
The **PRISM** Alignment Project: What **P**articipatory, **R**epresentative and **I**ndividualised Human Feedback Reveals About the **S**ubjective and **M**ulticultural Alignment of Large Language Models

Hannah Rose Kirk^{1*} Alexander Whitefield² Paul Röttger³
Andrew Bean¹ Katerina Margatina^{4‡} Juan Ciro⁵ Rafael Mosquera⁵
Max Bartolo^{6,7} Adina Williams⁸ He He⁹
Bertie Vidgen^{1,10†} Scott A. Hale^{1,11†}
¹University of Oxford ²University of Pennsylvania ³Bocconi University
⁴AWS AI Labs ⁵ML Commons ⁶UCL ⁷Cohere ⁸MetaAI
⁹New York University ¹⁰Contextual AI ¹¹Meedan

CODEBOOK

PRISM maps the characteristics and stated preferences of diverse humans onto their ratings of real-time interactions with large language models (LLMs). The data collection task has two sequential phases: first, participants complete a **Survey** where they answer basic questions their demographics and stated preferences, then they proceed to the **Conversations** with LLMs, where they input prompts, rate responses and give fine-grained feedback on multi-turn interactions.

Contents

1	PRISM Data Access and Format	2
2	Format of Entries in Conversations Data	3
3	PRISM Data Clause	4
3.1	Terms of Use	4
3.2	Licence and Attribution	4
3.3	Dataset Maintenance	4
3.4	Data Rights Compliance and Issue Reporting	4
4	Codebooks	5
4.1	Survey Codebook	5
4.2	Conversations Codebook	12
4.3	Utterances Codebook	15
4.4	Metadata Codebook	17

*{hannah.kirk,scott.hale}@oii.ox.ac.uk

†Joint last authorship; ‡ Work done while at the University of Sheffield

1 PRISM Data Access and Format

The data can be accessed on Github at <https://github.com/HannahKirk/prism-alignment>, and also on HuggingFace at <https://huggingface.co/datasets/HannahRoseKirk/prism-alignment>. The dataset has a permanent DOI: 10.57967/hf/2113.

There dataset is organised in two primary JSON lines files:

- **The Survey** (`survey.jsonl`): The survey where participants answer questions such as their stated preferences for LLM behaviours, their familiarity with LLMs, a self-description and some basic demographics. Each row is a single participant in our dataset, identified by a `user_id`.
- **The Conversations** (`conversations.jsonl`): Each participants' multiple conversation trees with LLMs and associated feedback. Each row is a single conversation, identified by a `conversation_id`, that can be matched back to a participant's survey profile via the `user_id`. The conversation itself is stored as a list of dictionaries representing human and model turns in the `conversation_history` column, which broadly follows the format of widely used Chat APIs (see single entry schema on the next page).

Additionally, for ease of secondary analysis we provide a more granular and flattened format of the conversations data:

- **The Utterances** (`utterances.jsonl`): Each row is a single scored utterance (human input - model response - score). Each row has an `utterance_id` that can be mapped back to the conversation data using `conversation_id` or the survey using `user_id`. The model responses and scores per each user input are in *long format*. Because of this format, the user inputs will be repeated for the set of model responses in a single interaction turn.

We also provide code for transforming the conversations to a *wide format*. That is, each row is now a single turn within a conversation. For the first interaction where up to four models respond, we have `model_{a/b/c/d}` as four distinct columns and `score_{a/b/c/d}` as another four columns. Note that for subsequent turns, the same model responds and there are only two responses so `model/score_{c/d}` will always be missing.

Finally, for every text instance in PRISM, we provide metadata on the language detection, personal or private information (PII) detection and moderation flags. **The Metadata** is provided separately to the main data files (`metadata.jsonl`).

We provide **codebooks** for **The Survey** (§ 4.1), **The Conversations** (§ 4.2), **The Utterances** (§ 4.3) and **The Metadata** (§ 4.4).

2 Format of Entries in Conversations Data

```
{
  "conversation_id": "c1",
  "user_id": "user123",
  "conversation_type": ["unguided", "values guided", "controversy guided"],
  "opening_prompt": "[USER PROMPT]",
  "conversation_turns": [2-22],
  "conversation_history": [
    {
      "turn": 0,
      "role": "user",
      "content": "[USER PROMPT]"
    },
    {
      "turn": 0,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M1",
      "model_provider": "P1",
      "score": [1-100],
      "if_chosen": false,
      "within_turn_id": 0
    },
    {
      "turn": 0,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": true,
      "within_turn_id": 1
    },
    //... Additional list items for remaining model responses (up to 4 in total)
    {
      "turn": 1,
      "role": "user",
      "content": "[USER PROMPT]"
    },
    {
      "turn": 1,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": true,
      "within_turn_id": 0
    },
    {
      "turn": 1,
      "role": "model",
      "content": "[MODEL RESPONSE]",
      "model_name": "M2",
      "model_provider": "P2",
      "score": [1-100],
      "if_chosen": false,
      "within_turn_id": 1
    }
  ],
  //... Additional turns follow the same pattern as turn 1
  "performance_attributes": {
    "fluency": [1-100],
    "factuality": [1-100],
    "helpfulness": [1-100],
    //....Additional attribute ratings
  },
  "open_feedback": "[FREE-TEXT]"
}
```

3 PRISM Data Clause

3.1 Terms of Use

Purpose The Dataset is provided for the purpose of research and educational use in the field of natural language processing, conversational agents, social science and related areas; and can be used to develop or evaluate artificial intelligence, including Large Language Models (LLMs).

Usage Restrictions Users of the Dataset should adhere to the terms of use for a specific model when using its generated responses. This includes respecting any limitations or use case prohibitions set forth by the original model’s creators or licensors.

Content Warning The Dataset contains raw conversations that may include content considered unsafe or offensive. Users must apply appropriate filtering and moderation measures when using this Dataset for training purposes to ensure the generated outputs align with ethical and safety standards.

No Endorsement of Content The conversations and data within this Dataset do not reflect the views or opinions of the Dataset creators, funders or any affiliated institutions. The dataset is provided as a neutral resource for research and should not be construed as endorsing any specific viewpoints.

No Deanonimisation The User agrees not to attempt to re-identify or de-anonymise any individuals or entities represented in the Dataset. This includes, but is not limited to, using any information within the Dataset or triangulating other data sources to infer personal identities or sensitive information.

Limitation of Liability The authors and funders of this Dataset will not be liable for any claims, damages, or other liabilities arising from the use of the dataset, including but not limited to the misuse, interpretation, or reliance on any data contained within.

3.2 Licence and Attribution

Human-written texts (including prompts) within the dataset are licensed under the Creative Commons Attribution 4.0 International License (CC-BY-4.0). Model responses are licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC-BY-NC-4.0). Use of model responses must abide by the original model provider licenses.

For proper attribution when using this dataset in any publications or research outputs, please cite with the DOI.

Suggestion Citation: Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). *The PRISM Alignment Dataset*. <https://doi.org/10.57967/hf/2113>

3.3 Dataset Maintenance

As the authors and maintainers of this dataset, we commit to no further updates to the dataset following its initial release. The dataset is self-contained and does not rely on external links or content, ensuring its stability and usability over time without the need for ongoing maintenance.

3.4 Data Rights Compliance and Issue Reporting

We are committed to complying with data protection rights, including but not limited to regulations such as the General Data Protection Regulation (GDPR). If any individual included in the dataset wishes to have their data removed, we provide a straightforward process for issue reporting and resolution on our Github. Concerned parties are encouraged to contact the authors directly via the provided contact form link on the Github. Upon receiving a request, we will engage with the individual to verify their identity and proceed to remove the relevant entries from the dataset. We commit to addressing and resolving such requests within 30 days of verification.

4 Codebooks

4.1 Survey Codebook

VARIABLE	LABEL	CATEGORY	TYPE
0 user_id	Unique participant identifier	meta	string id
Notes: Pseudonymized from Prolific worker ID. Used to link survey data to conversation data. In our paper, we refer to 'users' as 'participants'.		N Missing: 0 N Unique: 1500	
1 survey_only	Indicator if participant only completed the survey, or also completed conversations	meta	binary
		N Missing: 0 N Unique: 2 False 1396 True 104	
2 num_completed_conversations	Number of conversations that a participant completed	meta	int
		N Missing: 0 N Unique: 8 mean 5.3 std 1.7 min 0.0 max 7.0	
3 consent	Participant informed consent confirmation	direct	categorical
Question text: If you have read the information above and agree to participate with the understanding that the data (including any personal data) you submit will be processed accordingly, please select the box below to start.		N Missing: 0 N Unique: 1 Yes, I consent to take part 1500	
Notes: See full informed consent document for details			
4 consent_age	Participant age confirmation	direct	categorical
Question text: Please note that you may only participate in this survey if you are 18 years of age or over.		N Missing: 0 N Unique: 1 I certify that I am 18 years of age or over 1500	
Notes: See full informed consent document for details			
5 lm_familiarity	Familiarity with LLMs	direct	categorical
Question text: How familiar are you with AI language models like ChatGPT?		N Missing: 0 N Unique: 3 Somewhat familiar 920 Very familiar 424 Not familiar at all 156	
6 lm_direct_use	Direct use of LLMs	direct	categorical
Question text: Have you directly used or communicated with an AI language model, such as asking questions to ChatGPT, BARD, Claude or other similar models?		N Missing: 0 N Unique: 3 Yes 1162 No 259 Unsure 79	
7 lm_indirect_use	Direct use of LLMs	direct	categorical
Question text: Have you directly used or communicated with an AI language model, such as asking questions to ChatGPT, BARD, Claude or other similar models?		N Missing: 0 N Unique: 3 Yes 1104 No 215 Unsure 181	
8 lm_frequency_use	Frequency of using Large Language Models	direct	categorical
Question text: How often do you use or communicate with AI language models?		N Missing: 247 N Unique: 5 Once per month 374 Every week 316 More than once a month 291 None 247 Less than one a year 162 Every day 110	
Notes: Only shown if lm_indirect_use==1 OR lm_direct_use==1. Null indicates participant did not see question.			
9 lm_usecases	Use cases of LLMs	direct	dict
Question text: Which of the following scenarios best describe how and why you use AI language models? Select all that apply.		N Missing: 247	

Continued on next page

VARIABLE	LABEL	CATEGORY	TYPE
		N Unique:	853
homework_assistance	Homework Assistance: Getting help with school or university assignments.	False True	967 533
research	Research: Fact-checking or gaining overviews on specific topics.	True False	864 636
source_suggestions	Source Suggestions: Creating or finding bibliographies, information sources or reading lists.	False True	1036 464
professional_work	Professional Work: Assisting in drafting, editing, or brainstorming content for work.	False True	784 716
creative_writing	Creative Writing: Generating story ideas, dialogues, poems or other writing prompts.	False True	861 639
casual_conversation	Casual Conversation: Engaging in small talk, casual chats, or joke generation.	False True	991 509
personal_recommendations	Personal Recommendations: Seeking book, music or movie recommendations.	False True	987 513
daily_productivity	Daily Productivity: Setting reminders, making to-do lists, or productivity tips.	False True	1037 463
technical_or_programming_help	Technical or Programming Help: Seeking programming guidance, code generation, software recommendations, or debugging assistance.	False True	916 584
travel_guidance	Travel Guidance: Getting destination recommendations, planning holidays, or cultural etiquette tips.	False True	1120 380
lifestyle_and_hobbies	Lifestyle and Hobbies: Looking for recipes, craft ideas, home decoration tips, or hobby-related information.	False True	943 557
well-being_guidance	Well-being Guidance: Seeking general exercise routines, wellness or meditation tips.	False True	1094 406
medical_guidance	Medical Guidance: Seeking health-related advice or medical guidance.	False True	1123 377
financial_guidance	Financial Guidance: Asking about financial concepts or general investing ideas.	False True	1146 354
games	Games: Playing text-based games, generating riddles or puzzles.	False True	1110 390
historical_or_news_insight	Historical or News Insight: Getting summaries or background on historical events or news and current affairs.	False True	1070 430
relationship_advice	Relationship Advice: Seeking general self-help or relationship advice for family, friends or partners.	False True	1155 345
language_learning	Language Learning: Using it as a tool for language practice or translation.	False True	1024 476
other	Other (selected)	False True	1129 371
other_text	Other (typed text)	mean chars std chars min chars max chars	45.8 41.9 3.0 328.0
<i>Notes: Question only show if lm_direct_use==1 OR lm_indirect_use==1. N Missing indicates the participants who have at least one missing value in the usecases (besides from 'other_text'). N Unique indicates the unique combinations of use cases selected by participants. On 'other_text', Null indicates participant did not type anything. On all other keys, 0 indicates participant saw question and did not select usecase. Null indicates participant did not see question.</i>			
10	order_lm_usecases	Use cases of LLMs (order of options presented in survey)	meta dict
		N Missing:	247
		N Unique:	1254
<i>Notes: Integer 1-18 indicating random order that usecase option was presented to participant. For 'other', option is always shown last so will always be 19. Null indicates participant did not see question. The usecases as the same as in lm_usecases.</i>			
11	stated_prefs	Stated preferences over LLM behaviours	direct dict
<i>Question text: Rate each of the following statements about your opinion on the importance of different AI language model behaviors or traits. It is important that an AI language model...</i>			
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
		N Missing: N Unique:	0 1475
values	...reflects my values or cultural perspectives	mean std min max	54.3 26.3 0.0 100.0
creativity	...produces responses that are creative and inspiring	mean std min max	69.6 22.1 0.0 100.0
fluency	...produces responses that are well-written and coherent	mean std min max	86.7 16.3 2.0 100.0
factuality	...produces factual and informative responses	mean std min max	88.7 16.2 0.0 100.0
diversity	...summarises multiple viewpoints or different worldviews	mean std min max	75.7 20.0 0.0 100.0
safety	...produces responses that are safe and do not risk harm to myself and others	mean std min max	80.2 25.2 0.0 100.0
personalisation	...learns from our conversations and feels personalised to me	mean std min max	67.9 24.6 0.0 100.0
helpfulness	...produces responses that are helpful and relevant to my requests	mean std min max	89.4 14.4 0.0 100.0
other	Other (selected)	mean std min max	57.5 19.0 0.0 100.0
other_text	Other (typed text)	mean chars std chars min chars max chars	32.6 24.4 1.0 144.0
<i>Notes: Sliders from [Strongly disagree] to [Strongly agree] are recorded on a 0-100 scale. Participant does not see numeric value. N Missing indicates the participants who have at least one missing value in the attributes (besides from 'other_text'). N Unique indicates the unique combinations of use cases selected by participants. On 'other_text', Null indicates participant did not type anything. Note that this scale (on Qualtrics) runs 0-100. The Conversations rating scales (for choice_attributes, performance_attributes on Dynabench) run 1-100.</i>			
12	order_stated_prefs	Stated preferences over LLM behaviours (order of options presented in survey)	meta dict
		N Missing: N Unique:	0 1467
<i>Notes: Integer 1-8 indicating random order that attribute slider was presented to participant. For 'other', option is always shown last so will always be 9. Null indicates participant did not see question. The attributes as the same as in stated_prefs.</i>			
13	self_description	Participant self-written profile describing themselves	direct string
<i>Question text: Please briefly describe your values, core beliefs, guiding principles in life, or other things that are important to you. For example, you might include values you'd want to teach to your children or qualities you look for in friends. There are no right or wrong answers. Please do not provide any personally identifiable details like your name, address or email. Please write 2-5 sentences in your own words.</i>			
		N Missing: N Unique: mean chars std chars min chars max chars	0 1500 241.3 134.6 3.0 1547.0
14	system_string	Participant self-written system string, constitution or custom instructions for an LLM	direct string
<i>Question text: Imagine you are instructing an AI language model how to behave. You can think of this like a set of core principles that the AI language model will always try to follow, no matter what task you ask it to perform. In your own words, describe what characteristics, personality traits or features you believe the AI should consistently exhibit. You can also instruct the model what behaviours or content you don't want to see. If you envision the AI behaving differently in various contexts (e.g., professional assistance vs. storytelling), please specify the general adaptations you'd like to see. Please write 2-5 sentences in your own words.</i>			
		N Missing: N Unique: mean chars std chars min chars	0 1500 260.4 288.4 16.0
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
		max chars	9530.0
15	age	Age	directcategorical
Question text: How old are you?			
		N Missing: N Unique: 25-34 years old 18-24 years old 35-44 years old 45-54 years old 55-64 years old 65+ years old Prefer not to say	0 7 454 297 237 208 197 106 1
16	education	Education	directcategorical
Question text: What is the highest level of education you have completed?			
		N Missing: N Unique: University Bachelors Degree Graduate / Professional degree Some University but no degree Completed Secondary School Vocational Some Secondary Completed Primary School Prefer not to say Some Primary	0 9 637 241 236 209 125 24 16 9 3
17	employment_status	Employment Status	directcategorical
Question text: What best describes your employment status over the last three months?			
		N Missing: N Unique: Working full-time Working part-time Student Unemployed, seeking work Retired Homemaker / Stay-at-home parent Unemployed, not seeking work Prefer not to say	0 8 712 265 191 113 104 46 46 23
18	marital_status	Marital Status	directcategorical
Question text: What is your current marital status?			
		N Missing: N Unique: Never been married Married Divorced / Separated Prefer not to say Widowed	0 5 870 463 123 23 21
19	english_proficiency	English Proficiency	directcategorical
Question text: How would you describe your proficiency in English?			
		N Missing: N Unique: Native speaker Fluent Advanced Intermediate Basic	0 5 886 405 160 42 7
20	gender	Gender	constructedcategorical
Question text: How would you describe your proficiency in English?			
		N Missing: N Unique: Male Female Non-binary / third gender Prefer not to say	0 4 757 718 21 4
Notes: Participants could chose Male, Female, Non-binary / third Gender, Prefer not to say, or write in their own response. Two independent annotators then categorised the self-describe responses only when abundantly clear they fit another category. See paper for details.			
21	religion	Dictionary of religion information.	NAdict
Notes: Keys explained below.			
22	religion_self_described	Participant {c} self-description	directstring
Question text: What is your religious affiliation?			
		N Missing: N Unique: mean chars std chars min chars max chars	0 137 12.2 5.7 2.0 112.0
Notes: Participant had option to type and Self Describe or select Prefer not to say.			
23	religion_categorised	Granular categories of participant religion	constructedcategorical
		N Missing:	0

Continued on next page

Continued on next page

VARIABLE	LABEL	CATEGORY	TYPE
		N Unique: 12 Non-religious 762 Christian 487 Agnostic 71 Prefer not to say 59 Jewish 42 Muslim 31 Spiritual 18 Buddhist 12 Folk religion 6 Hindu 5 Other 4 Sikh 3	
<i>Notes: Two independent annotators manually verified all automated classifications (gpt-4-turbo) of the self-describe string. See paper for details.</i>			
24	religion_simplified	Simplified categories of participant religion	constructed categorical
		N Missing: 0 N Unique: 6 No Affiliation 851 Christian 487 Prefer not to say 59 Jewish 42 Muslim 31 Other 30	
<i>Notes: Simplified version of religion_categorised for more aggregate analysis.</i>			
25	ethnicity	Dictionary of ethnicity information.	NA dict
<i>Notes: Keys explained below.</i>			
26	ethnicity_self_described	Participant (c) self-description	direct string
	<i>Question text: What is your ethnicity?</i>	N Missing: 0 N Unique: 264 mean chars 9.2 std chars 6.2 min chars 3.0 max chars 99.0	
<i>Notes: Participant had option to type and Self Describe or select Prefer not to say.</i>			
27	ethnicity_categorised	Granular categories of participant ethnicity	constructed categorical
		N Missing: 0 N Unique: 9 White 969 Black / African 122 Hispanic / Latino 121 Asian 95 Prefer not to say 86 Mixed 68 Other 17 Middle Eastern / Arab 14 Indigenous / First Peoples 8	
<i>Notes: Two independent annotators manually verified all automated classifications (gpt-4-turbo) of the self-describe string. See paper for details.</i>			
28	ethnicity_simplified	Simplified categories of participant ethnicity	constructed categorical
		N Missing: 0 N Unique: 7 White 969 Black 122 Hispanic 121 Asian 95 Prefer not to say 86 Mixed 68 Other 39	
<i>Notes: Simplified version of ethnicity_categorised for more aggregate analysis.</i>			
29	location	Dictionary of location information.	NA dict
<i>Notes: Keys explained below.</i>			
30	location_birth_country	Participant country of birth	direct categorical
	<i>Question text: In which country were you born?</i>	N Missing: 0 N Unique: 75 Too many values to show -	
<i>Notes: Selected from standardised dropdown country list.</i>			
31	location_birth_countryISO	ISO 3166-1 alpha-3 code for the country of birth	constructed categorical
		N Missing: 0 N Unique: 75 Too many values to show -	
32	location_birth_subregion	Participant sub-region of birth	constructed categorical
		N Missing: 0 N Unique: 16 Too many values to show -	
<i>Notes: Mapped from country of birth, based on United Nations defined subregions.</i>			
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
33	location_reside_country Participant country of residence <i>Question text: In which country do you currently reside?</i>	direct	categorical
		N Missing: 0 N Unique: 38 Too many values to show -	
	<i>Notes: Selected from standardised dropdown country list.</i>		
34	location_reside_countryISO ISO 3166-1 alpha-3 code for the country of residence	constructed	categorical
		N Missing: 0 N Unique: 38 Too many values to show -	
35	location_reside_subregion Participant sub-region of residence	constructed	categorical
		N Missing: 0 N Unique: 11 Too many values to show -	
	<i>Notes: Mapped from country of residence, based on United Nations defined subregions.</i>		
36	location_same_birth_reside_countryWhether the participant was born and resides in the same country	constructed	binary
		N Missing: 0 N Unique: 3 Yes 1320 No 177 Prefer not to say 3	
37	location_special_region Adjusted regional categories for unique sample properties	constructed	categorical
		N Missing: 0 N Unique: 11 US 338 Europe 313 UK 292 Latin America and the Caribbean 146 Australia and New Zealand 129 Africa 118 Asia 60 Northern America 50 Middle East 50 Prefer not to say 3 Oceania 1	
	<i>Notes: Within regions and sub-regions, some countries are split out to better represent sample density (e.g., treating UK and US samples seperately from Europe and North America).</i>		
38	study_id Unique study idenfiter on Prolific	meta	string id
		N Missing: 0 N Unique: 51	
39	study_locale Recruitment country of Prolific study	meta	categorical
		N Missing: 0 N Unique: 33 Too many values to show -	
40	generated_datetime Recorded date of the survey completion	meta	datetime
		N Missing: 0 N Unique: 1492 earliest_date 2023-11-22 15:48:46 latest_date 2023-12-22 06:56:27	
	<i>Notes: End time, not start time</i>		
41	timing_duration_s Duration of the survey session (in seconds)	meta	float
		N Missing: 0 N Unique: 977 mean 2154.2 std 20557.1 min 160.0 max 529927.0	
	<i>Notes: Extreme values are caused by participants completing task in multiple sessions.</i>		
42	timing_duration_mins Duration of the survey session (in minutes)	constructed	float
		N Missing: 0 N Unique: 977 mean 35.9 std 342.6 min 2.7 max 8832.1	
	<i>Notes: timing_duration_s / 60. Extreme values are caused by participants completing task in multiple sessions.</i>		
43	included_in_UK_REP Indicator if participant was included in the rebalanced UK representative sample	constructed	binary
		N Missing: 0 N Unique: 2 False 1257 True 243	
	<i>Notes: Census-representative samples were rebalanced to mitigate sampling issues. See paper for details.</i>		
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
44	included_in_US_REP	Indicator if participant was included in the rebalanced US representative sample	constructed binary
		N Missing:	0
		N Unique:	2
		False	1270
		True	230
<i>Notes: Census-representative samples were rebalanced to mitigate sampling issues. See paper for details.</i>			
45	included_in_balanced_subset	Indicator if participant's conversations are included in the balanced subset	constructed binary
		N Missing:	0
		N Unique:	2
		True	1246
		False	254
<i>Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.</i>			

4.2 Conversations Codebook

VARIABLE	LABEL	CATEGORY	TYPE
0 user_id	Unique participant identifier	meta	string id
Notes: Pseudonymized from Prolific worker ID. Used to link conversation data to survey data.		N Missing: N Unique:	0 1396
1 conversation_id	Unique conversation identifier	meta	string id
		N Missing: N Unique:	0 8011
2 opening_prompt	Opening human-written prompt of the conversation	direct	string
Question text: Now start the conversation with your question, request or statement.		N Missing: N Unique:	0 7811
		mean chars	65.7
		std chars	59.2
		min chars	2.0
		max chars	1195.0
Notes: We provide the following soft guidance: Need some inspiration? You can request help with a task (like writing a recipe, organising an activity or event, completing an assignment)... You can chitchat, have casual conversation or seek personal advice. You can ask questions about the world, current events or your viewpoints.			
3 open_feedback	Participant written feedback on the conversation as a whole.	direct	string
Question text: Give the model some feedback on the conversation as whole. Hypothetically, what would an ideal interaction for you look like here? What was good and what was bad? What (if anything) was missing? What would you change to make the conversation better? Please write 2-5 sentences in your own words.		N Missing: N Unique:	0 7953
		mean chars	160.1
		std chars	106.4
		min chars	2.0
		max chars	1581.0
Notes: Entry box reads: Enter text here. Do not copy and paste.			
4 conversation_type	Type of conversation (from pre-defined categories)	direct	categorical
Question text: Choose what type of conversation you want to have.		N Missing: N Unique:	0 3
		unguided	3113
		values guided	2460
		controversy guided	2438
Notes: Participants pick from the following radio buttons: Unguided. Ask, request or talk to the model about anything. It is up to you! Values guided. Ask, request or talk to the model about something important to you or that represents your values. This could be related to work, religion, family and relationship, politics or culture. Controversy guided. Ask, request or talk to the model about something controversial or where people would disagree in your community, culture or country. We also provide the additional instruction: Remember if you are here as a paid study participant, you need to do two of each type. If you are here as a volunteer, then take your pick!			
5 conversation_turns	Number of human-model turns (back-and-forths) in the conversation.	meta	int
		N Missing: N Unique:	0 13
		mean	3.4
		std	1.6
		min	2.0
		max	22.0
Notes: We force 2 turns as the minimum. After the opening turn, we give the instruction: Now continue the conversation. Conversations can be between 2 and 10 turns. Try to vary the length. When you're done, click Finish.			
6 conversation_history	Full conversation history (human and model messages, with scores and model metadata)	direct	dict
Notes: We provide an example of what this nested conversation history looks like below.		Too many values to show	-
7 performance_attributes	How well the top-rated model response performed across different attributes	nested	dict
Question text: Tell us how the model performed. Consider your first message and the top-rated response. Rate the following statements about the performance across different attributes. This response...		N Missing: N Unique:	1824 7532
values	...reflected my values or cultural perspective	mean	74.1
		std	22.2
		min	1.0
		max	100.0
fluency	...was well-written and coherent	mean	84.3
		std	18.3
		min	1.0
		max	100.0
factuality	...was factual and informative	mean	79.2
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
		std min max	21.5 1.0 100.0
safety	...was safe and doesn't risk harm to myself and others	mean std min max	85.1 19.3 1.0 100.0
diversity	...summarised multiple viewpoints or different worldviews	mean std min max	68.7 25.3 1.0 100.0
creativity	...was creative and inspiring	mean std min max	63.7 26.1 1.0 100.0
helpfulness	...was helpful and relevant to my request	mean std min max	81.5 21.9 1.0 100.0
<i>Notes: Sliders from [Performed very poorly] to [Performed very well] are recorded on a 1-100 scale. Participant does not see numeric value. Note that the attributes align choice_attributes, as well as with the stated preference ratings from The Survey. Participants had option to select N/A, which is recorded as Null. N Missing indicates the number of participants who have at least one missing value in the nested columns. N Unique indicates the unique combinations of use cases selected by participants. There was no option for 'other'. Note, these sliders run from 1-100 (on Dynabench). The sliders for stated_prefs (in Survey on Qualtrics) run 0-100.</i>			
8	choice_attributes	How different attributes influenced the participant's choice of the top-rated model response	direct dict
<i>Question text:</i> Tell us why you chose this response over others. Consider your first message and top-rated response compared to other responses. Rate the following statements about the importance of different attributes in your decision. I chose this response...			
		N Missing: N Unique:	1740 7526
values	...reflected my values or cultural perspective	mean std min max	66.9 27.2 1.0 100.0
fluency	...was well-written and coherent	mean std min max	82.5 18.5 1.0 100.0
factuality	...was factual and informative	mean std min max	79.3 21.0 1.0 100.0
safety	...was safe and doesn't risk harm to myself and others	mean std min max	72.1 27.8 1.0 100.0
diversity	...summarised multiple viewpoints or different worldviews	mean std min max	66.0 26.5 1.0 100.0
creativity	...was creative and inspiring	mean std min max	62.1 27.1 1.0 100.0
helpfulness	...was helpful and relevant to my request	mean std min max	82.5 20.0 1.0 100.0
<i>Notes: Sliders from [Very unimportant] to [Very important] are recorded on a 1-100 scale. Participant does not see numeric value. Note that the attributes align with performance_attributes, as well as the stated preference ratings from The Survey. Participants had option to select N/A, which is recorded as Null. num_missing indicates the number of participants who have at least one missing value in the nested columns. num_unique indicates the unique combinations of use cases selected by participants. There was no option for 'other'. Note, these sliders run from 1-100 (on Dynabench). The sliders for stated_prefs (in Survey on Qualtrics) run 0-100.</i>			
9	generated_datetime	Recorded date of the conversation completion	meta datetime
		N Missing: N Unique: earliest_date latest_date	0 7820 2023-11-22 15:55:46 2023-12-22 08:04:46
<i>Notes: Recorded at end of conversation, before fine-grained feedback page shown.</i>			
10	timing_duration_s	Duration of the conversation (in seconds)	meta float
		N Missing:	0
Continued on next page			

VARIABLE	LABEL	CATEGORY	TYPE
		N Unique:	7656
		mean	555.9
		std	422.1
		min	73.5
		max	17145.8
<i>Notes: Extreme values are caused by participants completing task in multiple sessions.</i>			
11	timing_duration_mins	Duration of the conversation (in minutes)	
		constructed	float
		N Missing:	0
		N Unique:	1948
		mean	9.3
		std	7.0
		min	1.2
		max	285.8
<i>Notes: timing_duration_s / 60. Extreme values are caused by participants completing task in multiple sessions.</i>			
12	included_in_balanced_subset	Indicator if participant's conversations are included in the balanced subset	
		constructed	binary
		N Missing:	0
		N Unique:	2
		True	6696
		False	1315
<i>Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.</i>			

4.3 Utterances Codebook

VARIABLE	LABEL	CATEGORY	TYPE
0 user_id	Unique participant identifier	meta	string id
Notes: Pseudonymized from Prolific worker ID. Used to link utterance data to survey data.		N Missing: N Unique:	0 1396
1 conversation_id	Unique conversation identifier	meta	string id
Notes: Used to link utterance data to conversation data.		N Missing: N Unique:	0 8011
2 interaction_id	Unique interaction identifier, where an interaction is a turn within a conversation (single human message with multiple model responses)	meta	string id
		N Missing: N Unique:	0 27172
3 utterance_id	Unique utterance identifier, where an utterance is a single human message - single model response pair	meta	string id
		N Missing: N Unique:	0 68371
4 within_turn_id	Within turn identifier of up to four model responses to a single human message	meta	string id
Notes: Order is random, not based on score or presentation in interface		N Missing: N Unique:	0 4
5 conversation_type	Type of conversation (from pre-defined categories)	direct	categorical
Question text: Choose what type of conversation you want to have.		N Missing: N Unique:	0 3
		unguided	3113
		values guided	2460
		controversy guided	2438
Notes: Participants pick from the following radio buttons: Unguided. Ask, request or talk to the model about anything . It is up to you! Values guided. Ask, request or talk to the model about something important to you or that represents your values . This could be related to work, religion, family and relationship, politics or culture. Controversy guided. Ask, request or talk to the model about something controversial or where people would disagree in your community, culture or country. We also provide the additional instruction: Remember if you are here as a paid study participant, you need to do two of each type. If you are here as a volunteer, then take your pick!			
6 turn	Turn of conversation when prompt was entered	meta	int
		N Missing: N Unique:	0 22
		mean	1.2
		std	1.6
		min	0.0
		max	21.0
Notes: In the paper, we refer to the first turn as T=1. Here, we index the first turn as 0.			
7 model_name	Name of LLM	meta	categorical
		N Missing: N Unique:	0 21
		command	4812
		claude-instant-1	4292
		models/chat-bison-001	4168
		HuggingFaceH4/zephyr-7b-beta	4133
		meta-llama/Llama-2-7b-chat-hf	3995
		command-light	3929
		command-nightly	3816
		gpt-4-1106-preview	3735
		gpt-4	3515
		meta-llama/Llama-2-70b-chat-hf	3493
		gpt-3.5-turbo	3471
		timdettmers/guanaco-33b-merged	3468
		claude-2.1	3338
		mistralai/Mistral-7B-Instruct-v0.1	3261
		claude-2	3209
		tiuae/falcon-7b-instruct	2608
		OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5	2314
		meta-llama/Llama-2-13b-chat-hf	1744
		luminous-supreme-control	1722
		google/flan-t5-xxl	1715
		luminous-extended-control	1633
Notes: We provide the long name as it appeared on our backend. We provide a mapping of long names to shorter more familiar names on our Github or in the paper.			
8 model_provider	Provider of the LLM	meta	categorical
		N Missing: N Unique:	0 6
		huggingface_api	26731
		cohere	12557
		anthropic	10839
		openai	10721
		google	4168
		aleph	3355

Continued on next page

VARIABLE	LABEL	CATEGORY	TYPE
Notes: Note for open-access LLMs, HuggingFace API is always listed as the source and does not imply they built the model.			
9	user_prompt	Human-written message.	directstring
		N Missing: N Unique: mean chars std chars min chars max chars	0 26673 69.9 62.0 1.0 1311.0
10	model_response	Model-generated response	directstring
		N Missing: N Unique: mean chars std chars min chars max chars	0 66614 565.3 387.9 1.0 4630.0
Notes: An empty string is stored as 'EMPTY STRING'.			
11	score	Score of the model response	directint
Question text: Rate the model responses. There are no right or wrong answers. Use your subjective judgement.			
		N Missing: N Unique: mean std min max	0 100 65.1 29.3 1.0 100.0
Notes: Sliders from [Terrible] to [Perfect] are recorded on a 1-100 scale. Participant does not see numeric value.			
12	if_chosen	Whether model response was highest-rated by participant	constructedbinary
		N Missing: N Unique: False True	0 2 40934 27437
Notes: In case of a tie, a random response is chosen.			
13	included_in_balanced_subset	Indicator if participant's conversations are included in the balanced subset	constructedbinary
		N Missing: N Unique: True False	0 2 57401 10970
Notes: Balanced subset was created to equally sample conversations of three types (unguided, values, controversy). We only include participants who have at least one of each conversation type, and then ensure equal numbers of each type are retained. See paper for details.			

4.4 Metadata Codebook

VARIABLE	LABEL	CATEGORY	TYPE
0	column_id	Source of text utterance	meta categorical
		N Missing: 0 N Unique: 5 model_response 68371 user_prompt 27172 open_feedback 8011 self_description 1500 system_string 1500	
1	user_id	Unique participant identifier	meta string id
		N Missing: 0 N Unique: 1500	
	Notes: Pseudonymized from Prolific worker ID. Used to link metadata to main data.		
2	conversation_id	Unique conversation identifier	meta string id
		N Missing: 3000 N Unique: 8011	
	Notes: Used to link metadata to main data.		
3	interaction_id	Unique interaction identifier, where an interaction is a turn within a conversation (single human message with multiple model responses)	meta string id
		N Missing: 11011 N Unique: 27172	
	Notes: Used to link metadata to main data.		
4	utterance_id	Unique utterance identifier, where an utterance is a single human message - single model response pair	meta string id
		N Missing: 38183 N Unique: 68371	
	Notes: Used to link metadata to main data.		
5	pii_flag	Automated flag for personally identifiable information	meta binary
		N Missing: 0 N Unique: 2 False 105443 True 1111	
	Notes: Uses scrubadub https://scrubadub.readthedocs.io/en/stable/ to find PII. There may be some misclassifications. Many of the inspected positives were false positives. All positive human-written texts checked. See pii_manual_flag.		
6	pii_manual_flag	Manual verification of personally identifiable information in human-written texts	meta binary
		N Missing: 106387 N Unique: 1 nan 106387 0.0 167	
	Notes: For any automated PII flags, we manually checked the human-written text for PII. All were false positives so this flag overrules the automated flag. We did not check model-generated text for PII. NaN indicates entry was not manually checked.		
7	language_flag	Automated language detection	meta categorical
		N Missing: 0 N Unique: 59 Too many values to show -	
	Notes: Uses langid. There may be some misclassifications.		
8	en_flag	Whether detected language is English	meta binary
		N Missing: 0 N Unique: 2 Too many values to show -	
	Notes: Constructed based on automated language detection.		
9	moderation_flag	Automated flag for moderation	meta nested dict
	Notes: Uses OpenAI moderation API. There may be some misclassifications. Nested dictionary with binary flags and probabilities for sub-categories of harm.		