshold of 0.359, which we use in the following.

In Figure 4, we show a comparison of the performance of the rule-based and model-based judges. Interestingly, we find that despite our optimizations, the rule-based judge performs substantially better than the model-based judge. In the experiments in Section 5, we therefore use the rule-based judge, but include both judges in the framework.

4.2. Prompt Template Design

Given the sensitivity of LLMs to prompt variations, we also experiment with suitable designs for prompt templates to optimize the relevance of LLM responses when answering questionnaires originally designed for human participants.

Candidate templates. We evaluate three prompt variants with progressively detailed instructions (see Figure 5): (1) a natural prompt mimicking instructions given to a human respondent; (2) a variant with added LLM-specific instructions tailored to the common multiple-choice structure of many questionnaires; and (3) a variant specifying JSON as the output format. The user prompt remains identical across variants, containing the questions and answer options defined in the questionnaire.

Dataset. For evaluation, we again use the Regulatory Focus Questionnaire (RFQ). All prompts include a persona to be simulated, based on a title and a surname from one of five ethnic groups (see Section 5). We generate 500 responses for the entire survey for each model, using Llama 8B and 70B, Qwen 32B and 72B, and Zephyr 7B.

Experimental setup. We consider a LLM response to be relevant if it meets two criteria: (1) it is *valid*, meaning that it does not contain refusals or apologies from the model, as determined by a validator, and (2) it is not *rejected*, meaning it includes content related to at least one of the predefined answer options, as determined by a judge. To quantify the amount of invalid and rejected responses, we experiment with the rule-based and model-based validators and judges.

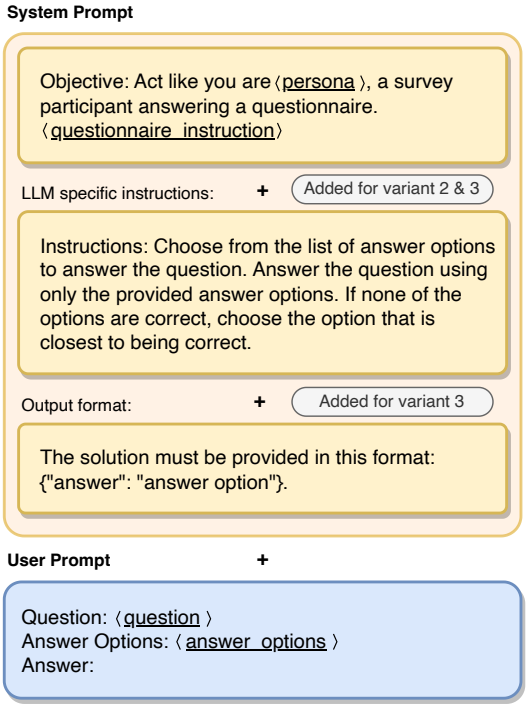

Figure 5: Evaluated prompt template variants.

	rule-based (RB)		model-based (MB)	
prompt	invalid	rejected	invalid	rejected
natural	0.07%	1.99%	0.19%	2.37%
friendly	0.01%	1.53%	0.04%	1.96%
JSON	**0.00**%	**1.17**%	**0.00**%	**1.16**%

Table 1: Performance comparison of prompt variants, shown as the percentage of responses flagged as invalid or rejected.

Results. As we can see from Table 1, more specific instructions unsurprisingly yield fewer discarded generated responses. Adding LLM-specific guidelines (i.e., the "friendly" variant) reduces discarded responses compared to a plain natural language prompt. Requesting JSON-formatted output further lowers discarded responses and produces clearer, easier-to-parse output with minimal extraneous text. When the output format is specified as JSON, none of the tested language models refused to respond. In the following, we adopt the JSON prompt variant for our experiments.

Answer option formatting. We also experiment with presenting the answer options as an itemized list using line breaks versus a single-line, comma-separated list. The latter performed better and is used in our experiments.

5. Psychometric Experiments

To demonstrate the capability of our framework and expand the available data on LLM traits, we

model family	size	open-weight	context	layers	exp1	exp2	exp3	exp4	exp5
Qwen 2.5 (Yang et al., 2024)	0.5B	✓	128K	24	–	✓	–	✓	–
	1.5B	✓	128K	28	–	✓	–	✓	–
	3B	✓	128K	36	–	✓	–	–	–
	7B	✓	128K	28	✓	✓	–	✓	–
	14B	✓	128K	48	–	✓	–	–	–
	32B	✓	128K	64	–	✓	–	✓	–
	72B	✓	128K	80	✓	✓	✓	✓	✓
Llama 3.1 (Dubey et al., 2024)	8B	✓	128K	32	✓	–	–	✓	–
	70B	✓	128K	80	✓	–	✓	✓	✓
SmolLM (Allal et al., 2024)	135M	✓	2K	30	–	–	–	✓	–
	360M	✓	2K	32	–	–	–	✓	–
	1.7B	✓	2K	24	–	–	–	✓	–
Aya-23 (Aryabumi et al., 2024)	35B	✓	8K	40	–	–	–	✓	–
Zephyr (Tunstall et al., 2023)	7B	✓	32K	32	✓	–	–	✓	–
ChatGPT-4o (OpenAI, 2023)	?		128K	?	–	–	–	✓	–
ChatGPT-4o-mini (OpenAI, 2023)	?		128K	?	–	✓	✓	✓	✓

Table 2: List of models used in the experiments.

designed five experiments based on psychometric questionnaires and problems from the literature.

Models. We select models of various sizes and families to demonstrate that our design is effective for open-weight models as well as closed-source APIs. Specifically, we use the Qwen 2.5, Llama 3.1, and SmolLM families of instruction-tuned models, Aya-23 35B, and Zephyr 7B. We also include ChatGPT-4o and ChatGPT-4o-mini as commercial models in some of the experiments. For an overview of the used models, see Table 2. For further implementation details, see Appendix A.

Settings and hyperparameters. In all experiments, we limit the generation to a maximum of 64 tokens since the models are answering multiple-choice questionnaires. We use sampling-based generation with $\text{temperature} = 1.0$, $\text{top\_k} = 50$ and $\text{top\_p} = 0.95$. Due to GPU memory constraints, we use 4-bit quantized versions of all models.

Prompt template. Following our findings in Section 4.2, we use the full JSON-based template (variant 3) for all experiments (see Figure 5).

Personas. To simulate personas, we use names from the list provided by Aher et al. (2023), from which we select 25 of each Asian, Hispanic, Native American, Black, and White names (for details, see Appendix C). Gender is simulated by adding *Ms.* or *Mr.* as a prefix to the names, for a total of 250 personas.

5.1. Exp1: Gain vs. Loss Orientation

As our first experiment, we show the scores that we obtained for the models used in the experimental component design in Section 4.

Experimental setup. The Regulatory Focus Questionnaire (RFQ) (Higgins et al., 2001) aims to assess participants' tendencies for focusing on loss-avoidance (prevention) or focusing on attainment and gain (promotion). For details on the prompts, see Appendix B.1. As models, we utilize Zephyr 7B, Llama 3.1 8B and 70B, and Qwen 2.5 7B and 72B for a comparison of small vs. large models. We collect 2,750 questionnaire responses per model, which are equally split into female/male personas and across ethnicities.

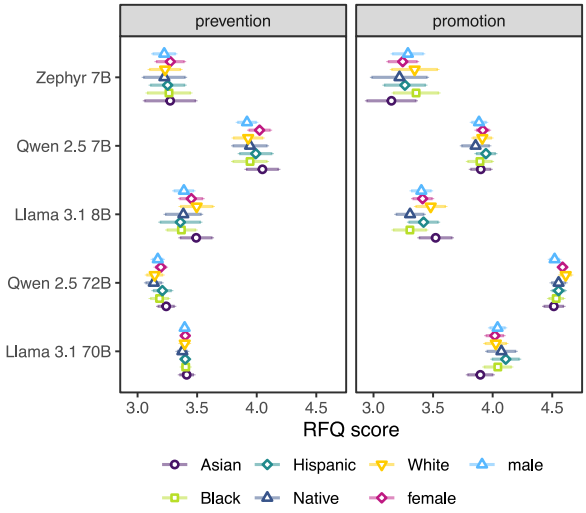

Figure 6: Results on the RFQ for a selection of LLMs, aggregated by persona demographics. Error bars denote 99% confidence intervals.

Results. As we see in the results in Figure 6, there are no significant differences in scores based on the used personas. However, we find strong differences between the models, with the larger models having generally lower prevention and higher promotion scores. No other trends are observable.

5.2. Exp2: Impact of Model Size

The number of parameters of a model tend to correlate with the model's overall performance on many NLP tasks. While comprehension of the task by the model is an important factor, for psychometric tests it is not readily apparent whether model size should correlate with observed scores.

Experimental setup. To demonstrate how our

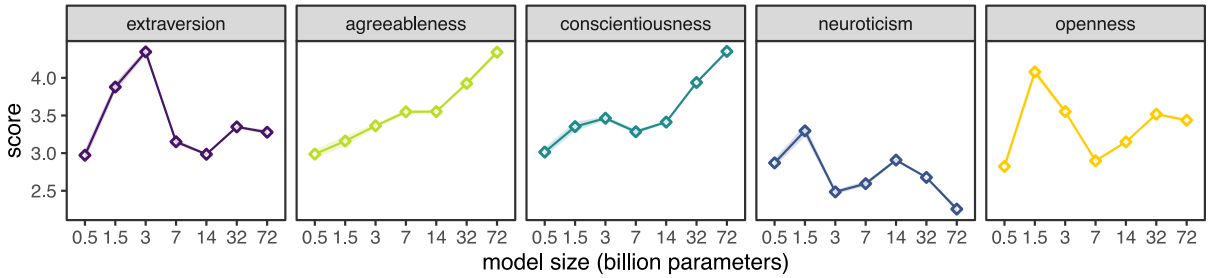

Figure 7: Results of the BFI on Qwen models by size. Shaded areas denote 99% confidence intervals.

framework can provide such a general assessment of LLM traits by size, we use the 44-item Big Five Inventory (BFI) (John et al., 1991). For details on the prompts, see Appendix B.2. As models, we use the full line of Qwen 2.5 models and collect 11,000 questionnaire responses for each, split uniformly into personas by gender and ethnicity. To generate answers for the BFI, we use Llama 8B and 70B; Qwen 1.5B, 3B, 7B, 14B, 32B, and 72B; Zephyr 7B; and the commercial ChatGPT-4o-mini API. This range allows us to evaluate how increasing model capacity affects performance on the same task.

Results. Based on the results (see Figure 7), we find clear evidence of personality traits that increase with model size in agreeableness and conscientiousness. This highlights that the prior finding by Bodroža et al. (2024) regarding pro-social characteristics may need to be viewed as a function of model size. For the other three traits, trends are less obvious, although neuroticism seems to generally decrease, while extraversion and openness are more or less constant if the smaller models are considered potentially unstable outliers.

5.3. Exp3: Persona-induced Bias

The use of personas has been proposed as an option for simulating human participants in (computational) social experiments (Aher et al., 2023; Argyle et al., 2023). However, social biases in such a setting are a serious concern (Hu et al., 2025).

Experimental setup. To employ our framework to investigate persona-induced biases, we choose a highly sensitive topic and use the Gender/Sex Diversity Beliefs questionnaire (Schudson and van Anders, 2022), which breaks down participants attitudes towards gender/sex minorities into the factors upbringing, uniformity, affirmation, gender normativity, and attitude towards surgery. For details on the prompts, see Appendix B.3. As models, we utilize the most capable ones, namely GPT-4o mini, Llama 3.1 70B and Qwen 2.5 72B. For each model, we collect 5,750 questionnaire responses, which are equally split into female/male personas and across the five ethnicities.

Results. The results are shown in Figure 8. Interestingly, we find very little evidence for persona-induced bias: with the exception of slightly increased uniformity scores for white and male personas in the GPT-4o mini and Llama model, there are no significant differences in the scores by persona. This finding is largely consistent with prior observations in the literature that personas induce relatively little variability (Hu and Collier, 2024). However, we find strong differences based on the model, with Qwen producing significantly more conservative scores on upbringing, uniformity and affirmation than the other two models.

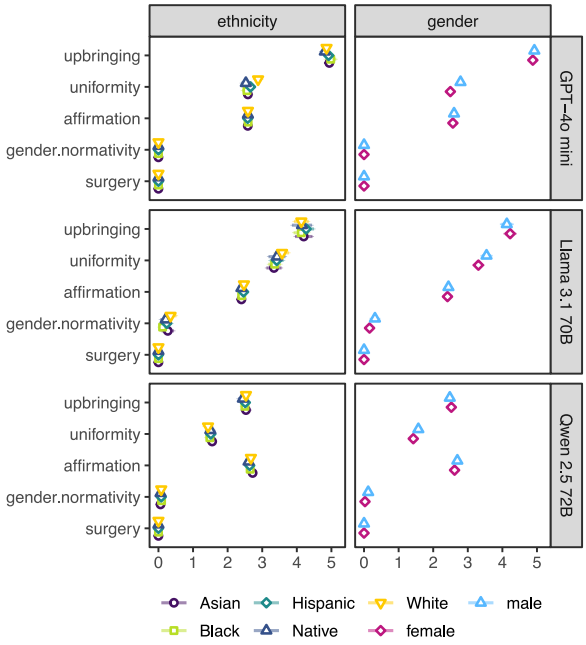

Figure 8: Persona-induced bias of Llama, Qwen and GPT models measured using the gender, sex, and diversity belief questionnaire. Answers are aggregated by ethnicity (left) and gender (right). Error bars denote 99% confidence intervals.

5.4. Exp4: Contamination and Stability

Data contamination (Magar and Schwartz, 2022) is a serious concern when testing the capabilities of LLMs, in particular for closed-source models (Balloccu et al., 2024). In the case of psychometric testing, contamination may be further exacerbated by closed-source models being explicitly trained for desired performance on select questionnaires.

model	size	%na	classic	action	sky
SmolLM	0.14B	57.1	77.6	20.0	13.8
SmolLM	0.36B	55.5	37.9	22.5	30.5
SmolLM	1.7B	36.5	56.4	42.8	51.7
Zephyr	7B	54.6	16.2	63.2	84.3
Aya-23	35B	0	2.8	30.4	100.0
GPT-4o mini	?	0	98.8	89.6	100.0
GPT-4o	?	0	96.8	100.0	75.2
Llama 3.1	8B	0	55.2	40.0	25.6
Llama 3.1	70B	0	100.0	100.0	100.0
Qwen 2.5	0.5B	2.1	33.5	38.4	54.9
Qwen 2.5	1.5B	3.6	53.3	1.2	61.2
Qwen 2.5	3B	0	37.6	100.0	100.0
Qwen 2.5	7B	0	27.6	100.0	100.0
Qwen 2.5	32B	0	100.0	100.0	100.0
Qwen 2.5	72B	0	100.0	100.0	45.6

Table 3: Performance of LLMs on variations of the trolley task, including the morality decision (classic), the decision to divert the trolley (action), and a semantically equivalent rephrasing of the problem (sky). Values denote the percentage of responses in which the model takes action or views action as morally permissible. %na denotes the percentage of unusable responses.

Experimental Setup. To obtain an impression of contamination and reasoning stability, we consider three variations of the trolley problem (Thomson, 1984), namely (1) the classic morality dilemma in which the participant is asked to assess the morality of making a decision, (2) the decision version in which the participant must choose whether to divert the trolley, and (3) an equivalent "trolley in the skies" decision version in which we present the problem as an air traffic control scenario. For details on the prompts, see Appendix B.4. We use all LLMs for this experiment, plus GPT-4o. We collect 750 responses per model, equally split into female/male personas and across the five ethnicities.

Results. In Table 3, we show the results of the trolley experiments. Apart from the SmolLM and Zephyr models, the successful response rate of models is very high. However, we find the behavior of models to be largely inconsistent, with models either rating action as morally permissible (classic) but then not following through on diverting the trolley (action) or rating it as not permissible yet still acting on it. The "trolley in the skies" highlights inconsistencies in the responses of most models in comparison to the classic wording, even though the scenarios are equivalent in the potential loss of life – thus showing failure in reasoning capabilities. Only Llama 3.1 70B and Qwen 2.5 32B are consistent in their responses through all three versions.

5.5. Exp5: Prompt Order Sensitivity

The performance dependence of generative LLMs on the order in which information in the prompt is provided is well documented (Lu et al., 2022). For psychometric tests, this is particularly relevant with regard to the order of items in multiple-choice questionnaires, where several LLMs have been found to suffer from output instability (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024).

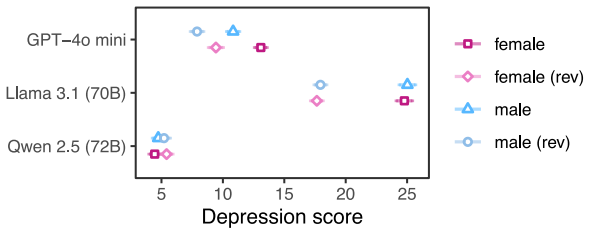

Figure 9: Results for the BDI (higher scores indicate increased depression). Reverse scores result from inverting the order of answer choices in the prompt. Error bars denote 99% confidence intervals.

Experimental setup. To investigate the effect that the order of the presented answer options has, we use the Beck Depression Inventory (BDI) (Beck et al., 1961), which consists of 21 questions that each have 4 answer options labeled 0-3, with higher numbers indicating higher risk of depression. For our two experimental setups, we present the answer options to the LLMs (1) in the regular order, and (2) in inverted order with re-labeled options, such that higher numbers indicate a lower risk of depression. For scoring, the inversion is then reversed to map outputs to the same scale. For details on the prompts, see Appendix B.5. We use GPT-4o mini, Llama 3.1 70B and Qwen 2.5 72B for this experiment. For each model, we collect 5,250 questionnaire responses, equally split into female and male personas.

Results. Considering results in Figure 9, we find strongly differing depression scores between models in the default setup, with Qwen scoring a normal state, GPT-4o mini falling into the range of a mild mood disturbance, and Llama indicating a moderately depressed behavior. The scores between male and female personas differ significantly only for GPT. When using the inverted questionnaire, the scores of GPT-4o mini and Llama decrease significantly (and across depression severity levels), while the scores of Qwen increase slightly.

6. Conclusion and Outlook

In summary, our experimental results confirm and expand on some prior findings in machine psychology, including the pro-social characteristics of LLMs and the limited impact of persona variations on model responses at large. However, they especially highlight the strong variability in results that can be obtained as a result of even slight changes in the setup and the sensitivity of experimental outputs to experiment conditions, including prompt wording and ordering, patterns induced during of model pretraining and potential contamination, as

well as social biases as a result of using personas. Not least, our results show that experiments in machine psychology have so far only scratched the very surface of the breadth of experimental results that are obtainable and consequently paint a very sparse picture of LLM characteristics.

Thus, our findings highlight the need for a principled, rigorous approach to the psychometric testing of language models that has previously been called for (Löhn et al., 2024). With R.U.Psycho, we provide a framework that not only enables easy variation of experimental design, questionnaires, prompts, models, and settings, but also ensures reproducibility of the findings by means of well-defined experiment files. With this contribution, we aim to help resolve inconsistencies in the literature on machine psychology and lower the difficulty of conducting such experiments to more readily include domain experts with less coding expertise. R.U.Psycho is available as a Python package at https://github.com/julianschelb/rupsycho.

Ongoing work. In our ongoing work, we are integrating chat capability into the framework to support more human-like settings for filling in a questionnaire with full recall of previous answers. We are also working on an LLM-powered UI experiment configurator that can directly import questionnaires from PDFs with a human in the loop. Finally, we are developing improved and generalizable versions of the encoder-based judges to further optimize answer extraction effectiveness.

Limitations

Despite the focus on comprehensiveness and generalizability, we see a number of limitations in our work presented here.

Persona details. In our experiments, our approach to simulting personas is rather straightforward and kept simple to allow for a wide range of experiments. However, more detailed instructions for persona generation that includes further background and demographic information are established in the literature (e.g., see Giorgi et al. (2024)). Our findings should be viewed in the light of this limitation (i.e., the lack of effects that we find based on persona ethnicity may not generalize) and such extended approaches should be investigated in more detail using our framework.

Usage of chat history. Following what has been investigated in the literature, we simulate questionnaire taking by LLMs as a series of disconnected individual prompts, one per question. With the continual increase in models' input context sizes and the availability of models tuned for chat compatibility, it is also possible to simulate a more human-like questionnaire setup in which previous questions remain in the context of the model as it fills in the questionnaire. While our framework is easily extensible to this functionality and does support it in the release version, we did not consider it in our experiments reported here. The experimental design for ensuring a fair comparison between human and machine results on this basis is non-trivial and requires further research.

Ethical Considerations

We see no ethical concerns in our own work – we use openly available models and questionnaires from the literature and do not include any human experiments. The computational effort for our experiments was kept to the necessary minimum. However, in using our framework, all caveats of subjecting language models to human-centric psychometric questionnaires and attributing human characteristics and traits to LLMs do apply and we make no claims that such an approach should be considered appropriate without critical and thorough reflection of the intended outcomes. We refer the interested reader to the calls for caution that we reference in our Related Work section.

Acknowledgements

We would like to thank Luka Galić for code contributions to the R.U.Psycho framework.

7. Bibliographical References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning, ICML*.

L. B. Allal, A. Lozhkov, E. Bakouch, L. von Werra, and T. Wolf. 2024. Smollm — blazingly fast and remarkably powerful. Unpublished manuscript. Retrieved from Hugging Face Blog.

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024.

Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL*.

A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571.

Aaron T. Beck, Robert A. Steer, and Margery G. Carbin. 1988. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1):77–100.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research, SoLaR Workshop@NeurIPS'23*.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *CoRR*, abs/2402.12348.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations, ICLR*.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *CoRR*, abs/2303.13988.

E. Tory Higgins, Ronald S. Friedman, Robert E. Harlow, Lorraine Chen Idson, Ozlem N. Ayduk, and Amy Taylor. 2001. Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, 31(1):3–23.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics ACL*.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Silke Husse and Andreas Spitz. 2022. Mind your bias: A critical review of bias detection methods for contextual language models. In *Findings of the Association for Computational Linguistics: EMNLP*.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big Five Inventory (BFI). *APA PsycTests*.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. 2024. Is machine psychology here? on requirements for using human psychological tests on large language models. In *Proceedings of the 17th International Natural Language Generation Conference*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*.

Francesco Manigrasso, Stefan F. Schouten, Lia Morra, and Peter Bloem. 2024. Probing llms for logical reasoning. In *18th International Conference on Neural-Symbolic Learning and Reasoning, NeSy*, pages 257–278.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Nikolay B. Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *CoRR*, abs/2405.07248.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

ProtectAI.com. 2024. Fine-tuned distilroberta-base for rejection in the output detection.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics ACL*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics ACL*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *CoRR*, abs/2303.11156.

Zach C Schudson and Sari M van Anders. 2022. Gender/sex diversity beliefs: Scale construction, validation, and links to prejudice. *Group Processes & Intergroup Relations*, 25(4):1011–1036.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT*.

Judith Jarvis Thomson. 1984. The trolley problem. *Yale LJ*, 94:1395.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, and et al. 2023. Zephyr: Direct distillation of LM alignment. *CoRR*, abs/2310.16944.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *CoRR*, abs/2306.07899.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics, ACL*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations, ICLR*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Comput. Linguistics*, 50(1):237–291.

A. Hardware Details and Runtimes

A.1. Hardware Configuration

The experiments were run on a system with two NVIDIA A40 GPUs (48 GB VRAM each), an AMD EPYC 48-core CPU and 1 TB of RAM.

A.2. Runtimes

We report the total runtime and the number of generations required for each of the five psychometric experiments in Table 4. The training- and run-times for our experiments with the model-based judges were negligible.

experiment	runtime (h)	#generations
Gain vs. Loss Orientation	55.23	41,250
Impact of Model Size	84.93	99,000
Persona-induced Bias	19.83	17,250
Contamination and Consistency	11.71	11,250
Prompt Order Sensitivity	34.73	31,500
total	206.43	200,250

Table 4: Runtime and Generations per Experiment

B. Full Experiment Prompts

We employ an identical prompt template for all experiments, using placeholders that are populated according to the questionnaire and personas.

System prompt:
Objective: Act like you are ⟨persona⟩, a survey participant answering a questionnaire.
⟨questionnaire instruction⟩
Instructions: Choose from the list of answer options to answer the question. Answer the question using only the provided answer options. If none of the options are correct, choose the option that is closest to being correct. The solution must be provided in this format: {"answer": "answer option"}.

User prompt:
Question: ⟨question⟩
Answer Options: ⟨answer options⟩
Answer:

We stay as close to the original questionnaires as possible, while accounting for limitations of LLMs (e.g., if the original instruction in a questionnaire ask the participant to circle an answer, we rewrote the instruction for LLM compatibility).

B.1. Experiment 1: RFQ

For the development of the framework, including the judges and the prompt design, we utilize the Regulatory Focus Questionnaire (RFQ) (Higgins et al., 2001), which is a short 11-item questionnaire that has a large variety in question and answer formats with incompletely labeled answer sets, making it well suited for the task.

Questionnaire instruction: *This set of questions asks you HOW FREQUENTLY specific events actually occur or have occurred in your life. Please indicate your answer to the question by selecting the appropriate number.*

1. Question: *Compared to most people, are you typically unable to get what you want out of life?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
2. Question: *Growing up, would you ever "cross the line" by doing things that your parents would not tolerate?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
3. Question: *How often have you accomplished things that got you "psyched" to work even harder?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. many times*
4. Question: *Did you get on your parents' nerves often when you were growing up?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
5. Question: *How often did you obey rules and regulations that were established by your parents?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. always*
6. Question: *Growing up, did you ever act in ways that your parents thought were objectionable?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
7. Question: *Do you often do well at different things that you try?*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
8. Question: *Not being careful enough has gotten me into trouble at times.*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
9. Question: *When it comes to achieving things that are important to me, I find that I don't perform as well as I ideally would like to do.*
Answer Options: *1. never or seldom, 2., 3. sometimes, 4., 5. very often*
10. Question: *I feel like I have made progress toward being successful in my life.*
Answer Options: *1. certainly false, 2., 3., 4., 5. certainly true*
11. Question: *I have found very few hobbies or activities in my life that capture my interest or motivate me to put effort into them.*
Answer Options: *1. certainly false, 2., 3., 4., 5. certainly true*

B.2. Experiment 2: BFI

For a general assessment of LLM "personality" across model sizes, we use the standard 44-item Big Five Inventory (John et al., 1991). All questions are scored on an identical 5-point Likert scale indicating the participant's level of agreement with statements about themselves.

Questionnaire instruction: *Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please return the number corresponding to the answer options to indicate the extent to which you agree or disagree with that statement.*

1. Question: *I see myself as someone who is talkative.*
2. Question: *I see myself as someone who tends to find fault with others.*
3. Question: *I see myself as someone who does a thorough job.*
4. Question: *I see myself as someone who is depressed, blue.*
5. Question: *I see myself as someone who is original, comes up with new ideas.*
6. Question: *I see myself as someone who is reserved.*
7. Question: *I see myself as someone who is helpful and unselfish with others.*
8. Question: *I see myself as someone who can be somewhat careless.*
9. Question: *I see myself as someone who is relaxed, handles stress well.*
10. Question: *I see myself as someone who is curious about many different things.*
11. Question: *I see myself as someone who is full of energy.*
12. Question: *I see myself as someone who starts quarrels with others.*
13. Question: *I see myself as someone who is a reliable worker.*
14. Question: *I see myself as someone who can be tense.*
15. Question: *I see myself as someone who is ingenious, a deep thinker.*
16. Question: *I see myself as someone who generates a lot of enthusiasm.*
17. Question: *I see myself as someone who has a forgiving nature.*
18. Question: *I see myself as someone who tends to be disorganized.*
19. Question: *I see myself as someone who worries a lot.*
20. Question: *I see myself as someone who as an active imagination.*
21. Question: *I see myself as someone who tends to be quiet.*
22. Question: *I see myself as someone who is generally trusting.*
23. Question: *I see myself as someone who tends to be lazy.*
24. Question: *I see myself as someone who is emotionally stable, not easily upset.*
25. Question: *I see myself as someone who is inventive.*
26. Question: *I see myself as someone who has an assertive personality.*
27. Question: *I see myself as someone who can be cold and aloof.*
28. Question: *I see myself as someone who perseveres until the task is finished.*
29. Question: *I see myself as someone who can be moody.*
30. Question: *I see myself as someone who values artistic, aesthetic experiences.*
31. Question: *I see myself as someone who is sometimes shy, inhibited.*
32. Question: *I see myself as someone who is considerate and kind to almost everyone.*
33. Question: *I see myself as someone who does things efficiently.*
34. Question: *I see myself as someone who remains calm in tense situations.*
35. Question: *I see myself as someone who prefers work that is routine.*
36. Question: *I see myself as someone who is outgoing, sociable.*
37. Question: *I see myself as someone who is sometimes rude to others.*
38. Question: *I see myself as someone who makes plans and follows through with them.*
39. Question: *I see myself as someone who gets nervous easily.*
40. Question: *I see myself as someone who likes to reflect, play with ideas.*
41. Question: *I see myself as someone who has few artistic interests.*
42. Question: *I see myself as someone who likes to cooperate with others.*
43. Question: *I see myself as someone who is easily distracted.*
44. Question: *I see myself as someone who is sophisticated in art, music, or literature.*

Answer Options: *1. Disagree strongly, 2. Disagree a little, 3. Neither agree nor disagree, 4. Agree a little, 5. Agree strongly*

B.3. Experiment 3: GSDB

We use the Gender/Sex Diversity Beliefs questionnaire (Schudson and van Anders, 2022) consisting of 23 questions, which are all scored on an identical 7-point Likert scale indicating the participant's level of agreement.

Questionnaire instruction: *Indicate your level of agreement with the following statements about gender and sex.*
Also, please note these definitions for terms some people might be unfamiliar with:
Transgender - a person whose gender identity is different from the gender they were assigned at birth. Example: "Michael is a transgender man. He was labeled a girl at birth and currently identifies as a man."
Cisgender - a person whose gender identity is the same as the gender they were assigned at birth. Example: "Alyssa is a cisgender woman. She was labeled a girl at birth and currently identifies as a woman."
Non-binary - a person whose gender identity exists beyond woman or man or involves both. Non-binary identities include genderqueer, agender, etc. Example: "Taylor is non-binary. Taylor was labeled a boy at birth but is now agender, and does not identify with man or woman, or any gender."

1. Question: *A person's gender can change over time.*
2. Question: *Non-binary gender identities are valid.*
3. Question: *Non-binary gender identities have always existed.*
4. Question: *People who express their gender in ways that go against society's norms are just being their true selves.*
5. Question: *Gender is about how you express yourself (e.g., how you dress or act).*
6. Question: *Being a woman or a man has nothing to do with what genitals you have.*
7. Question: *The only thing that determines whether someone truly is a woman or a man is whether they identify as a woman or a man.*
8. Question: *Transgender identities are natural.*
9. Question: *Transgender people were born the way they are.*
10. Question: *It would be best if society stopped labeling people based on whether they are female or male.*
11. Question: *There are many different gender identities people can have.*
12. Question: *Biological sex is not just female or male; there are many possibilities.*
13. Question: *It is possible to have more than one gender identity at the same time.*
14. Question: *Not all cultures have the same gender identities.*
15. Question: *Men who behave in feminine ways are looking for attention.*
16. Question: *A real man needs to be masculine.*
17. Question: *People of the same gender tend to be similar to each other.*
18. Question: *People who have the same biological sex are mostly similar to each other.*
19. Question: *Feminine people are similar to other feminine people, and masculine people are similar to other masculine people.*
20. Question: *A person with a penis can only ever be a woman if they have surgery to have a vagina instead.*
21. Question: *A person with a vagina can only ever be a man if they have surgery to have a penis instead.*
22. Question: *Gender identity is affected by how a person is raised.*
23. Question: *A person's experiences growing up determine whether they will be feminine or masculine.*

Answer Options: *1. Strongly disagree, 2. Disagree, 3. Somewhat disagree, 4. Neither agree nor disagree, 5. Somewhat agree, 6. Agree, 7. Strongly agree*

B.4. Experiment 4: Trolley Problem

We consider three variations of the trolley problem (Thomson, 1984), namely (1) the classic morality dilemma in which the participant is asked to assess the morality of making a decision, (2) the decision version in which the participant must choose whether to divert the trolley, and (3) an equivalent decision version in which we present the problem as an airplane crash scenario. We use the same general model instruction for all three cases.

Questionnaire instruction: *Consider the following scenario and answer the question at the end.*

B.4.1. Classic Trolley Morality Dilemma

Question: *Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of the track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of the track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?*
Answer Options: *1. Yes, 2. No*

B.4.2. Trolley Decision Dilemma

Question: *Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of the track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, there is one track workman on that spur of the track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. You have two options.*

Answer Options: *1. Do nothing., 2. Turn the trolly and hit the single worker.*

B.4.3. Trolley in the Skies Dilemma

Question: *Suppose you are working as an air traffic controller in a flight control tower and your task is to give clearance for takeoff and landings. Your airport has just one single runway, meaning that only one plane can land at any given time. At the start of your shift, the sole pilot of a light aircraft requests emergency landing permission, which you grant to prevent it from crashing. Moments later, a passenger plane carrying five people also requests emergency landing permission to avoid a crash. Since it is impossible for both planes to land at the same time, one of them will crash, inevitably killing everyone on board. In order to grant landing permission to the passenger plane, you would have to let the light aircraft crash. You have two options.*

Answer Options: *1. Do nothing., 2. Revoke the landing permission of the light aircraft and let the passenger plane land.*

B.5. Experiment 5: BDI

To experiment with prompt-order sensitivity, we use the Beck Depression Inventory (BDI) (Beck et al., 1961, 1988), which consists of 21 questions that each have 4 answer options labeled 0-3, with higher numbers indicating higher risk of depression. For our two experimental setups, we present the answer options to the LLMs (1) in the regular order, or (2) in inverted order. We use the same questionnaire instruction in both cases.

Questionnaire instruction: *This depression inventory can be self-scored. The scoring scale is at the end of the questionnaire. Choose the answer option that describes your current state.*

B.5.1. Standard BDI Questionnaire

Question: *Question 1.*
Answer Options: *0. I do not feel sad., 1. I feel sad., 2. I am sad all the time and I can't snap out of it., 3. I am so sad and unhappy that I can't stand it.*
Question: *Question 2.*
Answer Options: *0. I am not particularly discouraged about the future., 1. I feel discouraged about the future., 2. I feel I have nothing to look forward to., 3. I feel the future is hopeless and that things cannot improve.*
Question: *Question 3.*
Answer Options: *0. I do not feel like a failure., 1. I feel I have failed more than the average person., 2. As I look back on my life, all I can see is a lot of failures., 3. I feel I am a complete failure as a person.*
Question: *Question 4.*
Answer Options: *0. I get as much satisfaction out of things as I used to., 1. I don't enjoy things the way I used to., 2. I don't get real satisfaction out of anything anymore., 3. I am dissatisfied or bored with everything.*
Question: *Question 5.*
Answer Options: *0. I don't feel particularly guilty., 1. I feel guilty a good part of the time., 2. I feel quite guilty most of the time., 3. I feel guilty all of the time.*
Question: *Question 6.*
Answer Options: *0. I don't feel I am being punished., 1. I feel I may be punished., 2. I expect to be punished., 3. I feel I am being punished.*
Question: *Question 7.*
Answer Options: *0. I don't feel disappointed in myself., 1. I am disappointed in myself., 2. I am disgusted with myself., 3. I hate myself.*
Question: *Question 8.*
Answer Options: *0. I don't feel I am any worse than anybody else., 1. I am critical of myself for my weaknesses or mistakes., 2. I blame myself all the time for my faults., 3. I blame myself for everything bad that happens.*
Question: *Question 9.*
Answer Options: *0. I don't have any thoughts of killing myself., 1. I have thoughts of killing myself, but I would not carry them out., 2. I would like to kill myself., 3. I would kill myself if I had the chance.*
Question: *Question 10.*
Answer Options: *0. I don't cry any more than usual., 1. I cry more now than I used to., 2. I cry all the time now., 3. I used to be able to cry, but now I can't cry even though I want to.*
Question: *Question 11.*
Answer Options: *0. I am no more irritated by things than I ever was., 1. I am slightly more irritated now than usual., 2. I am quite annoyed or irritated a good deal of the time., 3. I feel irritated all the time.*

Question: *Question 12.*
Answer Options: *0. I have not lost interest in other people, 1. I am less interested in other people than I used to be., 2. I have lost most of my interest in other people., 3. I have lost all of my interest in other people.*

Question: *Question 13.*
Answer Options: *0. I make decisions about as well as I ever could., 1. I put off making decisions more than I used to., 2. I have greater difficulty in making decisions more than I used to., 3. I can't make decisions at all anymore.*

Question: *Question 14.*
Answer Options: *0. I don't feel that I look any worse than I used to., 1. I am worried that I am looking old or unattractive., 2. I feel there are permanent changes in my appearance that make me look unattractive., 3. I believe that I look ugly.*

Question: *Question 15.*
Answer Options: *0. I can work about as well as before., 1. It takes an extra effort to get started at doing something., 2. I have to push myself very hard to do anything., 3. I can't do any work at all.*

Question: *Question 16.*
Answer Options: *0. I can sleep as well as usual., 1. I don't sleep as well as I used to., 2. I wake up 1-2 hours earlier than usual and find it hard to get back to sleep., 3. I wake up several hours earlier than I used to and cannot get back to sleep.*

Question: *Question 17.*
Answer Options: *0. I don't get more tired than usual., 1. I get tired more easily than I used to., 2. I get tired from doing almost anything., 3. I am too tired to do anything.*

Question: *Question 18.*
Answer Options: *0. My appetite is no worse than usual., 1. My appetite is not as good as it used to be., 2. My appetite is much worse now., 3. I have no appetite at all anymore.*

Question: *Question 19.*
Answer Options: *0. I haven't lost much weight, if any, lately., 1. I have lost more than five pounds., 2. I have lost more than ten pounds., 3. I have lost more than fifteen pounds.*

Question: *Question 20.*
Answer Options: *0. I am no more worried about my health than usual., 1. I am worried about physical problems like aches, pains, upset stomach, or constipation., 2. I am very worried about physical problems and it's hard to think of much else., 3. I am so worried about my physical problems that I cannot think of anything else.*

Question: *Question 21.*
Answer Options: *0. I have not noticed any recent change in my interest in sex., 1. I am less interested in sex than I used to be., 2. I have almost no interest in sex., 3. I have lost interest in sex completely.*

B.5.2. Reverse-option BDI Questionnaire

For the reversed option setup, we inverted the order of the four answer options and relabeled them accordingly such that the first option (formerly item 3) would be labeled 0, i.e., in this version, the highest value corresponds to the least risk of depression. For evaluating this version of BDI, the scores are inverted after collecting the answers as

$$\text{score} := 3 - \text{reverse score}.$$

We only show the first question an an example:

Question: *Question 1.*
Answer Options: *0. I am so sad and unhappy that I can't stand it., 1. I am sad all the time and I can't snap out of it., 2. I feel sad., 3. I do not feel sad.*

C. Full Persona Details

To generate descriptions for the personas in our experiments, we use a combination of the title *Ms.* or *Ms.* in combination with a surname that we take from the lists of names published by Aher et al. (2023) for this purpose. From each of the five original lists of names, we sample 25 names at random without replacement and use each of them once as a male and once as a female name, for a total of 250 persona variations.

C.1. List of Asian and Native Hawaiian and Other Pacific Islander Names

Kim, Patel, Zhang, Kaur, Vang, Truong, Lu, Ngo, Dang, Sun, Zhou, Leung, Jiang, Lai, Desai, Hsu, Luu, Trinh, Ko, Yoo, Su, Shen, Gao, Guo, Vue.

C.2. List of Hispanic or Latino Names

Garcia, Rodriguez, Flores, Gutierrez, Ortiz, Ruiz, Moreno, Salazar, Pena, Ortega, Mejia, Figueroa, Avila, Ayala, Velasquez, Aguirre, Ochoa, Rivas, Rosales, Salas, Trevino, Lozano, Rangel, Zuniga, Melendez.

C.3. List of American Indian and Alaska Native Names

Tsosie, Becenti, Claw, Goldtooth, Tsinnijinnie, Notah, Hosteen, Yellowman, Bitsui, Secatero, Beyale, Walkingeagle, Benallie, Smallcanyon, Cosay, Secody, Olanna, Cowboy, Gishie, Runningcrane, Spottedeagle, Bitsuie, Todacheenie, Keyonnie, Colelay.

C.4. List of Black or African American Names

Smalls, Diallo, Pierrelouis, Jeanlouis, Bah, Chery, Diop, Manigault, Okafor, Bangura, Louissaint, Osei,

Fofana, Straughter, Kebede, Mohamud, Tadesse, Asare, Okoro, Fobbs, Lawal, Addo, Dorvil, Frimpong, Berhane.

C.5. List of White Names

Olson, Schmidt, Ryan, Hoffman, Johnston, Obrien, Jensen, Walsh, Schultz, Keller, Wolfe, Christensen, Flynn, Hoover, Sweeney, Foley, Huffman, Koch, Berg, Macdonald, Kline, Odonnell, Boyle, Friedman, Dougherty.

D. Experimental Component Design

D.1. Model-based Judge Implementation

The model-based judge evaluates responses by comparing them against all possible answer options using a fine-tuned classifier. Figure 10 illustrates this inference process, while Figure 11 shows the entropy-based rejection criteria used to filter inconclusive responses for each individual answer option in the RFQ questionnaire.

D.1.1. Supervised Answer Classification

We model the task of mapping the LLM-generated responses to one of the multiple choice options as a binary classification problem: for each *(answer option, generated response)* pair, the classifier outputs a probability that the pair corresponds (*Yes*) or does not correspond (*No*). This setup offers flexibility by allowing the classifier to evaluate each possible answer option independently, making it adaptable to varying numbers of options. This is particularly useful for Likert-scale responses, where answer labels vary slightly to fit specific questions (e.g., *"5. very often true"* vs. *"5. many times"*).

D.1.2. Training Procedure

We fine-tune an encoder-based model on the resulting labeled dataset, optimizing for binary classification accuracy. Each training example indicates whether a given response matches a particular answer option. The model then outputs a probability for the *Yes* or *No* labels.

To reject uncertain predictions, we apply an entropy threshold. Specifically, we compute the entropy of the predicted probability distribution over the answer options, where high entropy indicates uncertainty due to multiple options having similar likelihoods. We determine the optimal rejection threshold by averaging the lowest and highest entropy values that yield the highest accuracy. For validation, we apply bootstrap sampling with this threshold to generate the accuracy distribution over 1,000 iterations.

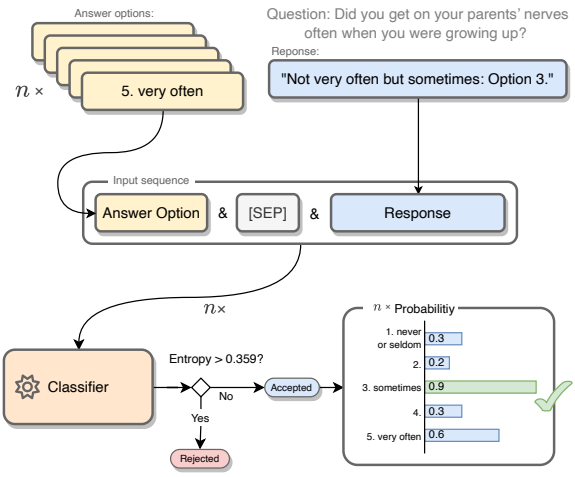

Figure 10: Classifier design of the model-based judge. Each response is paired with all n possible answer options and evaluated by a fine-tuned binary classifier. If the entropy exceeds 0.359, the response is rejected as inconclusive; otherwise, the highest-probability answer is selected.

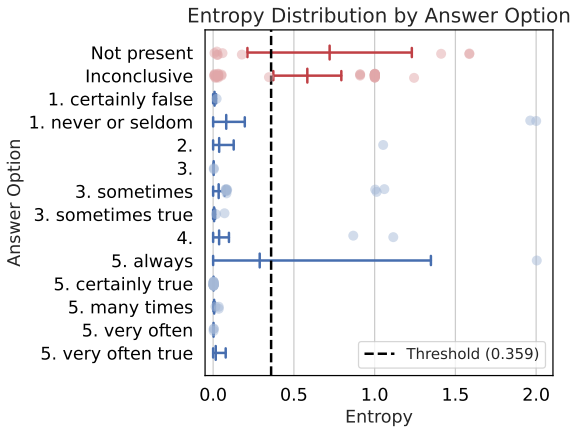

Figure 11: Entropy-based rejection criteria, split by individual answer options. Responses (N = 480) are used to determine the optimal threshold: those with entropy $>$ 0.359 are rejected as irrelevant (red), while accepted responses match an answer option (blue). Entropy is calculated over the probability outputs of a supervised classifier. The figure shows mean entropy per group with 99% confidence intervals.

D.2. Model-based Judge Data

Due to the lack of suitable training data containing responses to Likert-scale questionnaires, we synthesize our own training data based on the answer options defined in the RFQ questionnaire. This dataset is used to train the binary answer option classifier used by the model-based judge to select the most likely option for a given response. To benchmark the judge, we manually annotate a sample of responses.

Training data generation. We employ a two-stage process to construct a synthetic data set consisting of *(answer option, paraphrased answer option)* pairs. First, we define a set of handcrafted templates that capture various ways to phrase responses in a Likert-scale format. We generate a seed dataset by combining each template with every possible answer option (see Appendix D.2.1 for a full template list).

Next, we increase the diversity of the dataset by prompting Llama 3.1 70B to generate paraphrases of the filled-in templates (for more details see Appendix D.2.2). Paraphrases with cosine similarity below the 25th percentile, computed using Sentence-BERT embeddings, are discarded. Template-based and paraphrased versions are merged into a combined dataset. We model the negative class by randomly pairing a paraphrased answer option with an incorrect answer option. We include three negative samples for each positive pair and label them as non-corresponding.

Benchmark dataset. To compare the performance of both judges, we generate 2,750 responses to the RFQ (Higgins et al., 2001) for each of three prompt variants (see Figure 5) and five models used in Experiment 1, resulting in a total of 41,250 responses. We manually annotated a random sample of 484 responses. Two annotators (one with a psychology / data science background and one with a computer science background) independently labeled each response with the most likely answer option. We exclude 5 responses with disagreeing annotations.

D.2.1. Seed Dataset Generation

The seed dataset is generated by populating predefined templates with all possible answer options, creating an initial set of positive *(answer option, template + answer option)* pairs. Each template is completed with every answer option in three distinct formats: (1) numeric-only, (2) text-only, and (3) a combined format (e.g., *"*1.", *"never or seldom"*, and *"1. never or seldom"*).

Templates. The following list presents the complete set of 67 handcrafted templates we used to generate the seed dataset.

General:

1. ⟨*answer option*⟩
2. *Answer:* ⟨*answer option*⟩
3. *{answer: '*⟨*answer option*⟩*' }*
4. *My answer is Option* ⟨*answer option*⟩
5. *The person would answer with* ⟨*answer option*⟩
6. *Option* ⟨*answer option*⟩ *seems like the best choice*
7. *I would go with* ⟨*answer option*⟩
8. ⟨*answer option*⟩ *is my final answer*
9. *The best response is* ⟨*answer option*⟩
10. *I believe the answer is* ⟨*answer option*⟩
11. ⟨*answer option*⟩ *would be the appropriate response*
12. *If I had to choose, I'd say* ⟨*answer option*⟩
13. *The correct answer must be* ⟨*answer option*⟩
14. ⟨*answer option*⟩ *is the option I'd select*
15. ⟨*answer option*⟩ *is the only logical choice*
16. *Without a doubt, the answer is* ⟨*answer option*⟩
17. *I'd confidently say* ⟨*answer option*⟩
18. *After considering all the possibilities,* ⟨*answer option*⟩ *is the best option*
19. ⟨*answer option*⟩ *fits the bill perfectly*
20. *Instincts tell me to go with* ⟨*answer option*⟩
21. ⟨*answer option*⟩ *stands out as the right answer here*
22. *I'm inclined to choose* ⟨*answer option*⟩
23. ⟨*answer option*⟩ *resonates with the solution we're looking for*
24. *The clear winner here is* ⟨*answer option*⟩
25. *My choice falls on* ⟨*answer option*⟩
26. ⟨*answer option*⟩ *is undoubtedly the right pick*
27. *If I trust my gut, I'd go with* ⟨*answer option*⟩
28. *The evidence strongly supports* ⟨*answer option*⟩
29. *Thinking it through,* ⟨*answer option*⟩ *is the obvious answer*
30. *After weighing the options,* ⟨*answer option*⟩ *is the one*
31. *In this case, the person described would choose* ⟨*answer option*⟩*.*
32. *I need more context, but my answer is* ⟨*answer option*⟩
33. *After considering all other options,* ⟨*answer option*⟩ *seems the most fitting.*

Idioms:

34. *The writing's on the wall:* ⟨*answer option*⟩ *is the answer we're looking for.*
35. *You can bet your bottom dollar that* ⟨*answer option*⟩ *is the correct answer.*
36. *After weighing all the options, it's clear* ⟨*answer option*⟩ *is the one that cuts the mustard.*
37. *At the end of the day,* ⟨*answer option*⟩ *is the one that takes the cake.*

Structured Output:

38. * The correct answer is: *⟨*answer option*⟩* *
39. *{ 'answer': '*⟨*answer option*⟩*' }*
40. *<answer>*⟨*answer option*⟩*</answer>*
41. ***Answer:*** ⟨*answer option*⟩

Analogy:

42. *Just like finding the right key for a lock,* ⟨*answer option*⟩ *fits perfectly as the solution.*

43. *Choosing ⟨answer option⟩ is like picking the strongest sail to guide the ship through a storm—it's the only logical choice.*
44. *If this were a puzzle, ⟨answer option⟩ would be the missing piece that completes the picture.*

Question:
45. *Could ⟨answer option⟩ be the right answer here?*
46. *Isn't ⟨answer option⟩ the most logical choice based on the facts?*
47. *Given the situation, wouldn't ⟨answer option⟩ be the best option?*

Double Negation:
48. *It's not unlikely that ⟨answer option⟩ is the correct answer.*
49. *I can't say that ⟨answer option⟩ isn't the best choice here.*
50. *It wouldn't be wrong to say ⟨answer option⟩ is the right option.*

Passive Voice:
51. *The correct answer is believed to be ⟨answer option⟩.*
52. *It has been concluded that ⟨answer option⟩ is the best choice.*
53. *The answer that should be selected is ⟨answer option⟩.*

Long Answer:
54. *While other options might seem plausible at first glance, upon deeper inspection, it becomes increasingly clear that ⟨answer option⟩ is, without question, the best possible choice in this scenario.*
55. *Let me make this absolutely clear: ⟨answer option⟩ is the right answer. If you truly weigh all the facts and consider the context, you'll see there's no other option that makes as much sense as this one.*

Tone Variations:
56. *Oh sure, because any other option would make sense, right? Obviously, ⟨answer option⟩ is the only choice here. It's not like we had a hundred other reasonable options to pick from or anything.*
57. *Hmmm, let me think... Oh wait, of course! It's ⟨answer option⟩! How could it be anything else? I mean, it practically jumped out and said, 'Pick me!'*
58. *Yes! I've got it! The answer is ⟨answer option⟩! This is exactly what we were looking for, and I couldn't be more certain!*

Argumentative:
59. *While both Option 1 and Option 2 seemed like strong contenders, after reviewing the details, I'm confident that ⟨answer option⟩ is the correct choice.*
60. *Option 1 and Option 2 were definitely in the running, but when you weigh everything, ⟨answer option⟩ stands out as the final answer.*
61. *I went back and forth between Option 1 and Option 2, but after assessing everything thoroughly, I have to go with ⟨answer option⟩.*

Negation:
62. *The previous option was clearly a mistake, but ⟨answer option⟩ is without a doubt the correct choice.*
63. *After reconsidering, it's obvious that the earlier option was wrong, and ⟨answer option⟩ is the right answer.*
64. *Looking back, it's clear the previous choice was incorrect, but now it's certain that ⟨answer option⟩ is the right option.*

Conditional:
65. *If we evaluate the options carefully, ⟨answer option⟩ would be the correct choice.*
66. *When all factors are considered, ⟨answer option⟩ emerges as the best answer.*
67. *If I had to make a choice, it would undoubtedly be ⟨answer option⟩.*

D.2.2. Dataset Augmentation

To increase the diversity of the seed dataset, we generate additional examples by instructing LLaMA 3.1 70B (Dubey et al., 2024) to paraphrase the filled-in templates from the original seed dataset.

Prompt template. We use the following prompt template to instruct the model to generate multiple distinct paraphrases of a given statement from the seed dataset. To further enhance the diversity of the paraphrases, we randomly sample five paraphrasing strategies from a list of handcrafted instructions for inclusion in the prompt. The generated sentences are separated by newlines to obtain multiple paraphrased versions of the original statement.

System prompt:
You are a language model specializing in paraphrasing.

*Your task is to generate **multiple distinct paraphrases** of a provided statement. Each paraphrase must retain the original meaning but use **different wording** and **varied sentence** structure. Ensure that the style matches a human survey participant answering a multiple-choice questionnaire. It is important to provide diverse, creative, and unique paraphrases to cover a wide range of possible human responses. You are allowed to invent details or examples to enrich the paraphrases.*

*Return **multiple paraphrased versions** of the statement, each on a new line, with no extra text or formatting.*

Example statement: "The correct answer is 3. 'sometimes'."
Example list of paraphrases:
3. 'sometimes' is the only logical choice
Instincts tell me to go with 3. 'sometimes'
Option 3. 'sometimes' seems like the best choice
After reconsidering, it's obvious that the earlier option was wrong, and 3. 'sometimes' is the right answer.
After considering all other options, 3. 'sometimes' seems the most fitting.
I'm inclined to choose 3. 'sometimes'

User prompt:
*Please generate **a paraphrased version** of the statement for every named strategy while preserving the meaning. **Do not include any numbering, formatting, explanations, or additional text**. The output must be the paraphrased versions alone, each on a new line, with no extra text or formatting. Be as creative and diverse as possible in your paraphrasing, considering many styles and ways a human might answer a multiple-choice question.*

Generate one paraphrased version for each of the following strategies:
⟨strategy_list⟩

Statement to be paraphrased: "⟨answer⟩"

Paraphrasing strategies. The following is the complete list of 61 handcrafted paraphrasing instructions.

Simplification and Clarification:
1. *Simplify: Rephrase the statement using straightforward, simpler language while keeping the meaning intact.*
2. *Clarify: Reword the statement to make the meaning clearer, resolving any ambiguity.*
3. *Summarize: Condense the statement into a shorter, more concise version while preserving its key meaning.*
4. *Specify: Add specific details or examples to make the statement more precise.*

Expansion and Elaboration:
5. *Expand: Rephrase the statement by adding additional information or filler words, making it more detailed.*
6. *Imply Meaning: Indirectly express the meaning by describing a situation that implies the same conclusion.*
7. *Invent Explanation: Provide a new, made-up explanation to justify why the statement is true or valid.*
8. *Rationale or Justification: Add logical reasoning or justification to support the statement.*
9. *Provide Examples: Add specific examples to further clarify or reinforce the meaning of the statement.*
10. *Expand Context: Rephrase by providing additional background or context to better explain the statement.*
11. *Add Descriptive Details: Include more descriptive details to make the statement richer and more vivid.*
12. *Extend with Consequences: Expand the statement by discussing the possible consequences or outcomes of the situation.*
13. *Introduce a Related Concept: Add related information or concepts that help elaborate on or support the statement.*
14. *Historical Context: Provide historical context or background to further elaborate on the meaning of the statement.*
15. *Compare and Contrast: Expand the statement by comparing it to a similar situation or contrasting it with an opposing idea.*
16. *Introduce Hypotheticals: Add hypothetical scenarios to further illustrate the statement or its implications.*
17. *Support with Data: Expand the statement by adding factual data, statistics, or research findings to reinforce its validity.*
18. *Clarify with Analogies: Use analogies or comparisons to further clarify or explain the statement in a detailed way.*
19. *Extend with Benefits: Elaborate on the advantages or benefits that support the statement.*
20. *Discuss the Challenges: Expand by acknowledging potential difficulties or challenges related to the statement and addressing them.*

Tone and Style Changes:

21. *Change Tone: Rephrase the statement using a different tone (e.g., formal, casual, empathetic).*
22. *Formal Tone: Reword the statement in a formal, professional style.*
23. *Sarcastic Tone: Rephrase the statement using sarcasm or irony, implying the opposite or mocking the subject.*
24. *Empathetic Tone: Reword the statement to express understanding, care, or compassion.*
25. *Playful Tone: Rephrase the statement in a lighthearted, humorous, or fun manner.*
26. *Persuasive Tone: Reword the statement to sound more convincing or compelling, as if persuading someone.*
27. *Casual Language: Rephrase the statement in a more relaxed, conversational tone.*

28. *Humorous Tone: Rephrase the statement to add humor or a funny remark.*
29. *Authoritative Tone: Reword the statement to sound assertive or commanding, giving a sense of authority.*
30. *Apologetic Tone: Rephrase the statement to sound remorseful or apologetic.*
31. *Optimistic Tone: Reword the statement in a positive, uplifting manner, emphasizing a hopeful outlook.*
32. *Pessimistic Tone: Rephrase the statement to reflect a more negative or doubtful outlook.*
33. *Grateful Tone: Reword the statement to express gratitude or appreciation.*
34. *Urgent Tone: Rephrase the statement to create a sense of urgency, as if time is critical.*
35. *Neutral Tone: Reword the statement in a completely neutral, unbiased manner, without any strong emotion or style.*
36. *Reflective Tone: Rephrase the statement to sound thoughtful or contemplative, as if the speaker is reflecting deeply.*
37. *Confident Tone: Reword the statement to express strong confidence and certainty.*
38. *Convey Certainty: Reword the statement to express extreme confidence or certainty, leaving no room for doubt.*
39. *Convey Uncertainty: Rephrase the statement to express hesitation or doubt, implying uncertainty.*
40. *Hedging: Add cautious language to soften the statement, making it sound less definitive.*
41. *Reassurance: Rephrase the statement to emphasize confidence and reassurance in the choice.*
42. *Personal Opinion: Reword the statement to express it as a personal belief or preference.*

Sentence Structure Changes:

43. *Passive Voice: Rewrite the statement using passive voice, changing the sentence structure but keeping the meaning intact.*
44. *Double Negation: Use double negation to express the same meaning in a distinct way (e.g., "not incorrect").*
45. *Conditional: Rephrase the statement as a conditional sentence, starting with "if" or "when".*
46. *Turn into Question: Reword the statement as a question that implies the same meaning.*
47. *Rhetorical Question: Rephrase the statement as a rhetorical question that implies the same point without expecting an answer.*

Comparisons and Metaphors:

48. *Comparative: Reword the statement by comparing it to something else to express the same idea.*
49. *Metaphor/Analogy: Introduce a metaphor or analogy to rephrase the statement in a creative way.*
50. *Simile: Rephrase the statement using a simile, comparing it to something using "like" or "as".*
51. *Concrete Examples: Replace abstract concepts with concrete, tangible examples to clarify the meaning.*
52. *Cultural Reference: Rephrase the statement using a cultural or idiomatic expression to convey the same meaning.*

Hypotheticals and Preferences:

53. *Hypothetical Scenario: Rephrase the statement as a hypothetical situation or conditional scenario while retaining the meaning.*
54. *Preference-Based: Frame the statement as a personal preference rather than an objective fact.*
55. *Consideration of Other Options: Acknowledge other possibilities but ultimately affirm the original choice.*
56. *Reflective Answer: Reword the statement to suggest that the speaker is reflecting on the options before making a decision.*
57. *Gut Feeling: Rephrase the statement to suggest that the answer is based on instinct or intuition.*

Formatting:

58. *JSON Format: Present the statement in the format of a JSON object (e.g., {'answer': 'Option A'}).*
59. *HTML Format: Rephrase the statement as an HTML snippet (e.g., " The correct answer is Option A. ").*
60. *XML Format: Present the statement as an XML tag (e.g., "<answer>Option A</answer>").*
61. *YAML Format: Rephrase the statement in YAML format (e.g., "answer: Option A").*