S²BENCH: A SURVEY-GROUNDED BENCHMARK FOR LLM-BASED SOCIAL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-driven social simulations enable controlled and repeatable experiments that would be costly, slow, or unethical with real people. Evaluating whether these simulations faithfully reflect human behavior remains challenging; existing work largely relies on real-world survey responses as quantifiable records of human choices. Yet current survey-based benchmarks are limited in three ways. (1) They evaluate only static settings with fixed respondents and fixed questions, whereas real survey environments keep introducing both. (2) They release only a processed dataset without a unified evaluation protocol, so every comparison must be reimplemented. (3) Their coverage and scale remain narrow. To address these limitations, we introduce ***S²Bench***, a large-scale benchmark for survey response prediction. It defines four settings by crossing seen and unseen respondents with seen and unseen questions. We also provide a public evaluation platform that scores submitted methods under identical conditions and ranks them on a live leaderboard, and release the benchmark construction pipeline. Applied to 15 real-world survey datasets across 234 waves, the pipeline yields nearly 0.4 billion respondent–question–answer triples. We evaluate prompting- and training-based methods across 13 LLMs and find that even the strongest models reach only 40–55% accuracy. Although scaling and reasoning modes improve performance, the gains diminish quickly, suggesting that the ability to simulate a person has improved far more slowly than general abilities such as knowledge and reasoning. Fine-tuning transfers well to new respondents, but much less to new questions.

1 INTRODUCTION

A core goal of social science is to understand and address problems concerning people and society, yet studying such problems often requires interventions on real people, whose consequences are experienced in the real world and cannot always be reversed after the fact. Social simulation instead experiments on computational models (Lazer et al., 2009; Squazzoni et al., 2020), yet modeling society remains difficult because human behavior is shaped by beliefs, values, and interpretations rather than fixed laws. Large language models (LLMs) (Achiam et al., 2023; Wei et al., 2022; Bai et al., 2023; Kosinski, 2023) offer a way to narrow this gap, and arguably reshape what social simulation can be. Trained on vast amounts of human-produced text, they have absorbed how people speak, reason, and justify their choices, and thus provide a strong human-like foundation on which simulated individuals can be built (Argyle et al., 2023; Park et al., 2023; Aher et al., 2023).

A central problem in LLM-based social simulation is evaluation: do simulated agents behave like the people they represent? Answering this requires records of what real people said and did, for which social surveys are a convenient source. Survey research administers standardized questionnaires to large, representative samples, producing respondent–question matrices of discrete answers together with demographic and attitudinal backgrounds (Dillman et al., 2014; Tourangeau et al., 2013; Williams and Brick, 2018). Each matrix cell is a real choice from a fixed option set, so a model prediction can be scored directly as correct or incorrect. The corresponding LLM task renders a respondent's background as a persona, presents a target question with its options, and predicts that person's answer. Survey response prediction is therefore widely used to evaluate social simulation (Santurkar et al., 2023; Moon et al., 2024; Suh et al., 2025; Lin et al., 2025; Wang et al., 2025; Hu et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with existing benchmarks. **SETTINGS**: number of evaluation settings (1 means static evaluation only). **PIPELINE**: source-agnostic automated construction pipeline. **LEADERBOARD**: public submission interface with an automatically updated leaderboard. **WAVES**: number of survey waves integrated. **SCALE**: number of respondent–question–answer triples.

BENCHMARK	SETTINGS	PIPELINE	LEADERBOARD	WAVES	SCALE
OpinionQA (Santurkar et al., 2023)	1	✗	✗	1	80K
SubPOP (Suh et al., 2025)	1	✗	✗	2	73K
AlignSurvey (Lin et al., 2025)	1	✗	✗	4	600K
SocioBench (Wang et al., 2025)	1	✗	✗	10	277K
SimBench (Hu et al., 2026)	1	✗	✗	20	10M
Ours (S^2Bench)	**4**	✓	✓	**234**	**0.4 billion**

Existing survey-based benchmarks share three limitations. *(1) Static evaluation settings.* They assume fixed respondents and questions and evaluate only held-out cells of an observed matrix. Real surveys continually introduce new respondents, as a Big Five personality inventory is administered to successive cohorts, and new questions, as the same residents are repeatedly surveyed on newly emerging community-governance issues. Without separating these cases, a benchmark cannot distinguish transferable representations of people and questions from interpolation within one matrix, and may overstate deployment performance. *(2) No shared interface for comparison.* Releasing processed data without a unified evaluation protocol forces researchers to reimplement data loading, prompts, and metrics, making comparisons expensive and fragile. An open interface is needed to score submitted methods under identical conditions on a public leaderboard. *(3) Limited scale.* Existing benchmarks contain only a few survey waves and at most a few hundred thousand response triples, insufficient for cross-country, cross-domain, longitudinal, or large-scale training analyses.

We introduce ***S^2Bench*** to address these limitations. *First*, a two-axis framework crosses seen and unseen respondents with seen and unseen questions, yielding four settings: missing value imputation, dynamic respondent generalization, dynamic question generalization, and dynamic respondent–question generalization. It separates the two kinds of novelty and makes static evaluation one case rather than the whole task. *Second*, a public platform runs submitted methods with the same data, prompts, and metrics and reports results on a live leaderboard by dataset and setting, turning comparison into submission rather than reimplementation. We also release a source-agnostic construction pipeline covering codebook parsing, question and option normalization, persona construction, and split generation, enabling others to reproduce and extend the data. *Third*, applying the pipeline to 15 public survey datasets across 234 waves yields nearly 0.4 billion valid respondent–question–answer triples, about forty times the largest previously released benchmark. The data span political attitudes, social values, religion, family, economics, inequality, institutional trust, environmental concern, and subjective well-being. Table 1 compares S^2Bench with recent LLM-based social survey benchmarks.

Across prompting and training-based methods on 13 LLMs, four findings stand out. First, even the strongest models reach only 40–55% accuracy. Second, scale helps with rapidly diminishing returns, while reasoning adds only a few points at substantial computational cost. Third, person-simulation ability improves much more slowly than general knowledge and reasoning, possibly because faithful simulation must reflect diverse values and bounded human cognition rather than only agreeable or optimal choices. Fourth, fine-tuning transfers better to new respondents than to new questions, plausibly because human thinking generalizes across people whereas questions are more diverse.

Contributions. We define four evaluation settings by crossing seen and unseen respondents with seen and unseen questions, together with baseline methods, metrics, and protocols. We provide a public evaluation platform and release a source-agnostic construction pipeline for reproducible and extensible comparison. Applying the pipeline to 15 survey datasets across 234 waves yields the largest survey response prediction benchmark to date, with nearly 0.4 billion respondent–question–answer triples, on which we evaluate 13 mainstream models.

2 FORMULATION OF SURVEY-BASED SOCIAL SIMULATION

We evaluate a social simulator by asking it to reproduce the answers that real people gave to a real questionnaire. This section fixes the notation used throughout the paper. Calligraphic capitals denote sets, lowercase italics their elements, and a hat a model prediction rather than an observation.

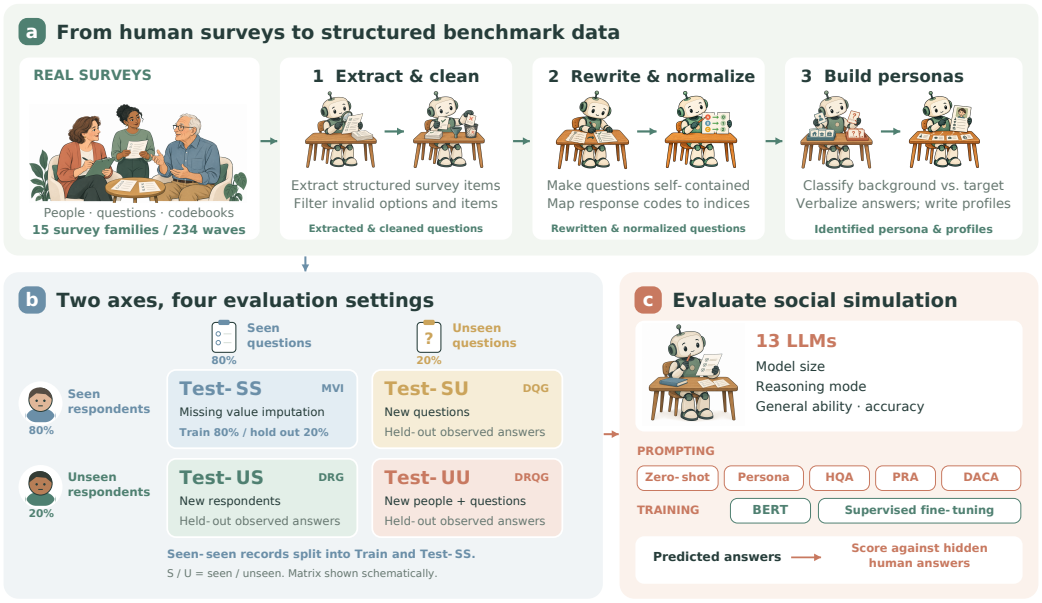

Figure 1: Overview of S^{2}Bench. (a) Survey questions and respondent attributes are processed into normalized questions and textual personas. (b) A two-axis respondent–question split framework. (c) Models are evaluated across prompting and training methods and key capability dimensions.

Respondents, questions, answers, and personas. A survey administers a questionnaire to a sample of people, giving a respondent set $\mathcal{R} = \{r_1, \ldots, r_N\}$ and a target-question set $\mathcal{Q} = \{q_1, \ldots, q_M\}$. Each question $q \in \mathcal{Q}$ carries a respondent-facing text t_q and an *ordered* option list $\mathcal{A}_q = (a_{q,0}, \ldots, a_{q,m_q-1})$ of length m_q, where $a_{q,k}$ is the text of the k-th option; the order matters, because many survey scales are ordinal. Answering q means selecting one index from the *answer space* $\mathcal{Y}_q = \{0, 1, \ldots, m_q - 1\}$, and we write $y_{r,q} \in \mathcal{Y}_q$ for the answer that respondent r actually gave to q. Surveys are rarely complete, so we let $\mathcal{O} \subseteq \mathcal{R} \times \mathcal{Q}$ collect the pairs whose answer was observed and take the survey data to be the set of triples $\mathcal{D} = \{(r, q, y_{r,q}) : (r, q) \in \mathcal{O}\}$, equivalently a partially observed *answer matrix* $\mathbf{Y}$ of size $N \times M$ whose entry $\mathbf{Y}_{ij}$ equals y_{r_i,q_j} when $(r_i, q_j) \in \mathcal{O}$ and is left blank otherwise. Respondents and questions, the two axes of this matrix, are the two axes along which S^{2}Bench evaluates generalization. Part of every questionnaire records who the respondent is rather than what they think. We write $\mathcal{Q}_{\text{per}}$ for these background questions, kept separate from the target questions $\mathcal{Q}$, and verbalize each respondent's background answers into a natural-language *persona profile* p_r. The profile is what a simulator is told about r; it never contains an answer to a target question.

Context. Beyond the persona, a simulator may be shown other cells of the answer matrix. For a target pair (r, q) there are two natural sources: the same respondent's answers to other questions, $\mathcal{C}^{\text{his}}_{r,q} = \{(q', y_{r,q'}) \in \mathcal{D} : q' \in \mathcal{Q},\ q' \neq q\}$, and other respondents' answers to the same question, $\mathcal{C}^{\text{pop}}_{r,q} = \{(r', y_{r',q}) \in \mathcal{D} : r' \in \mathcal{R},\ r' \neq r\}$. A prompting method is a choice of context $c_{r,q}$ from these sources: $c_{r,q} = \emptyset$, $c_{r,q} = \mathcal{C}^{\text{his}}_{r,q}$, $c_{r,q} = \mathcal{C}^{\text{pop}}_{r,q}$, or $c_{r,q} = \mathcal{C}^{\text{his}}_{r,q} \cup \mathcal{C}^{\text{pop}}_{r,q}$. Section 4 evaluates exactly these four choices.

The simulation task. A *simulator* f_θ is a model with parameters θ that reads a persona, a context, and a question, and returns one option index:

$$\hat{y}_{r,q} = f_\theta\big(p_r,\ c_{r,q},\ t_q,\ \mathcal{A}_q\big) \ \in\ \mathcal{Y}_q. \tag{1}$$

Given a test set $\mathcal{T} \subseteq \mathcal{D}$, we score the simulator by how often it recovers the recorded answer,

$$\mathrm{Acc}(f_\theta; \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{(r,q,y_{r,q}) \in \mathcal{T}} \mathbb{1}\big[\hat{y}_{r,q} = y_{r,q}\big]. \tag{2}$$

The evaluation settings of Section 3.1 differ only in which test set $\mathcal{T}$ is used: whether the respondent r, the question q, both, or neither were available while fitting θ. (Appendix B.1)

3 S^{2}BENCH: A COMPREHENSIVE BENCHMARK FOR SOCIAL SIMULATION

This section presents S^{2}Bench in three parts. Section 3.1 describes the construction pipeline, which turns raw survey materials into the prediction task of Section 2 and then partitions the answer matrix into the four evaluation settings. Section 3.2 describes the public platform, which lets users compare existing methods, build new benchmarks from new surveys, and submit new methods for evaluation. Section 3.3 reports the statistics of the resulting benchmark: its scale and topical coverage across 15 survey families and 234 waves, the variation in respondents, questions, and answer-matrix density across waves, and the composition of the training set and the four test sets.

3.1 CONSTRUCTION PIPELINE

As shown in Figure 1, the pipeline has three stages. *Question processing* extracts survey items from raw materials, removes invalid options and unusable items, rewrites questions into self-contained forms, and normalizes recorded responses into option indices. *Persona processing* separates background questions from target questions and converts each respondent's background answers into a textual profile. *Benchmark processing* partitions the respondent–question answer matrix along both axes to define the four evaluation settings. Every stage is agnostic to the survey source, so a new survey can be added without new source-specific code. The complete prompts used at each LLM-based step are given in Appendix C.

Question processing. This stage turns the files released with a survey into normalized questions and answers, in five steps that each consume the output of the previous one (Appendix B.2).

(i) Parsing. A survey is distributed as a set of *raw sources*, the files released with it: questionnaire documents that show the questions as respondents saw them, codebooks that list every variable with its label and answer codes, statistical metadata, and tabular data files that store one row of coded answers per respondent. The *format* f of a source is its file type (e.g., PDF, DOCX, or CSV). Since formats differ in structure, a format-specific parser ϕ_f converts each source x into plain text $z = \phi_f(x)$ that keeps the survey content, namely question wording, variable labels (e.g., `TRUSTGOV`), and value labels (e.g., `1 = Just about always`), and discards file structure. A codebook can run to hundreds of pages, more than an LLM reads reliably in one pass, so we split z into consecutive chunks $z_1, \ldots, z_n$.

(ii) Extraction. An LLM extractor Ext_θ reads each chunk and returns the survey items it contains as *question objects*, i.e., structured records with a question identifier (e.g., `V201`), the respondent-facing text t_q, the question type (e.g., single-choice, multiple-response, open-ended, or numeric), the raw option list $\mathcal{A}_q^{\mathrm{raw}}$ with the response code of each option, and source metadata recording where the item was found. A chunk without a usable item returns the empty set, and the union over chunks gives the raw question set $\mathcal{Q}^{\mathrm{raw}} = \bigcup_j \mathrm{Ext}_\theta(z_j)$.

(iii) Cleaning. Raw option lists mix substantive answers with *missing-value labels*, which record that no substantive answer was given (e.g., "Don't know" or "Refused"). An LLM classifier collects these labels across the survey into an invalid-label set $\mathcal{M}$, and we remove them from every option list, $\mathcal{A}_q = \{a \in \mathcal{A}_q^{\mathrm{raw}} : a \notin \mathcal{M}\}$, which leaves $m_q = |\mathcal{A}_q|$ options. On the cleaned questions, an LLM quality reviewer Flt_θ issues a decision $d_q = \mathrm{Flt}_\theta(q, \mathcal{A}_q) \in \{\mathrm{keep}, \mathrm{remove}\}$, removing items that are incomplete, ambiguous, or structurally defective. Because the prediction task asks for exactly one option, we further keep only single-choice questions with 2 to 12 options, excluding open-ended, numeric, unsupported, and multiple-response items (e.g., "Select all that apply"):

$$\mathcal{Q}^{\mathrm{clean}} = \{q \in \mathcal{Q}^{\mathrm{raw}} : q \text{ is single-choice},\ d_q = \mathrm{keep},\ 2 \le m_q \le 12\}. \tag{3}$$

(iv) Rewriting. Many retained items are written to be read in context and are unclear on their own. A codebook may give only a *telegraphic label* (e.g., "Trust in national govt – how often right"), and a questionnaire item may contain a *deictic reference* to its surroundings (e.g., "And what about the second one?"). Since a simulator sees each question in isolation, an LLM rewriter turns t_q into a self-contained question $\tilde{t}_q = \mathrm{Rew}_\theta(t_q)$, preserving the original language, meaning, and option order.

(v) Normalization. Data files store each answer as the response code of the chosen option, and these codes are rarely consecutive (e.g., 1–4 for valid options and 8 or 9 for missing values). For each question we define an index map $g_q : \mathcal{C}_q^{\mathrm{raw}} \to \mathcal{Y}_q$ from the codes $\mathcal{C}_q^{\mathrm{raw}}$ of the retained

options to zero-based indices in option order (e.g., $1 \mapsto 0, \ldots, 4 \mapsto 3$). Applying g_q to the data files turns every recorded answer into an index, and an answer whose code belongs to a removed option receives no index and is treated as unobserved. The stage thus outputs the processed questions $(q, \tilde{t}_q, \mathcal{A}_q, g_q)$, which form the normalized question set $\mathcal{Q}_{\text{all}}$, together with the answer triples $\mathcal{D}_{\text{all}} = \{(r, q, y_{r,q}) : r \in \mathcal{R},\ q \in \mathcal{Q}_{\text{all}}\}$ of all observed answers.

Persona processing. A questionnaire mixes two kinds of questions: *background questions* record who a respondent is (e.g., age or education), while *substantive questions* ask what the respondent thinks or does (e.g., trust in government). This stage separates the two and turns each respondent's background answers into a profile, in three steps (Appendix B.3).

(i) Classification. An LLM classifier Cls_θ assigns each $q \in \mathcal{Q}_{\text{all}}$ a persona score $s_q = \text{Cls}_\theta(q) \in [0, 1]$, which is high when q describes respondent attributes rather than substantive attitudes. With threshold $\tau_{\text{per}} = 0.8$, the persona-question set is $\mathcal{Q}_{\text{per}} = \{q \in \mathcal{Q}_{\text{all}} : s_q \geq \tau_{\text{per}}\}$ and the target-question set is $\mathcal{Q} = \mathcal{Q}_{\text{all}} \setminus \mathcal{Q}_{\text{per}}$; only target questions enter the answer matrix to be predicted.

(ii) Verbalization. For each respondent r, we replace every observed answer to a persona question by the text of the chosen option, giving the *persona facts* $\mathcal{P}_r = \{(q, a_{q,y}) : (r, q, y) \in \mathcal{D}_{\text{all}},\ q \in \mathcal{Q}_{\text{per}}\}$ (e.g., $(q^{\text{edu}}, \text{Bachelor's degree})$).

(iii) Profile writing. An LLM profile writer turns these facts into a third-person paragraph $p_r = \text{Prf}_\theta(\mathcal{P}_r)$. The writer may use only facts in $\mathcal{P}_r$ and may not infer attitudes or add unobserved traits, so p_r carries background information alone and never reveals an answer to a target question.

Benchmark processing. This stage splits the answers to target questions into one training set and four test sets, in three steps (Appendix B.4).

(i) Answer matrix. Restricting $\mathcal{D}_{\text{all}}$ to target questions gives $\mathcal{D}_{\mathcal{Q}} = \{(r, q, y_{r,q}) \in \mathcal{D}_{\text{all}} : q \in \mathcal{Q}\}$, which is the survey data $\mathcal{D}$ and the partially observed answer matrix $\mathbf{Y}$ of Section 2.

(ii) Holdout sampling. To test whether a model generalizes to people and questions it has never seen, we hold out entire rows and columns of $\mathbf{Y}$. Under a fixed seed, we sample 20% of the respondents as the unseen respondent set $\mathcal{R}_{\text{un}} \subset \mathcal{R}$ and 20% of the target questions as the unseen question set $\mathcal{Q}_{\text{un}} \subset \mathcal{Q}$, and take the complements $\mathcal{R}_{\text{se}} = \mathcal{R} \setminus \mathcal{R}_{\text{un}}$ and $\mathcal{Q}_{\text{se}} = \mathcal{Q} \setminus \mathcal{Q}_{\text{un}}$ as the seen sets.

(iii) Partition. Crossing the two axes divides the matrix into four blocks,

$$\mathcal{D}_{\sigma_r \sigma_q} = \{(r, q, y_{r,q}) \in \mathcal{D}_{\mathcal{Q}} : r \in \mathcal{R}_{\sigma_r},\ q \in \mathcal{Q}_{\sigma_q}\}, \qquad \sigma_r, \sigma_q \in \{\text{se}, \text{un}\}, \tag{4}$$

where the first subscript refers to respondents and the second to questions; for example, $\mathcal{D}_{\text{unse}}$ holds the answers of unseen respondents to seen questions. The three blocks with at least one unseen axis serve directly as test sets, Test-US $= \mathcal{D}_{\text{unse}}$, Test-SU $= \mathcal{D}_{\text{seun}}$, and Test-UU $= \mathcal{D}_{\text{unun}}$. The seen–seen block is split at the record level: 20% of its records become Test-SS, and the remainder becomes the training set $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{sese}} \setminus$ Test-SS. This yields one training set and four test sets, each corresponding to a distinct kind of generalization. Test-SS is *missing value imputation* (MVI), where both the respondent and the question have been seen. Test-US is *dynamic respondent generalization* (DRG), where the question is familiar but the respondent is new. Test-SU is *dynamic question generalization* (DQG), where the respondent is familiar but the question is new. Test-UU is *dynamic respondent–question generalization* (DRQG), where neither has been seen. The static evaluation of prior benchmarks corresponds to Test-SS alone.

3.2 Evaluation Platform

A static-file benchmark fixes its surveys at release and provides no shared mechanism to add surveys or evaluate new methods under one protocol. Our public platform provides three functions that let both data and method comparisons grow after release.

Function 1: Comparing existing methods (Figure 7). Each *run* evaluates one method on one survey wave under one setting. The live leaderboard provides a *model view* for foundation models under fixed prompting and an *algorithm view* for complete prediction systems, matching the two tracks of Function 3. Users can filter and search runs and sort multiple metrics. Only owner-published runs appear, and results remain separated by dataset and setting for controlled comparison.

Function 2: Building new benchmarks (Figure 8). Users upload a ZIP of up to 500 MB containing the survey's raw sources, such as its codebook and data file, and select the processing LLM and

respondent sample size. The service executes eleven steps across the question, persona, and benchmark processing stages of Section 3.1, reports status and logs, maintains a task history, and returns a downloadable benchmark in the standardized S^{2}Bench format. Thus, new surveys can be processed without installing the pipeline locally.

Function 3: Evaluating new methods (Figures 9 and 10). The platform scores submissions centrally while keeping test answers hidden. The *model API track* evaluates an OpenAI-compatible model endpoint using platform-defined prompts, parsing, and metrics, enabling new foundation models to be compared under a fixed procedure. The *open system track* evaluates any prediction service as a black box and optionally provides labeled training data, personas, normalized questions, and protocol documentation, allowing fine-tuned models, retrieval systems, or other end-to-end methods. Both tracks report live progress, and owners choose whether to publish completed results. Details appear in Appendix D; protocol pseudocode and interface figures are in Appendix D.

3.3 Benchmark Statistics

We summarize S^{2}Bench at the family, wave, and split levels; the corresponding figures and complete analysis are provided in Appendix E, and data quality is evaluated in Section 4.2.

Family level. Figure 11 summarizes the family-level coverage of S^{2}Bench. The 15 survey families contain 234 waves and 381,442,527 valid respondent–question–answer triples, built from 4,075,760 respondent entries and 30,753 question entries summed over waves. The survey families span five regions and more than four decades. ESS, WVS, and ATP contribute 20.9%, 20.9%, and 20.5% of all triples, while the other twelve families contribute 37.7%. The largest 23 waves account for 54% of all triples, indicating that the benchmark is not dominated by a small number of waves.

Wave level. Figure 12 shows that a wave has medians of 9,568 respondents, 98 target questions, and 565,684 triples, although sizes vary by more than three orders of magnitude. Answer matrices have a median density of 70.5%, and the median respondent answers 66 target questions; in 91% of waves, the average respondent answers at least 20. These distributions provide substantial variation along both generalization axes while retaining enough observed context for prediction.

Split level. Figure 13 shows that the two-axis protocol assigns 51.3% of triples to training and 12.8%, 16.0%, 15.9%, and 4.0% to Test-SS, Test-US, Test-SU, and Test-UU. Even Test-UU contains 15,123,031 triples; it has at least 10,000 triples in 70% of waves and at least 1,000 in 97%. Only four waves with three or four questions cannot form Test-SU or Test-UU. Thus, the four settings are large enough for both aggregate and per-wave comparisons.

4 Experiments

We study six research questions across the four evaluation settings. **(RQ1)** Are the construction pipeline's outputs faithful to the source surveys? **(RQ2)** How accurately do current methods predict survey responses? **(RQ3)** How does model scale affect prediction accuracy? **(RQ4)** Does reasoning mode improve prediction? **(RQ5)** Does general capability translate into simulation performance? **(RQ6)** Does fine-tuning transfer to unseen respondents and questions?

4.1 Experimental Setup

Evaluated LLMs. We evaluate a diverse set of LLMs spanning API-based models and open-weight models. The API-based models include DeepSeek-V4-Pro, DeepSeek-V4-Flash, Mimo-v2.5, Mimo-v2.5-Pro, LongCat-2.0, GPT-4o-mini, and GPT-5-mini. The open-weight models include Qwen2.5-72B-Instruct, Qwen3-32B, Qwen3-30B-A3B, Llama-3.3-70B-Instruct, Gemma-3-27B-it, and Mistral-Small-3.1-24B-Instruct. All models are evaluated with temperature set to 0.

Evaluation Methods and Metrics. We compare prompting and training-based methods. Prompting progresses from *Zero-shot* and *Persona* to HQA, PRA, and DACA, which add the target respondent's history, other respondents' answers to the target question, or both; full prompts are provided in Appendix G. Training-based methods comprise a ModernBERT-base classifier and SFT of Qwen3-8B and Qwen2.5-7B-Instruct on persona–question–answer triples, with specifications in Appendix F. Our primary metric is accuracy, the proportion of predictions that exactly match observed answers.

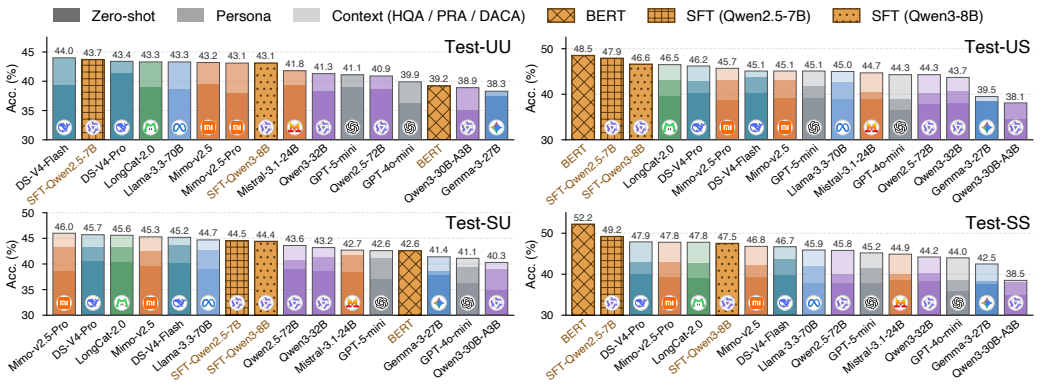

Figure 2: Overall accuracy of prompting and training-based methods across four evaluation settings.

Implementation details. All LLM-based processing steps in the construction pipeline use DeepSeek-V4-Pro. For evaluation, we use the most recent publicly available wave from each of the 15 survey datasets. To reduce computational costs, we uniformly sample 500 respondents from each wave before applying the respondent–question split. To mitigate position bias, we apply option shuffling when prompting LLMs. All random sampling operations use a fixed seed of 42 to ensure reproducibility.

4.2 Reliability of the Construction Pipeline (RQ1)

Because five construction steps rely on an LLM, we audit their faithfulness on WVS Wave 7. Three PhD candidates independently label the outputs under blinded conditions. We assess all 251 retained questions for extraction, cleaning, rewriting, and persona identification, together with the first 200 profiles in processing order. As shown in Figure 14, every output is judged appropriate by all three annotators, yielding 100% appropriateness. We attribute this result to the tightly constrained nature of the operations: extraction and cleaning copy or filter source content, rewriting must preserve meaning and option order, persona identification is a binary decision, and profile writing may use only provided facts. The prompts make these constraints explicit, leaving little room for open-ended generation. Detailed criteria and uncertainty analysis are provided in Appendix H.1.

4.3 Overall Performance (RQ2)

We evaluate all 13 LLMs with five prompting methods and three training-based methods on the most recent wave of each survey family. Context-based prompts are used only where their required information is available: HQA on Test-SU/SS, PRA on Test-US/SS, and DACA on Test-SS. Figure 2 summarizes accuracy; full results appear in Appendix I. Zero-shot accuracy is stable across the four settings (0.382–0.385 on average). Persona raises it to 0.410–0.417, while HQA reaches 0.436 on Test-SU and 0.434 on Test-SS, PRA reaches 0.441 on Test-US and Test-SS, and DACA reaches 0.452 on Test-SS. The strongest prompted models reach 0.478–0.479 with DACA. Training is strongest on seen questions: BERT obtains 0.522 on Test-SS and 0.485 on Test-US, ahead of SFT on Qwen2.5-7B (0.492 and 0.479). This advantage disappears for unseen questions: on Test-SU, the best HQA result (0.460) exceeds SFT (0.445) and BERT (0.426); on Test-UU, Persona (0.440) is comparable to SFT (0.437), while BERT falls to 0.392. The stability of Zero-shot suggests that these differences are not artifacts of test-set difficulty, while the gains from HQA, PRA, and DACA show that respondent history and population answers provide complementary information. More importantly, method rankings change with respondent and question visibility: training-based methods exploit seen questions effectively but transfer poorly to unseen ones, whereas prompted LLMs degrade much less. However, the best method correctly predicts about half of held-out responses.

4.4 Effect of Model Scale (RQ3)

To isolate scale from differences in training data and architecture, we evaluate seven Qwen2.5 models (0.5B–72B) and six Qwen3 models (0.6B–32B) with Persona, HQA, PRA, and DACA on Test-SS, micro-averaging accuracy over all 15 datasets by their numbers of response instances. Figure 3 shows that accuracy rises steeply among small models and then saturates. Under Persona, Qwen2.5 improves from 0.276 at 0.5B to 0.407 at 14B, but reaches only 0.419 at 72B; HQA follows the same pattern (0.283, 0.422, and 0.437). PRA peaks at 32B (0.457) before declining at 72B (0.447), while

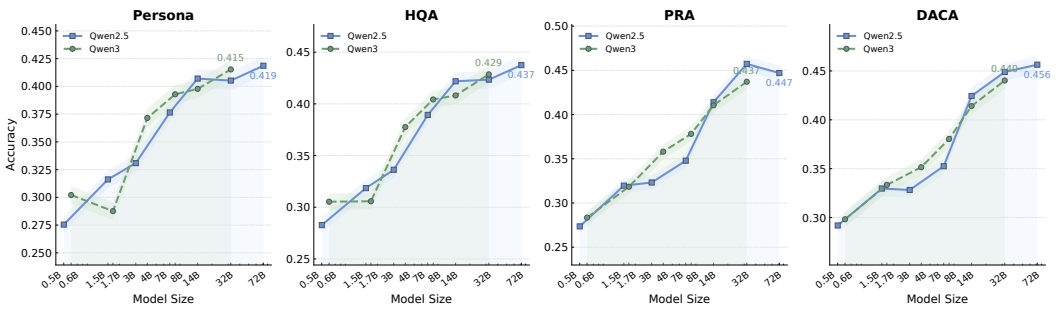

Figure 3: Accuracy scaling with model size for the Qwen2.5 (0.5B–72B) and Qwen3 (0.6B–32B) families, evaluated on four prompting settings (Persona, HQA, PRA, DACA) on the Test-SS split.

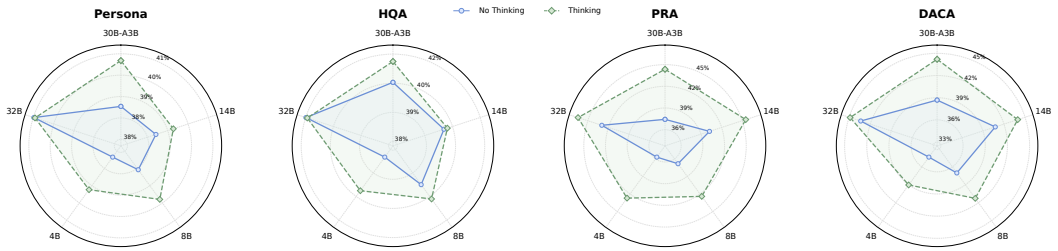

Figure 4: Accuracy with and without reasoning (thinking) mode for Qwen3 models of varying sizes, evaluated on four prompting settings (Persona, HQA, PRA, DACA) on the Test-SS split.

DACA improves only from 0.449 to 0.456. Qwen3 shows a similar, occasionally non-monotonic trajectory. Models up to 3B remain at 0.27–0.34, suggesting that a minimum capacity is required to use persona and context, but beyond 14B information matters more than parameters: moving Qwen2.5-72B from Persona to DACA adds 3.7 points, roughly three times its Persona gain from 14B to 72B. Moreover, SFT on Qwen2.5-7B reaches 0.492, outperforming every prompted model on Test-SS. Thus, a mid-sized model with rich context or task-specific training is more cost-effective than a much larger model with Persona alone.

4.5 Impact of Reasoning Mode (RQ4)

We compare five Qwen3 models (4B, 8B, 14B, 30B-A3B, and 32B) with and without thinking mode under Persona, HQA, PRA, and DACA on WVS Test-SS, using the same model weights in both modes. Figure 4 shows that thinking improves almost every model–method pair, with the largest gains for smaller or mixture-of-experts models and the context-rich PRA and DACA methods. For Qwen3-32B, however, Persona and HQA change little, while DACA improves from 43.35% to 44.86% (+1.51 points). This modest gain increases evaluation time on the same 5,135 requests from about 6 minutes to 2 hours 45 minutes, a 27× slowdown. The pattern suggests that reasoning mainly helps models aggregate information already present in rich prompts and can partly substitute for scale, rather than adding knowledge about respondents; once the model is sufficiently large, the remaining uncertainty about individual answers is difficult to resolve through more reasoning. Given its limited gain and substantial cost, thinking mode is hard to justify for large-scale survey simulation.

4.6 General Capability and Simulation Performance (RQ5)

We relate models' MMLU-Pro scores from the leaderboard[1] to Test-SS accuracy under HQA, PRA, and DACA, fitting regressions across models whose MMLU-Pro scores range from about 62 to 84. Figure 5 shows positive but moderate and statistically weak associations: Pearson's r is 0.55 for HQA ($p = 0.062$), 0.52 for PRA ($p = 0.081$), and 0.53 for DACA ($p = 0.073$), while Spearman's ρ ranges from 0.45 to 0.52 ($p = 0.081$–0.138). Across the 22-point MMLU-Pro range, fitted simulation accuracy rises by only about 4 points; under DACA, some models scoring 66–69 on MMLU-Pro reach roughly 48% accuracy, matching models above 80. General capability is therefore a weak

[1] https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

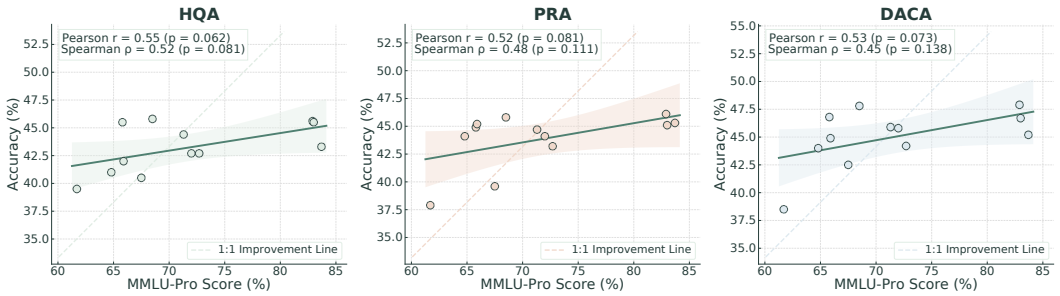

Figure 5: MMLU-Pro score versus Test-SS accuracy under HQA, PRA, and DACA. Solid lines show linear fits; dashed lines mark 1:1 improvement.

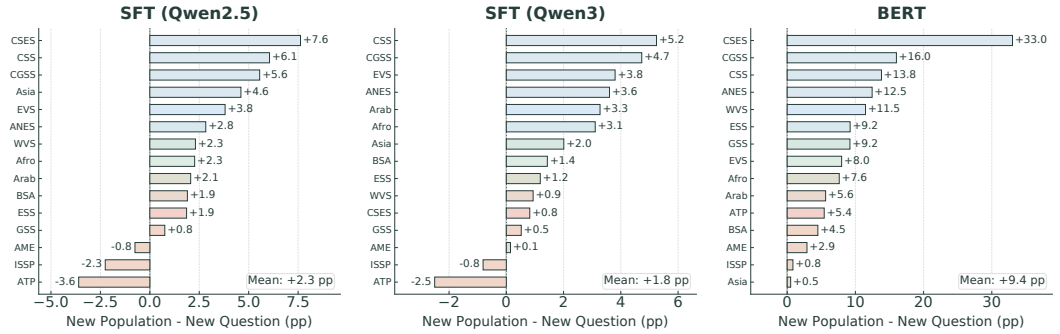

Figure 6: Difference in persona-adjusted gains between unseen respondents (Test-US) and unseen questions (Test-SU). Positive values indicate better transfer to new respondents.

proxy for simulation quality: expert knowledge and multi-step reasoning are only loosely related to calibrated knowledge of social-group opinions and its transfer to individuals, while pretraining coverage and post-training may also matter. Survey simulation must consequently be measured directly and improved through targeted training rather than expected from general progress.

4.7 Transferability of Fine-Tuning Gains (RQ6)

We compare BERT and SFT on Qwen2.5-7B and Qwen3-8B between Test-US (unseen respondents) and Test-SU (unseen questions), subtracting the Persona baseline within each dataset and setting to control for intrinsic difficulty. Figure 6 shows that fine-tuning transfers better to new respondents for most datasets and all three methods. The mean Test-US minus Test-SU gain is $+9.4$ points for BERT, $+2.3$ for SFT Qwen2.5-7B, and $+1.8$ for SFT Qwen3-8B; only a few SFT results reverse the sign, notably on ATP and ISSP. This asymmetry suggests that fine-tuning primarily learns the answer distributions of seen questions, which generalize across respondents but not to questions whose distributions were never observed. BERT depends most strongly on such question-specific learning, whereas SFT retains useful LLM priors for new questions. The smaller gaps on thematically focused surveys such as ISSP and ATP further suggest that coherence among questions aids transfer. Fine-tuning historical data is therefore well suited to extending an existing survey to new samples, while forecasting responses to genuinely new questions remains an open problem. Detailed experimental analyses for RQ1–RQ6 are provided in Appendix H.

5 Conclusion

We introduced S^2Bench, a survey response prediction benchmark spanning 15 datasets, 234 waves, nearly 0.4 billion respondent–question–answer triples, and four evaluation settings across diverse surveys and generalization settings. The strongest models reach only 40–55% accuracy; scaling shows diminishing returns, reasoning offers modest gains at high cost, and general capability correlates only weakly with simulation accuracy. Fine-tuning transfers better to unseen respondents than to unseen questions. A remaining limitation is potential training-data contamination; future work should use newly collected or access-restricted survey data for stronger evaluation (see Appendix J).

AI USE STATEMENT

Generative AI was used as part of the research methodology and for limited authoring assistance. In benchmark construction, DeepSeek-V4-Pro performed five constrained operations: question extraction, question cleaning, question rewriting, persona-question identification, and persona-profile writing. LLMs were also the models under study in the prompting and supervised fine-tuning experiments, where they generated survey-response predictions that were scored against observed human answers. Separately, generative AI tools assisted with language editing, translation, formatting, and software troubleshooting. They were not used to originate the research idea, formulate the method, design the experiments, alter the reported results, or determine the interpretation of the findings. The authors reviewed all AI-generated outputs and assisted edits and take responsibility for the final manuscript, data, results, code, and other artifacts.

ETHICS STATEMENT

This work uses publicly released social-survey datasets and publicly available language models, subject to their respective terms of use and licenses. We do not collect new data from human participants. The simulated responses are research artifacts and should not be treated as substitutes for informed responses from real individuals or as authoritative statements about any demographic group. Because language-model outputs can reproduce biases or errors present in their training data and prompts, the reported simulations should be interpreted with appropriate caution.

REPRODUCIBILITY STATEMENT

We provide the formal problem definition in Section 2, the benchmark construction pipeline in Section 3.1 with all LLM-based processing prompts in Appendix C, the evaluated models and experimental settings in Section 4.1, the evaluation prompt templates in Appendix G, and the training configurations of the training-based methods in Appendix F. All 15 source surveys are publicly available, and we release the benchmark data together with the source-agnostic construction pipeline. Random sampling uses a fixed seed of 42, all LLMs are evaluated at temperature 0 with option shuffling to mitigate position bias, and the evaluation protocol of the public platform, which centrally scores submitted model endpoints and prediction systems under identical data, prompts, and metrics, is documented in Section 3.2 and Appendix D.

REFERENCES

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR, 2023.

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

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety, 2023. URL https://arxiv.org/abs/2306.11247.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Marcel Binz and Eric Schulz. Turning large language models into cognitive models. *ArXiv*, abs/2306.03917, 2023. URL https://api.semanticscholar.org/CorpusID:259095948.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Dwaipayan Roy. Language models trained on media diets can predict public opinion. *ArXiv*, abs/2303.16779, 2023. URL https://api.semanticscholar.org/CorpusID:257805183.

Don A Dillman, Jolene D Smyth, and Leah Melani Christian. Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana*, 17:161–176, 2014.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *ArXiv*, abs/2402.01306, 2024. URL https://api.semanticscholar.org/CorpusID:267406810.

Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17001–17019, 2024.

Ori Heffetz and Daniel B Reeves. Difficulty of reaching respondents and nonresponse bias: Evidence from large government surveys. *Review of Economics and Statistics*, 101(1):176–191, 2019.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors, 2026. URL https://arxiv.org/abs/2510.17516.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, 2023.

Graham Kalton. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2):125–141, 2009.

Jaehyung Kim and Yiming Yang. Few-shot personalization of llms with mis-aligned responses. *ArXiv*, abs/2406.18678, 2024. URL https://api.semanticscholar.org/CorpusID:270764657.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4(169):2, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL https://api.semanticscholar.org/CorpusID:261697361.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*, 2024.

Chenxi Lin, Weikang Yuan, Zhuoren Jiang, Biao Huang, Ruitao Zhang, Jianan Ge, Yueqian Xu, and Jianxing Yu. Alignsurvey: A comprehensive benchmark for human preferences alignment in social surveys, 2025. URL https://arxiv.org/abs/2511.07871.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M Chan. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR, 2023.

Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *ArXiv*, abs/2209.12106, 2022. URL https://api.semanticscholar.org/CorpusID:252531526.

Flaminio Squazzoni, J Gareth Polhill, Bruce Edmonds, Petra Ahrweiler, Patrycja Antosz, Geeske Scholz, Émile Chappin, Melania Borit, Harko Verhagen, Francesca Giardini, et al. Computational models that matter during a global pandemic outbreak: A call to action. *Jasss*, 23(2), 2020.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Ren Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *ArXiv*, abs/2402.11060, 2024. URL https://api.semanticscholar.org/CorpusID:267750999.

Roger Tourangeau, Frederick George Conrad, and Mick Couper. *The science of web surveys*, volume 10. Oxford University Press New York, 2013.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models, 2025. URL https://arxiv.org/abs/2510.11131.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Douglas Williams and J Michael Brick. Trends in us face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6(2):186–211, 2018.

James D Wright, Peter V Marsden, et al. Survey research and social science: History, current practice, and future prospects. *Handbook of survey research*, pages 3–26, 2010.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *ArXiv*, abs/2310.11523, 2023. URL https://api.semanticscholar.org/CorpusID:264289064.

Zhibin Zhou, Yaoqi Li, and Junnan Yu. Exploring the application of llm-based ai in ux design: an empirical case study of chatgpt. *Human–Computer Interaction*, pages 1–33, 2024.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, pages 4653–4661, 2025.

A Related Work

↩ Back to Introduction

A.1 LLMs as Proxies for Survey Respondents

Conventional survey fieldwork is slow, expensive, and difficult to scale (Wright et al., 2010; Heffetz and Reeves, 2019; Kalton, 2009), which has motivated a line of research asking whether LLMs can act as substitutes for human respondents (Zhou et al., 2024; Zhu et al., 2025; Wang et al., 2024; Li et al., 2024; Santurkar et al., 2023; Abdurahman et al., 2024). Work in this space falls into two broad camps. One camp keeps model weights untouched and instead conditions generation at inference time, steering outputs toward a target demographic or attitudinal profile through persona-style instructions (Santurkar et al., 2023; Hwang et al., 2023; Simmons, 2022; Kim and Yang, 2024; Sun et al., 2024; Moon et al., 2024). The other camp updates model parameters directly, tuning on real survey records so that individual- or subgroup-level response patterns are absorbed into the model itself (Chu et al., 2023; He et al., 2024; Kwon et al., 2023; Zhao et al., 2023; Suh et al., 2025). Both camps, however, are typically validated on a single fixed pool of respondents and questions drawn from one survey; S^2Bench differs by explicitly separating respondents and questions seen during training from those held out, so that a method's performance can be attributed to genuine generalization rather than memorization of the exact respondents or items it was evaluated on.

A.2 Benchmarks for Survey-Style Evaluation of LLMs

A separate line of work asks how closely LLM outputs track human preferences in general (Köpf et al., 2023; Aroyo et al., 2023; Lambert et al., 2025; Ethayarajh et al., 2024), but purpose-built testbeds for social-survey response prediction are still uncommon; Table 1 situates the closest resources from both the cognitive-science and survey-methodology traditions against S^2Bench. Resources originating from cognitive science tend to probe reasoning through scenarios that are authored or simulated rather than drawn from fielded surveys (Binz and Schulz, 2023; Binz et al., 2024; Abdurahman et al., 2024; Dominguez-Olmedo et al., 2024), which suits controlled hypothesis testing but leaves the demographic breadth and multi-wave structure of real survey programs less directly addressed. Resources built directly on survey data, such as OpinionQA, Anthology, and SubPOP (Kirk et al., 2024; Santurkar et al., 2023; Moon et al., 2024; Suh et al., 2025), are closer in spirit but are each anchored to a small, fixed set of items and predominantly U.S. respondent pools. To the best of our knowledge, we did not find prior benchmarks that systematically evaluate both respondent and question generalization axes in a unified framework. S^2Bench is built to close this gap along three axes: it is assembled through a reusable extraction-and-normalization pipeline rather than a one-off script tied to a single source, so new survey programs can be folded in without redesigning the benchmark; it spans 15 datasets and 234 waves rather than one survey instrument, giving broader topical and temporal coverage than prior single-source resources; and it defines four respondent/question generalization settings rather than one static split, measuring the failure mode, generalization to unseen respondents or unseen questions, that earlier benchmarks leave unexamined.

B Worked Examples

↩ Back to Section 2

B.1 Worked Example: Survey Response Prediction

Take a respondent r whose background answers yield the persona profile

> p_r = "A woman aged 40 to 44 living in a mid-sized city. She holds a bachelor's degree, works full time in the private sector, is married with two children, and attends religious services a few times a year."

Take a target question q with t_q = "How much of the time do you think you can trust the national government to do what is right?" and the ordered options

$$\mathcal{A}_q = (\underbrace{\text{Just about always}}_{a_{q,0}}, \underbrace{\text{Most of the time}}_{a_{q,1}}, \underbrace{\text{Only some of the time}}_{a_{q,2}}, \underbrace{\text{Never}}_{a_{q,3}}),$$

so $m_q = 4$ and $\mathcal{Y}_q = \{0, 1, 2, 3\}$. Suppose the context supplies one cell from each source: this respondent answered "Somewhat dissatisfied" about the state of the economy, and another respondent with a similar background answered $a_{q,2}$ to q. The simulator receives $(p_r, c_{r,q}, t_q, \mathcal{A}_q)$ and must return a single index. It returns $\hat{y}_{r,q} = 2$, meaning "Only some of the time". The woman's recorded answer is $y_{r,q} = 2$, so this instance counts as correct in Equation (2). Had it returned $\hat{y}_{r,q} = 1$, the instance would count as wrong, even though the two options are adjacent on the scale.

B.2 Worked Example: Question Processing

We trace the target question q of the worked example in Section 2 back to its raw source.
Parsing. The raw source x is a PDF codebook (f = PDF), and its textual representation $z = \phi_f(x)$ contains the entry

```
V201 TRUSTGOV. Trust in national govt - how often right
1 Just about always   2 Most of the time   3 Only some
of the time
4 Never   8 Don't know   9 Refused
```

Extraction. Ext_θ turns this entry into a single-choice question object $q \in \mathcal{Q}^{\mathrm{raw}}$ with identifier V201, telegraphic text t_q = "Trust in national govt – how often right", and raw options

$$\mathcal{A}_q^{\mathrm{raw}} = (\text{Just about always}, \text{Most of the time}, \text{Only some of the time}, \text{Never}, \text{Don't know}, \text{Refused}).$$

Cleaning. "Don't know" and "Refused" belong to $\mathcal{M}$, so

$$\mathcal{A}_q = (\text{Just about always}, \text{Most of the time}, \text{Only some of the time}, \text{Never}), \qquad m_q = 4.$$

The reviewer returns $d_q = \mathrm{keep}$, and since q is single-choice with $2 \le m_q \le 12$, we have $q \in \mathcal{Q}^{\mathrm{clean}}$. By contrast, an item such as "Which of these news sources do you use? (Select all that apply)" is a multiple-response item and is excluded.
Rewriting. Rew_θ resolves the telegraphic label into $\tilde{t}_q$ = "How much of the time do you think you can trust the national government to do what is right?", keeping the four options and their order unchanged.
Normalization. The raw codes of the retained options are mapped as $g_q(1) = 0$, $g_q(2) = 1$, $g_q(3) = 2$, $g_q(4) = 3$, so the respondent r of Section 2, recorded with code 3, has the answer $y_{r,q} = 2$, while an answer coded 8 or 9 receives no index and is treated as unobserved.
Output. The tuple $(q, \tilde{t}_q, \mathcal{A}_q, g_q)$ enters $\mathcal{Q}_{\mathrm{all}}$, and her recorded code 3 enters $\mathcal{D}_{\mathrm{all}}$ as $(r, q, 2)$.

B.3 Worked Example: Persona Processing

Besides the trust question q above, $\mathcal{Q}_{\mathrm{all}}$ contains questions such as q^{sex} "What is your sex?", q^{age} "What is your age group?", q^{edu} "What is the highest level of education you have completed?", and q^{rel} "How often do you attend religious services?".
Classification. Cls_θ assigns $s_{q^{\mathrm{sex}}} = 0.98$, $s_{q^{\mathrm{age}}} = 0.97$, $s_{q^{\mathrm{edu}}} = 0.96$, $s_{q^{\mathrm{rel}}} = 0.85$, and $s_q = 0.03$. With $\tau_{\mathrm{per}} = 0.8$, the first four questions enter $\mathcal{Q}_{\mathrm{per}}$, whereas q remains in the target set $\mathcal{Q}$.
Verbalization. For the respondent r of Section 2, the observed persona answers give

$$\mathcal{P}_r = \{(q^{\mathrm{sex}}, \text{Female}),\ (q^{\mathrm{age}}, \text{40–44}),\ (q^{\mathrm{edu}}, \text{Bachelor's degree}),\ (q^{\mathrm{emp}}, \text{Full time, private sector}),$$
$$(q^{\mathrm{mar}}, \text{Married}),\ (q^{\mathrm{chi}}, \text{Two children}),\ (q^{\mathrm{urb}}, \text{Mid-sized city}),\ (q^{\mathrm{rel}}, \text{A few times a year})\}.$$

Profile writing. Prf_θ writes

> p_r = "A woman aged 40 to 44 living in a mid-sized city. She holds a bachelor's degree, works full time in the private sector, is married with two children, and attends religious services a few times a year."

Every statement in p_r is backed by a fact in $\mathcal{P}_r$. Her answer $y_{r,q} = 2$ to the trust question is not a persona fact, so it never appears in the profile.

B.4 Worked Example: Two-Axis Benchmark Split

Answer matrix. Consider a toy survey with $\mathcal{R} = \{r_1, \ldots, r_5\}$ and $\mathcal{Q} = \{q_1, \ldots, q_5\}$, where r_1 is the respondent r above and q_1 is the trust question q. Its observed answers $\mathcal{D}_\mathcal{Q}$ form the matrix below, in which "·" marks an unobserved cell.
Holdout sampling and partition. Sampling at ratio 0.2 gives $\mathcal{R}_{\text{un}} = \{r_5\}$ and $\mathcal{Q}_{\text{un}} = \{q_5\}$, so the rules divide the matrix into four blocks:

$\mathbf{Y} =$

	q_1	q_2	q_3	q_4	q_5
r_1	$\boxed{2}$	0	3	·	1
r_2	1	·	1	1	0
r_3	2	2	·	$\boxed{0}$	·
r_4	3	$\boxed{3}$	2	1	2
r_5	1	·	0	2	1

Test and training sets. The top-left block $\mathcal{D}_{\text{sese}}$ holds 13 observed cells; 20% of them, the three boxed cells, become Test-SS, and the other 10 form $\mathcal{D}_{\text{train}}$. The row of r_5 under q_1–q_4 gives Test-US (3 cells), the column of q_5 for r_1–r_4 gives Test-SU (3 cells), and the corner cell (r_5, q_5) gives Test-UU (1 cell). The running instance $(r, q) = (r_1, q_1)$ is a boxed cell and hence an MVI instance: the training set contains both r's answers to other questions (q_2 and q_3) and other respondents' answers to q (r_2 and r_4), which are exactly the two context sources used in the worked example of Section 2. Had q been sampled into $\mathcal{Q}_{\text{un}}$, the same cell would fall in Test-SU, and no answer to q would be available as context.

C Data Processing Prompts

↩ Back to Section 3.1

This section provides the complete system prompts used in the S^{2}Bench construction pipeline.

C.1 Question Extraction Prompt

The question extraction module converts heterogeneous source files (codebooks, questionnaires, metadata) into structured question objects. The LLM receives format-aware textual representations and extracts questions with their response options.

Question Extraction System Prompt

```
You are a professional survey codebook extraction assistant.

Extract survey questions and response options from the provided
source chunk. Return strict JSON with exactly this schema:
{
  "questions": [
    {
      "question_id": "variable name, item id, or question number if
 visible",
      "question_text": "full respondent-facing question text",
```

```
      "question_type": "Single Choice | Multiple Choice | Open Text
 | Numeric | Scale | Unknown",
      "question_options": ["1. Option A", "2. Option B"],
      "question_label": "short variable label if available",
      "source_notes": "brief uncertainty or parsing note"
    }
  ]
}

Rules:
- Prefer respondent-facing wording over administrative labels.
- Matrix/grid questions MUST be extracted row-by-row. Create one
question object per row item with the shared response scale.
- Classify all scale questions as "Single Choice".
- Return JSON only without Markdown or explanations.
```

C.2 Option Cleaning Prompt

After extraction, response options are canonicalized and deduplicated. An LLM classifier identifies invalid options (missing values, refusals, administrative codes) that should be removed before benchmark construction.

Option Cleaning System Prompt

```
You classify survey response options for cleaning.

Your job is to identify option labels that should be removed before
 using survey questions for modeling or benchmark construction.

Remove labels that are non-substantive missing-value or
administrative options, including but not limited to:
- don't know / do not know / DK
- no answer / not answered / unanswered
- refused / refusal / prefer not to answer
- not applicable / inapplicable / N/A
- missing / system missing / invalid / reserved / error
- skipped / not asked / breakoff / partial interview
- interviewer-only instructions or codebook placeholders

Keep substantive response labels, including:
- yes/no, agree/disagree, support/oppose, approve/disapprove
- neutral, neither, mixed, about the same
- numeric scale endpoints or middle labels
- other/specify when it represents a respondent-selectable
substantive answer

Return strict JSON:
{
  "invalid_options": [
    {
      "label_key": "exact label_key from input",
      "canonical_label": "label shown in input",
      "category": "missing | refused | not_applicable | skipped |
administrative | other_invalid",
      "confidence": 0.0,
      "reason": "short reason"
    }
  ],
  "notes": "short notes if needed"
}
```

```
Use only label_key values present in the input. If unsure whether a
 label is substantive, keep it out of invalid_options.
```

C.3 Question Filtering Prompt

The quality-control filter removes questions unsuitable for benchmark use. An LLM reviewer examines each question for completeness and usability.

Question Filtering System Prompt

```
You are a strict quality-control reviewer for survey benchmark
questions.

Decide whether each question must be removed. A kept question must
be a complete, self-contained single-choice question with a fully
specified, mutually interpretable response set.

REMOVE a question when any of these conditions applies:
1. UNRESOLVED PLACEHOLDERS: Contains template substitutions such as
 <programme>, <country>, [INSERT ...], {placeholder}, XXX,
underscores/blanks, or local/national wording that has not been
filled in.
2. INCOMPLETE OR ABBREVIATED OPTIONS: The response set is visibly
truncated or only illustrative, for example an ellipsis ("..." or
"..."), "etc.", "and so on", "examples include", or omitted
categories.
3. MULTI-FIELD FORM DISGUISED AS SINGLE CHOICE: The record actually
 requires several separate inputs or controls, such as separate day
/month/year drop-downs, an address, a start and end date, or
ranking several items.
4. OPEN-ENDED OR NUMERIC ENTRY DISGUISED AS CHOICE: The respondent
must type free text, a name, exact date, exact amount, exact age,
code, or other value and the listed "options" are field labels,
examples, units, ranges, or input instructions.
5. OPTIONS ARE NOT RESPONSES: The options are subquestion names,
variable/field names, questionnaire sections, column headings,
interviewer codes, programming commands, or routing instructions.
6. MISSING REQUIRED CONTEXT OR MATERIAL: The question cannot be
answered without a missing showcard, image, video, vignette,
preceding passage, randomized text, country-specific list, external
 lookup table, or another question whose content was not embedded
during rewriting.
7. ADMINISTRATIVE OR NON-RESPONDENT ITEM: It is an interviewer
observation, timing or paradata field, consent script, introduction,
 instruction, validation message, computed/derived variable, or
questionnaire routing item.
8. MALFORMED OR SEMANTICALLY BROKEN: The text is corrupted,
substantially incomplete, internally contradictory, has duplicate/
indistinguishable options, or its options cannot sensibly answer
the question.
9. TOO MANY OPTIONS: The question has more than 12 response options.


KEEP ordinary substantive survey questions, including those with
valid scales, yes/no answers, complete category lists, or a valid "
Other" option.

Return strict JSON:
{
  "questions": [
    {
      "question_id": "exact question_id from input",
```

```
      "remove_question": true,
      "category": "unresolved_placeholder | incomplete_options |
multi_field | open_ended | non_response_options | missing_context |
 administrative | malformed | too_many_options | other_invalid |
keep",
      "confidence": 0.0,
      "reason": "short evidence-based reason"
    }
  ]
}
```

C.4 Question Rewriting Prompt

The rewriting module converts telegraphic labels and matrix fragments into self-contained questions. Each question must be independently understandable without requiring questionnaire context.

Question Rewriting System Prompt

```
You are an expert survey methodologist. Your task is to rewrite
survey question text and response options so that each question is
fully self-contained and can be answered directly by its options.

You will receive a JSON payload with two fields:
- "context_questions": neighbouring questions provided for
reference ONLY. Do NOT rewrite them.
- "target_questions": the questions you must rewrite.

Rules:
1. SELF-CONTAINED: The rewritten question must be independently
understandable -- no references like "the following", "this aspect",
 "the above", "each of the items below", "as mentioned", etc.
2. INTERROGATIVE: The result must be a proper question ending with
"?". If the original is a statement or label (e.g. "Age", "
Employment status"), convert it to a natural question (e.g. "How
old are you?", "What is your current employment status?").
3. ONE ITEM ONLY: A target may be one row split from a matrix or
multi-select list. Rewrite it as a question about that row's sub-
item only. Remove list-level instructions that no longer apply to a
 standalone row.
4. PRESERVE MEANING: Keep the full semantic content that applies to
 the standalone item. Do not drop the sub-item topic.
5. REWRITE OPTIONS: Rewrite every response option so that it
directly and naturally answers the rewritten standalone question.
Convert extraction/codebook wording into respondent-facing
semantics when needed.
6. OPTION LABELS ONLY: Return only the rewritten option label,
without its numeric code or prefix. Keep exactly the same option
count and order. Never merge, add, remove, or reorder options.
7. PRESERVE SCALE LANGUAGE: If the original names specific scale
points, keep the scale's distinctions and meaning.
8. MINIMAL CHANGES: If the question or an option is already self-
contained and natural, copy that part unchanged. Set rewritten=
false only when neither the question text nor any option label
needs a semantic change.
9. LANGUAGE: Preserve the original language. Do not translate.
10. USE CONTEXT: Use context_questions to understand the survey
topic and maintain consistency in phrasing across neighbouring
questions, but only output rewrites for target_questions.

Return strict JSON -- only entries for target_questions, not
context_questions:
{
```

```
  "questions": [
    {
      "question_id": "<exact question_id from target_questions>",
      "rewritten_text": "<final question text ending with ?>",
      "rewritten_options": [
        {"option_index": 0, "rewritten_label": "<label only, no
code>"}
      ],
      "rewritten": true
    }
  ]
}

Set rewritten=true only when you actually changed the question text
 or at least one option label.
```

C.5 Persona Identification Prompt

The persona identifier classifies questions as background/demographic variables versus substantive attitudes or opinions.

Persona Identification System Prompt

```
You classify survey questions as persona/background variables.

A persona question describes relatively stable respondent
attributes or background context that can be used to characterize
the respondent. Mark as persona when the question is about
demographics, household or family background, education, employment
 status, occupation, income, region, migration/background,
citizenship, language, religion, ethnicity/race, gender/sex, age,
marital status, household composition, or similar profile
information.

Do NOT mark as persona when the question asks for attitudes,
opinions, preferences, evaluations, vote choices, policy positions,
 values, trust, satisfaction, media use opinions, candidate ratings,
 issue stances, or target outcomes to be predicted.

Return strict JSON:
{
  "questions": [
    {
      "question_id": "exact question_id from input",
      "is_persona": true,
      "category": "demographic | socioeconomic | household |
geography | religion | ethnicity | education | employment |
other_background | non_persona",
      "confidence": 0.0,
      "reason": "short evidence-based reason"
    }
  ],
  "notes": "optional short notes"
}

Use only question_id values present in the input. If unsure, prefer
 is_persona=false.
```

C.6 Persona Profiling Prompt

For each respondent, the profiling module generates a natural-language description from their answers to persona questions.

Persona Profiling System Prompt

```
You are a survey data analyst. Given a respondent's answers to
background/demographic questions, write a single fluent descriptive
 paragraph about this person.

Rules:
1. Write in the SAME language as the question text -- do not
translate.
2. Do NOT use "you are" or "I am". Write in third-person
descriptive style, e.g. "A 35-year-old male who ...".
3. Weave all provided facts naturally into one coherent paragraph.
Do not list them.
4. Do not speculate beyond what the data shows.
5. Return ONLY the paragraph text. No JSON, no headers, no extra
explanation.
```

D EVALUATION PLATFORM DETAILS

↩ Back to Section 3.2

A benchmark released as static files supports comparison only among the surveys and methods it already contains. We therefore open S^{2}Bench through a public platform with three functions. The first, *comparing existing methods*, is what a conventional benchmark offers: it shows how the evaluated methods perform on the released data. The other two extend it along its two axes. *Building new benchmarks* lets the data grow by applying the construction pipeline of Section 3.1 to new surveys, and *evaluating new methods* lets the set of compared methods grow by scoring any submitted method under the same conditions as every existing entry. With these extensions, neither the benchmark nor the comparison is fixed at release time.



Figure 7: Leaderboard of the evaluation platform (Function 1), showing model runs on the ANES 2024 wave under Test-SS. 1 Filters for survey dataset, wave, evaluation setting, and method. 2 Switch between the model view and the algorithm view. 3 Search by model, method, or publisher. 4 Sortable metric columns, where ↑ marks metrics for which higher is better and ↓ those for which lower is better. 5 Prompting method and publication time of each run.

Function 1: Comparing existing methods. The basic unit of comparison is a *run*, i.e., one evaluation of one method on one survey wave under one evaluation setting. Completed runs are displayed on a continuously updated *leaderboard* (Figure 7) with two views 2: the *model view* lists foundation models evaluated with a platform-defined prompting method, and the *algorithm view* lists complete

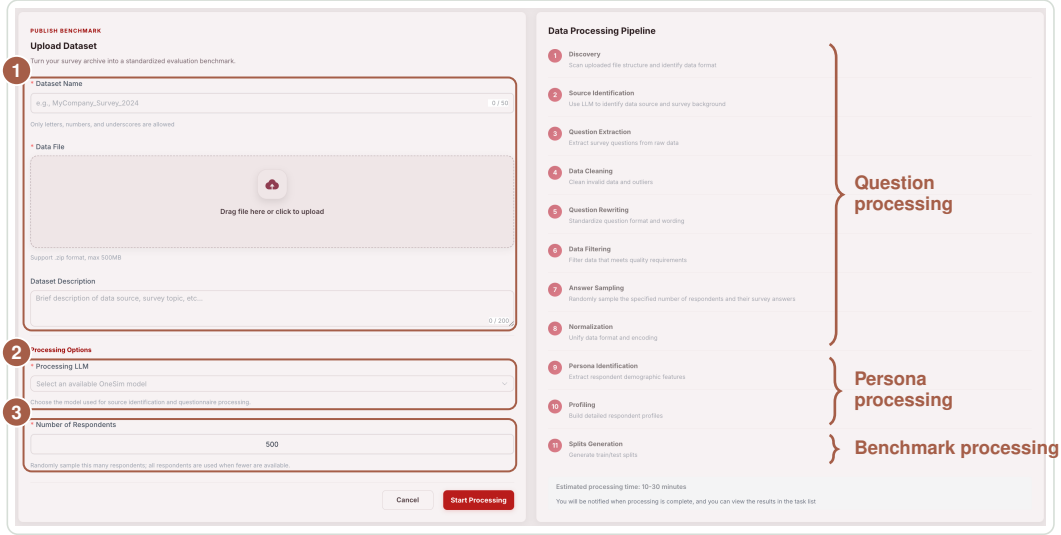

Figure 8: Benchmark construction service of the evaluation platform (Function 2). Left: the upload form, with 1 the survey archive (a ZIP file of up to 500 MB) and its name and description, 2 the processing LLM, and 3 the number of respondents to sample. Right: the eleven steps the service runs, grouped by the braces into the three stages of Section 3.1.

prediction systems; these correspond to the two submission tracks of Function 3. Users can filter runs by survey dataset (e.g., the World Values Survey), wave, evaluation setting (Test-SS, Test-US, Test-SU, or Test-UU), and method 1, and can search by model, method, or publisher 3. Each row records the method and its publication time 5 and can be sorted by three groups of metrics 4. Accuracy and weighted and macro F1 measure whether each predicted option matches the observed one, where weighted F1 averages over options in proportion to their frequency and macro F1 weights all options equally. Mean absolute error (MAE) measures how far the predicted option index lies from the observed one, which matters on ordinal scales. Earth mover's distance (EMD), Jensen–Shannon divergence, total variation distance, and KL divergence compare the distribution of predicted answers over the options with the distribution of human answers. A run appears on the leaderboard only after its owner explicitly publishes it and can later be withdrawn, which keeps private development runs out of the public record. Because results remain broken down by dataset, wave, and setting, foundation models and complete systems can be compared in one interface without collapsing results across survey families or generalization settings.

Function 2: Building new benchmarks. This function lets users turn a new survey into a benchmark in the format of S^2Bench directly on the platform, without installing or running the construction pipeline themselves. As shown in Figure 8, a user uploads a *survey archive* 1, i.e., a single ZIP file of up to 500 MB that bundles the raw sources of one survey (e.g., its codebook and data file), optionally adds a description, and sets two options: the *processing LLM* 2, which performs the LLM-based steps of the pipeline such as extraction, rewriting, and profile writing, and the number of respondents to sample 3, which bounds the size and cost of the resulting benchmark. The service then runs the three stages of Section 3.1 in order, displayed on the right of Figure 8 as eleven finer-grained steps: question processing extracts, cleans, rewrites, and normalizes the questions and answers; persona processing separates background questions from target questions and writes the respondent profiles; and benchmark processing partitions the answer matrix into the training set and the four test sets. While a job runs, stage-level status and logs show which stage it has reached and what each stage has produced, and a persistent task history records completed, running, and failed jobs. When processing finishes, the resulting benchmark is returned as a downloadable ZIP in the same standardized format as S^2Bench. The pipeline thus becomes a reusable service through which anyone can reproduce the benchmark or extend it with new survey sources.

Function 3: Evaluating new methods. Comparing a new method with published results would otherwise require reimplementing the data loading, prompts, and metrics. The platform removes

this cost by scoring every submission centrally under one protocol, summarized in Algorithm 1. Its central principle is that the human answers in the test sets remain *hidden*: a submitted method receives only *test instances*, each consisting of a respondent's persona profile and a target question with its options, and returns one predicted option index per instance, which the platform then scores against the hidden answers. Two submission tracks implement this protocol and differ in how much of the prediction process the submitter controls (Figures 9 and 10).

Algorithm 1 Centralized Evaluation Protocol

```
# E: hidden (persona, question) instances; H: hidden human answers
if track == MODEL_API:
    model_name, api_base, api_key = submission
    prompt_method = configuration
elif track == OPEN_SYSTEM:
    # Any internal optimization, algorithm, tool, or agent design is
    allowed.
    algorithm_name, predict_endpoint, token = submission

predictions = []
for persona, question in E:
    if track == MODEL_API:
        prompt = canonical_prompt(prompt_method, persona, question)
        answer = call_model_api(api_base, model_name, api_key, prompt)
    else:
        request = {"persona": persona, "question": question}
        answer = post(predict_endpoint, request)  # POST /predict
    contract
    predictions.append(validate(answer))

scores = central_score(predictions, H)
return aggregate(scores, by=("dataset", "wave", "split"))
```

(a) Model API track. This track evaluates a foundation model on its own, with everything else fixed by the platform. The submitter registers an *OpenAI-compatible API*, i.e., a web endpoint that accepts requests in the OpenAI chat-completion format, which most commercial providers and open-source serving frameworks support, by giving its base URL, API key, and model identifier (1). The submitter then selects a survey dataset, a wave, and an evaluation setting (2), together with one of the prompting methods of Section 4 (3). The platform permits only prompting methods whose context is available in the selected setting. For example, PRA supplies other respondents' answers to the target question, and in Test-SU no such answers exist in the training data because the question is unseen, so PRA cannot be selected there. For each instance, the platform builds the request from the canonical prompt of the chosen method and fixes the remaining procedure, including answer parsing and scoring. The API key is held only in service memory for the duration of the run and is never written to evaluation records or logs. This track offers a low-effort and reproducible way to compare newly released proprietary or open-weight models without requiring their developers to reimplement the benchmark.

(b) Open system track. This track evaluates a complete *prediction system*, i.e., any service that maps a persona profile and a target question to an option index, such as a fine-tuned model, a retrieval-based method, or a multi-agent workflow. For a selected dataset and wave (2), the submitter may download a *public bundle* (3) that contains the labeled training examples, the respondent persona profiles, the normalized questions and options, and documentation of the interface protocol. Using the bundle is optional, and a system may use any subset of it or none at all; test answers are never included. The submitter hosts the system as a web service and submits an algorithm name, the service's prediction URL, and, if needed, an authentication token (1). The platform then sends batches (4) of hidden test instances to the service through `POST /predict` and expects one option index per instance. Beyond this input–output contract, the system is treated as a black box, so submitters may transform the public data arbitrarily, fine-tune or otherwise adapt models, add statistical or retrieval-based components, combine multiple models and tools, or design specialized agent architectures and inference workflows.

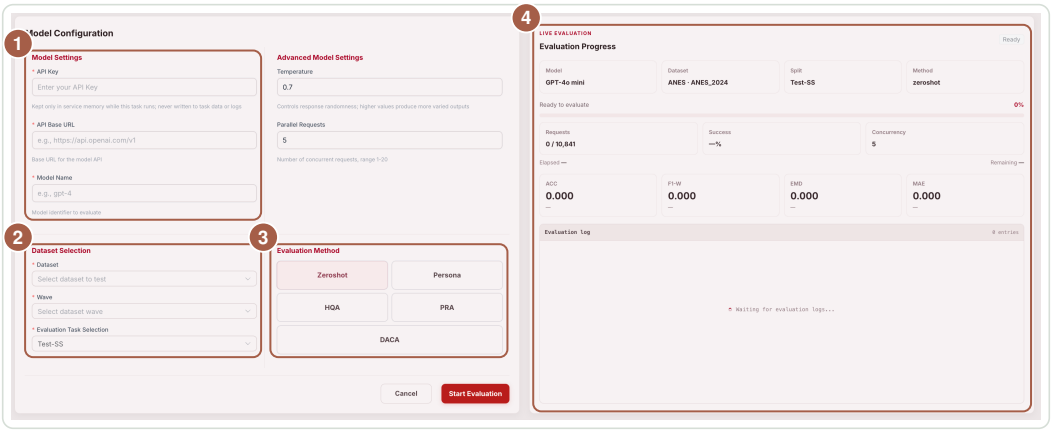

Figure 9: Model API track of the evaluation platform (Function 3a). 1 OpenAI-compatible API credentials, where the key is kept only in service memory. 2 Dataset, wave, and evaluation setting. 3 Prompting method. 4 Live progress, running metrics, and logs.

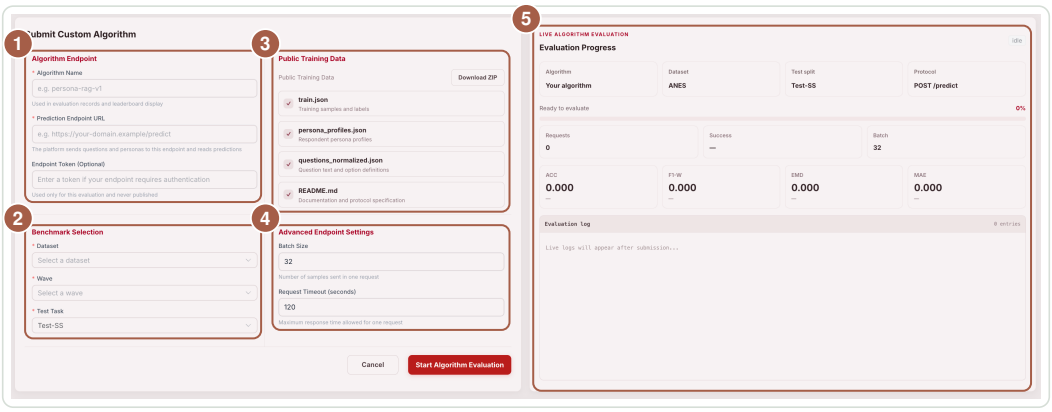

Figure 10: Open system track of the evaluation platform (Function 3b). 1 Algorithm name, prediction endpoint, and optional token. 2 Dataset, wave, and evaluation setting. 3 Optional public bundle with training data, persona profiles, normalized questions, and protocol documentation. 4 Batch size and timeout of the `POST /predict` requests. 5 Live progress, running metrics, and logs.

In both tracks, the service reports live progress, request counts, success rate, logs, and running metric estimates while a run is in progress (4 in Figure 9 and 5 in Figure 10). Once a run completes, its result files can be downloaded, and the owner decides whether to publish it to the leaderboard of Function 1 or withdraw it. Together, the two tracks separate *model capability* from *system capability*: the model API track compares models under a shared prompting framework, while the open system track measures the best performance attainable through end-to-end method design.

E Benchmark Details

↩ Back to Section 3.3

E.1 Benchmark Statistics

We characterize the resulting benchmark at three levels of granularity, each answering a different question about the data. At the *family level*, we ask what the benchmark covers: how much data each of the 15 survey families contributes and which regions and topics they span. At the *wave level*,

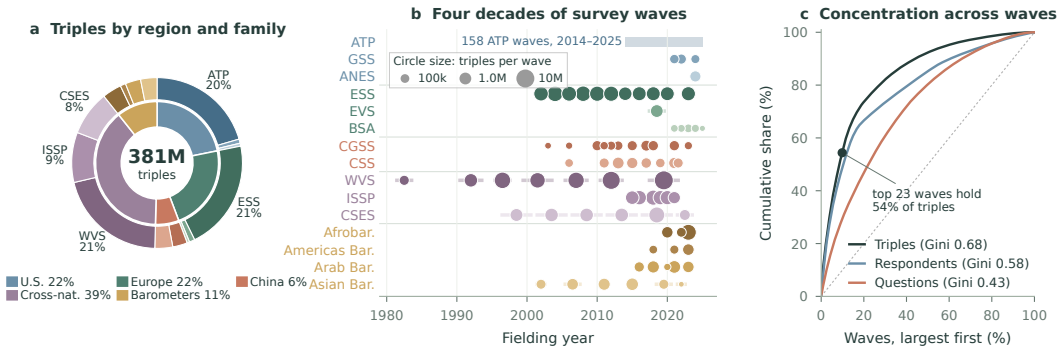

Figure 11: **Family-level coverage of S^{2}Bench. a**, Share of the 381M triples by region (inner ring) and survey family (outer ring); families below 4.5% are unlabeled. **b**, Fielding year of every wave, one row per family and grouped by region; circle area grows with the square root of the number of triples, horizontal bars mark multi-year fielding periods. Years follow the documentation of each survey programme, two CSS waves from 2021 are drawn side by side, and the 158 ATP waves are shown as a single span over 2014–2025 because per-wave fielding dates are not part of the released data. **c**, Lorenz curves of triples, respondents, and questions over the 234 waves sorted from largest to smallest; the dotted diagonal marks a perfectly even distribution.

we look inside the unit from which each answer matrix is built, a single wave, and ask whether its size and density support evaluation along both the respondent and the question axis. At the *split level*, we ask how the triples are finally distributed over the training set and the four test sets, and whether every evaluation setting is large enough for reliable comparison. The three levels are nested, since families consist of waves and each wave's answer matrix is partitioned into the five sets, so together they describe the benchmark from its overall composition, through its internal structure, to the data actually used for training and evaluation. Table 2 summarizes the family and split levels, Figures 11–13 visualize each level from several complementary angles, and the per-wave statistics are listed in Tables 4–15. The quality of the processed data itself is verified by a human evaluation in Section 4.2.

Family level: scale and coverage. Applying the pipeline to 15 publicly available survey families yields 234 waves and 381,442,527 valid respondent–question–answer triples, built from 4,075,760 respondent entries and 30,753 question entries summed over waves. These families span national election studies, general social surveys, cross-national value surveys, and regional barometers across Europe, Africa, Asia, the Middle East, and the Americas, and their content ranges over political attitudes, social values, religion, family life, economic conditions, social inequality, trust in institutions, environmental concern, and subjective well-being. As Table 2 shows, the data are not dominated by any single family. Families differ greatly in their number of waves: the American Trends Panel contributes 158 of the 234 waves, followed by the European Social Survey and the Chinese General Social Survey with 11 waves each, while the European Values Study and the American National Election Studies contribute a single wave each. Triples, however, are spread more evenly: the three largest families, the European Social Survey, the World Values Survey, and the American Trends Panel, hold 20.9%, 20.9%, and 20.5% of all triples, and the remaining twelve families hold the other 37.7%. Figure 11a regroups the same triples by region, which shows that no single society dominates either: cross-national programmes fielded in many countries contribute 38.9%, the United States and Europe contribute 21.8% and 22.4%, the four regional barometers covering Africa, Latin America, the Arab world, and Asia contribute 10.7%, and China contributes 6.2%. The benchmark is also broad in time (Figure 11b): its waves were fielded over more than four decades, from the first World Values Survey wave in 1981–1984 to waves collected in the mid-2020s, so predictors are tested on attitudes from very different historical contexts rather than on a single snapshot. Within this breadth, data volume is concentrated but not degenerate (Figure 11c): the largest 10% of waves (23 waves) hold 54% of all triples, giving a Gini coefficient of 0.68 over waves, whereas questions are spread more evenly (Gini 0.43), because the many small ATP waves each contribute their own questions. For this reason, Appendix I also breaks the results down by individual survey, so that the largest waves do not silently determine the overall ranking. A full taxonomy of the source surveys is given in Appendix E.2 and Table 3.

Table 2: Composition of S^{2}Bench by survey family. *Waves* counts survey administrations; *Resp.* and *Qs* count respondent entries and question entries summed over waves, so a panel respondent recurs once per wave in which they appear. *Triples* is the number of valid respondent–question–answer triples in millions, *Share* is each family's percentage of all triples, and the last five columns split these triples into the training set and the four test sets of Section 3.1, where SS, US, SU, and UU abbreviate Test-SS, Test-US, Test-SU, and Test-UU. Full survey names are given in Table 3.

Family	Composition					Split of triples (M)				
	Waves	Resp.	Qs	Triples	Share	Train	SS	US	SU	UU
ESS	11	518,597	3,053	79.85	20.93	40.43	10.11	12.63	13.35	3.34
WVS	7	446,767	1,587	79.73	20.90	40.97	10.24	12.81	12.57	3.14
ATP	158	1,137,734	14,664	78.15	20.49	40.27	10.07	12.58	12.19	3.05
ISSP	6	795,413	483	36.16	9.48	18.82	4.70	5.88	5.41	1.35
CSES	6	430,971	662	32.39	8.49	16.67	4.17	5.21	5.07	1.27
Afrobarometer	3	147,351	300	13.89	3.64	7.10	1.78	2.22	2.23	0.56
CSS	8	75,244	2,999	12.90	3.38	6.60	1.65	2.05	2.08	0.52
Asian Barometer	6	100,072	836	11.86	3.11	6.13	1.53	1.91	1.83	0.46
ArabBarometer	5	84,878	1,081	11.13	2.92	5.71	1.43	1.79	1.74	0.44
CGSS	11	105,518	2,151	10.60	2.78	5.44	1.36	1.70	1.69	0.42
EVS	1	59,437	76	3.98	1.04	2.02	0.51	0.63	0.66	0.16
AmericasBarometer	3	86,152	221	3.92	1.03	2.01	0.50	0.63	0.63	0.16
GSS	3	12,167	1,426	3.00	0.79	1.53	0.38	0.48	0.49	0.12
ANES	1	5,521	542	2.11	0.55	1.08	0.27	0.34	0.33	0.08
BSA	5	69,938	672	1.77	0.46	0.93	0.23	0.29	0.26	0.07
Total	**234**	**4,075,760**	**30,753**	**381.44**	**100.00**	**195.72**	**48.93**	**61.13**	**60.52**	**15.12**

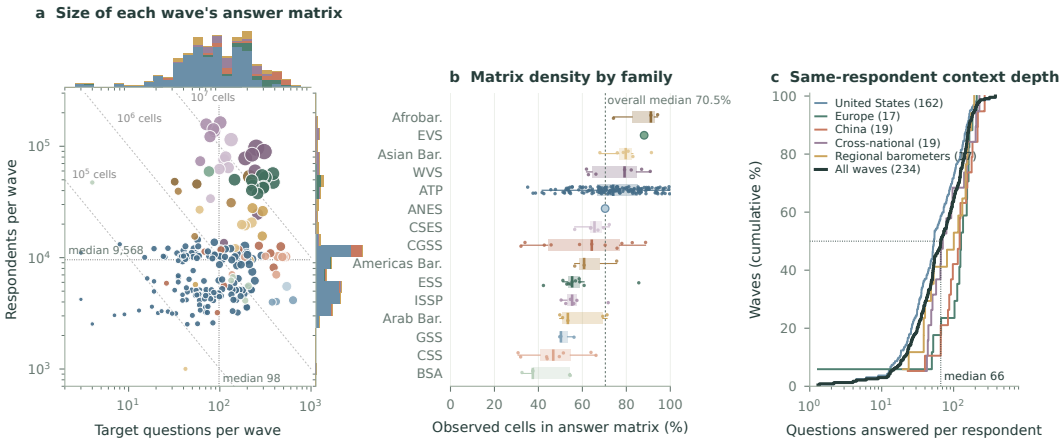

Figure 12: **Wave-level structure of the answer matrices. a**, Each circle is one wave, placed by its number of target questions and respondents (log scales), colored by family as in Figure 11, with area growing with the square root of its number of triples. Dashed lines mark matrices with 10^5, 10^6, and 10^7 cells, dotted lines mark the medians, and the marginal histograms are stacked by region. **b**, Density of each wave's answer matrix, i.e., the fraction of respondent–question cells that hold a valid answer, by family. Boxes show the interquartile range and the median, and each dot is one wave; the two single-wave families (EVS and ANES) are shown as outlined circles. **c**, Cumulative distribution over waves of the average number of questions answered per respondent, which bounds the same-respondent context available to a predictor; numbers in parentheses count waves.

Wave level: matrix size and density. Within this overall composition, individual waves differ widely along both axes, which is what makes the two-axis split informative rather than cosmetic. Figure 12a makes this visible, with waves ranging from small surveys with only a few questions or about a thousand respondents to large waves with hundreds of questions and tens of thousands of respondents. Quantitatively, a wave holds a median of 9,568 respondents (minimum 1,003, maximum 163,361)

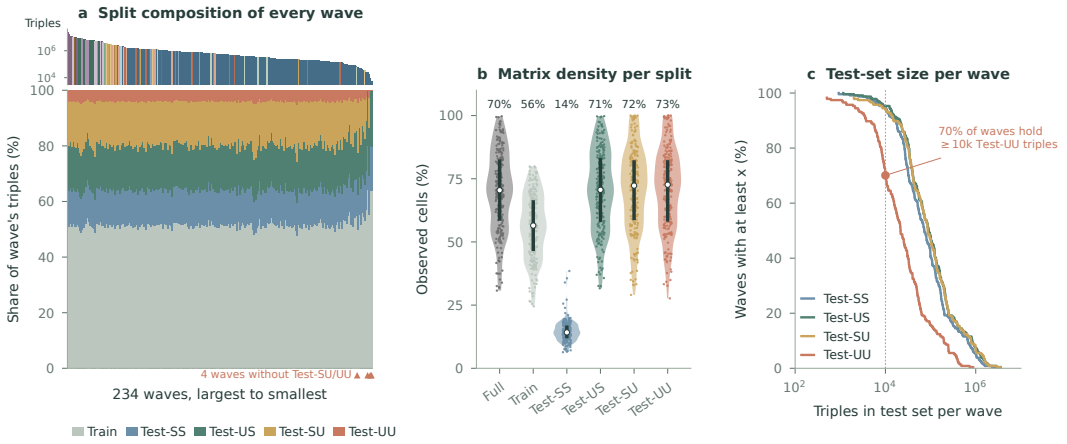

Figure 13: **Split-level statistics of the two-axis protocol. a**, Share of each wave's triples assigned to the training set and the four test sets, with waves sorted by size; the strip above gives each wave's number of triples (log scale, colored by family), and red triangles mark the four waves with only three or four questions, which are too small to hold out a question subset and therefore have no Test-SU or Test-UU. **b**, Matrix density of every wave within each set. Violins show the distribution over waves, dots are individual waves, black bars give the interquartile range, and white dots and the numbers above give the median. **c**, Fraction of waves whose test set contains at least a given number of triples (log scale); the dashed line marks 10,000 triples, and the red dot marks the share of waves whose Test-UU reaches this size.

and a median of 98 target questions (minimum 3, maximum 635); 109 waves have at least 10,000 respondents, and 113 have at least 100 questions. Accordingly, the median wave contributes 565,684 triples, against a minimum of 5,780 and a maximum of 22,179,427, spanning more than three orders of magnitude in matrix size. The two axes are also only loosely coupled: the 158 ATP waves form a horizontal band with 2,524 to 14,497 respondents but anywhere from 3 to 285 questions, whereas the cross-national programmes occupy the upper right with tens of thousands of respondents and hundreds of questions, so matrices with many respondents and few questions coexist with matrices with fewer respondents and many questions. The resulting answer matrices are dense but far from complete, with a median observed-cell rate of 70.5% and an interquartile range of 58.4% to 82.7%, so predicting held-out cells is a natural task rather than an artificial one. Density varies systematically by family (Figure 12b): its median ranges from 37.5% for British Social Attitudes and 46.7% for the Chinese Social Survey to 88.1% for the European Values Study and 91.2% for Afrobarometer, while the 158 ATP waves alone cover almost the full range from 35% to nearly 100%. Density also holds at the level of individual respondents (Figure 12c): in the median wave a respondent answers 66 target questions, and in 91% of waves the average respondent answers at least 20, which is what makes the same-respondent context of Section 2 available at all. This context is deepest in the European and Chinese surveys, with a median of 127 and 121 answered questions per respondent, and shallowest in the U.S. panel waves, with a median of 53, so the benefit of respondent history can itself be compared across regions. Full wave-level statistics are given in Table 4.

Split level: training and test sets. Partitioning these matrices with the two-axis protocol assigns 51.3% of all triples to training and the rest to the four test sets: 12.8% to Test-SS, 16.0% to Test-US, 15.9% to Test-SU, and 4.0% to Test-UU. This asymmetry follows from the design: Test-UU is the intersection of two 20% holdouts and is therefore the smallest, whereas Test-SS is a 20% record-level sample of the largest block. Even so, every setting is large in absolute terms, with the smallest, Test-UU, still holding 15,123,031 triples. The same ordering also holds within every family, where the training set is the largest block and Test-UU the smallest, so per-family comparisons across settings are not driven by one dominant source. Figure 13a shows that this consistency extends to individual waves: the training share of the median wave is 51.5%, and 90% of waves lie between 49.8% and 54.6%, regardless of wave size. The only exceptions are four waves with three or four questions (three ATP waves and the 2025 British Social Attitudes wave), where a 20% question holdout is empty, so these waves contribute to Test-SS and Test-US only. The split also shapes

the density of each set in a predictable way (Figure 13b): Test-US, Test-SU, and Test-UU keep the density of the full matrix (median 71%–73%), because each takes all observed cells of a respondent or question block, whereas the training set drops to 56% and Test-SS to 14%, because Test-SS removes a random 20% of the cells from the seen block. Consequently, in Test-SS a predictor always sees most of the same respondent's other answers, while in the unseen-respondent settings it sees none, which is exactly the contrast the two axes are designed to measure. Finally, the test sets are large within individual waves, not only in aggregate (Figure 13c): Test-SS, Test-US, and Test-SU contain at least 10,000 triples in 94% or more of waves, and even Test-UU does so in 70% of waves and contains at least 1,000 triples in 97% of them, so per-wave accuracies are estimated from enough answers to be stable. Every question carries between 2 and 12 options after cleaning, so accuracy is comparable across settings without rescaling.

E.2 Dataset Taxonomy

To systematically characterize the diversity of survey datasets included in S^2Bench, we adopt a multi-dimensional classification framework covering temporal structure, thematic focus, sampling unit, geographic scope, and target population. This taxonomy enables precise positioning of each dataset within the broader landscape of empirical social research and clarifies the range of survey methodologies our benchmark encompasses.

E.2.1 Temporal Structure

We classify datasets by their temporal design, following the standard typology in survey methodology. This dimension is mutually exclusive and collectively exhaustive (MECE) for traditional human-subject surveys, determined by asking: (1) Is data collected at multiple time points? (2) Are the same individuals tracked over time?

- **Single Cross-Sectional.** Data collected at a single time point, with no follow-up tracking or experimental intervention. Used for static snapshot analysis at a specific moment.
- **Repeated Cross-Sectional.** Data collected across multiple time points, but each wave samples a *different* independent group of respondents. Used for macro-level trend analysis without tracking individual changes.
- **Longitudinal Panel.** The *same* individuals, households, or organizations are tracked and re-interviewed across multiple time points. Designed for causal inference and micro-level trajectory analysis.

E.2.2 Thematic Focus

This dimension defines the primary research domain and core measurement content of each dataset, determining its disciplinary applicability (e.g., economics, education, sociology).

- **Politics & Governance.** Political attitudes, institutional trust, electoral behavior, public policy evaluation. Subfields: democracy, ideology, government efficacy, international relations.
- **Society & Values.** Social stratification, cultural identity, religion, gender norms, life satisfaction. Subfields: social mobility, civic morality, environmental awareness.
- **Economics & Finance.** Income distribution, household assets & liabilities, consumption & saving behavior, labor market performance. Subfields: microeconometrics, financial literacy, inequality.
- **Demography & Health.** Population structure (fertility, mortality, migration), physical/mental health, healthcare, aging. Subfields: epidemiology, elderly care, life course.
- **Education & Human Capital.** Educational attainment, cognitive/non-cognitive ability tests, school environment, educational equity. Subfields: student development, education economics, skill formation.
- **Comprehensive & Family Dynamics.** Multi-module coverage including all above themes, typically sampled at the household level, studying complex intra-family interactions. Subfields: interdisciplinary research, intergenerational transfers, family division of labor.

E.2.3 SAMPLING UNIT

This dimension specifies the smallest analytical unit in the dataset, defining the level at which observations are recorded.

- **Individual.** Each row represents a single person, recording their physical, psychological, behavioral, attitudinal, or economic characteristics. Used in psychology, political science, and micro labor economics.
- **Household.** Each row represents a family unit, recording total household income, assets, structure, and living conditions. Used in household finance, poverty measurement, and intergenerational mobility research.
- **Organization / Firm.** Each row represents an enterprise, school, nonprofit, or government agency, recording operational status, staffing, or financial statements. Used in industrial economics, management, and education policy evaluation.
- **Community / Village.** Data reflects the macro environment of a geographic or administrative community, such as local infrastructure, customs, price levels, or clan influence. Used in neighborhood effect studies and macro-micro moderation analysis.

E.2.4 GEOGRAPHIC SCOPE

This dimension characterizes the spatial coverage of the sampling frame, determining the level of geographic generalizability.

- **Global / Cross-national.** Standardized questionnaires fielded simultaneously in two or more sovereign countries/regions with cross-national comparability. Used in comparative politics, cross-cultural research, and globalization impact assessment.
- **National.** Sampling frame covers all major geographic regions within a single country, and weighted samples are representative of the national population. Used in national policy evaluation and domestic social change analysis.
- **Regional / Local.** Survey limited to specific provinces, states, cities, or targeted regions (e.g., contiguous poverty-stricken areas). Used in local policy pilot evaluation and regional economic or cultural research.

E.2.5 TARGET POPULATION

This dimension defines the age, demographic, or socioeconomic group that the survey specifically targets.

- **General Population.** Target group is typically all adult residents aged 18 and above in the country (or region), with no special age or occupational restrictions. Covers the general public with full legal capacity.
- **Children & Adolescents.** Questionnaires specifically designed for minors or youth in educational stages. Covers infants, primary/secondary school students, and college students.
- **Aging Population.** Questionnaires specifically designed for middle-aged and elderly groups (typically 45 or 50 years and above), focusing on aging and health. Covers retirees, empty-nesters, and the oldest-old.
- **Specific Groups.** Targeted sampling based on specific socioeconomic status, occupation, or household registration. Covers migrant populations (rural-to-urban migrants), specific industry labor forces, ethnic minorities, women, etc.

Table 3: Representative survey datasets organized by source family, with five-dimensional classification tags. Each tag color denotes a fixed classification dimension: data structure, core theme, unit of analysis, spatial scope, and target population.

Dataset	Description	Five-Dimensional Tags
Foundational comparative and national surveys		
ANES	The American National Election Studies is a long-running benchmark for studying U.S. electoral behavior, candidate evaluations, party identification, and political attitudes.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: National** **Target: General pop.**
GSS	The General Social Survey tracks U.S. social attitudes over time with highly comparable variables for research on values, religion, family, inequality, and social change.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: National** **Target: General pop.**
WVS	The World Values Survey provides cross-national measures of cultural values, democratic attitudes, gender norms, environmental beliefs, and self-expression values.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
ATP	The American Trends Panel is Pew Research Center's probability-based online panel for timely studies of U.S. public opinion, technology use, polarization, and global views.	**Panel** **Theme: Social & Values** **Unit: Individual** **Scope: National** **Target: General pop.**
ESS	The European Social Survey offers rigorously harmonized biennial data on European social attitudes, welfare preferences, immigration attitudes, trust, and work-family balance.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
ISSP	The International Social Survey Programme coordinates annual topical modules across member countries, enabling comparative research on inequality, environment, health, and other themes.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
CSES	The Comparative Study of Electoral Systems links post-election individual surveys with institutional context to study turnout, electoral systems, party competition, and political efficacy.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
EVS	The European Values Study focuses on European value change, secularization, labor ethics, family norms, and identity, with partial questionnaire overlap with WVS.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
Regional opinion barometers		
Americas Bar.	AmericasBarometer is a leading source for public opinion in Latin America and the Caribbean, especially on democracy, corruption, security, and rule of law.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
Afrobar.	Afrobarometer provides micro-level public opinion data across African countries on democracy, governance quality, political participation, public services, and international influence.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
Asian Bar.	Asian Barometer studies political values, democratic legitimacy, social capital, authority, and globalization across East, Southeast, and South Asian societies.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
Arab Bar.	Arab Barometer captures public opinion in the Middle East and North Africa on religion, public life, gender equality, youth unemployment, and political change.	**Repeated CS** **Theme: Politics & Governance** **Unit: Individual** **Scope: Cross-national** **Target: General pop.**
National panel, census, and specialized surveys		
BSA	British Social Attitudes is the U.K. counterpart to GSS, tracking views on welfare, class, the NHS, Brexit-related attitudes, and long-term social change.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: National** **Target: General pop.**

Continued on next page

Table 3 (continued)

Dataset	Description	Five-Dimensional Tags
CGSS	The Chinese General Social Survey is a national academic survey for studying social stratification, household registration, internet use, values, and social transformation in China.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: National** **Target: General pop.**
CSS	The Chinese Social Survey collects nationally representative data on stratification, public services, social trust, consumption, digital life, and public well-being in China.	**Repeated CS** **Theme: Social & Values** **Unit: Individual** **Scope: National** **Target: General pop.**

Table 4: Full-pair benchmark split statistics over 381,442,527 valid respondent–question pairs (Part 1/12). Full counts all valid pairs before splitting; Test-SS, Test-US, Test-SU, and Test-UU follow the two-axis split settings in Section 3.1. Density is the fraction of observed pairs among all user–question combinations in the split.

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ANES 2024	Full	2,105,203	5,521	542	381.3	3884.1	70.4%	ATP W163	Full	261,690	5,097	60	51.3	4361.5	85.6%
	Train	1,083,704	4,417	434	245.3	2497.0	56.5%		Train	134,835	4,078	48	33.1	2809.1	68.9%
	Test-UU	82,415	1,104	106	74.7	777.5	70.4%		Test-UU	10,224	1,019	12	10.0	852.0	83.6%
	Test-US	338,402	1,104	434	306.5	779.7	70.6%		Test-US	42,026	1,019	48	41.2	875.5	85.9%
	Test-SU	329,757	4,416	108	74.7	3053.3	69.1%		Test-SU	40,897	4,078	12	10.0	3408.1	83.6%
	Test-SS	270,925	4,415	434	61.4	624.3	14.1%		Test-SS	33,708	4,077	48	8.3	702.2	17.2%
ATP W1	Full	39,356	3,308	23	11.9	1711.1	51.7%	ATP W10	Full	72,528	3,147	30	23.0	2417.6	76.8%
	Train	21,444	2,647	19	8.1	1128.6	42.6%		Train	39,634	2,518	24	15.7	1651.4	65.6%
	Test-UU	1,141	518	4	2.2	285.2	55.1%		Test-UU	2,098	628	6	3.3	349.7	55.7%
	Test-US	6,707	661	19	10.1	353.0	53.4%		Test-US	12,377	629	24	19.7	515.7	82.0%
	Test-SU	4,704	2,091	4	2.2	1176.0	56.2%		Test-SU	8,511	2,508	6	3.4	1418.5	56.6%
	Test-SS	5,360	2,323	19	2.3	282.1	12.1%		Test-SS	9,908	2,481	24	4.0	412.8	16.6%
ATP W100	Full	734,868	14,497	57	50.7	12892.4	88.9%	ATP W101	Full	539,369	5,128	137	105.2	3937.0	76.8%
	Train	378,099	11,598	46	32.6	8219.5	70.9%		Train	277,036	4,103	110	67.5	2518.5	61.4%
	Test-UU	28,792	2,899	11	9.9	2617.5	90.3%		Test-UU	21,295	1,025	27	20.8	788.7	76.9%
	Test-US	118,104	2,899	46	40.7	2567.5	88.6%		Test-US	86,560	1,025	110	84.4	786.9	76.8%
	Test-SU	115,349	11,598	11	9.9	10486.3	90.4%		Test-SU	85,220	4,103	27	20.8	3156.3	76.9%
	Test-SS	94,524	11,597	46	8.2	2054.9	17.7%		Test-SS	69,258	4,103	110	16.9	629.6	15.3%
ATP W102	Full	210,150	10,236	38	20.5	5530.3	54.0%	ATP W103	Full	123,373	9,388	35	13.1	3524.9	37.5%
	Train	108,175	8,189	31	13.2	3489.5	42.6%		Train	67,268	7,510	28	9.0	2402.4	32.0%
	Test-UU	8,219	2,047	7	4.0	1174.1	57.4%		Test-UU	3,629	1,873	7	1.9	518.4	27.7%
	Test-US	33,765	2,047	31	16.5	1089.2	53.2%		Test-US	20,429	1,877	28	10.9	729.6	38.9%
	Test-SU	32,948	8,187	7	4.0	4706.9	57.5%		Test-SU	15,231	7,494	7	2.0	2175.9	29.0%
	Test-SS	27,043	7,961	31	3.4	872.4	11.0%		Test-SS	16,816	6,643	28	2.5	600.6	9.0%
ATP W104	Full	248,629	10,441	43	23.8	5782.1	55.4%	ATP W105	Full	242,600	3,581	70	67.7	3465.7	96.8%
	Train	131,477	8,353	35	15.7	3756.5	45.0%		Train	123,582	2,865	56	43.1	2206.8	77.0%
	Test-UU	8,654	2,088	8	4.1	1081.8	51.8%		Test-UU	9,913	716	14	13.8	708.1	98.9%
	Test-US	41,235	2,088	35	19.7	1178.1	56.4%		Test-US	38,678	716	56	54.0	690.7	96.5%
	Test-SU	34,394	8,350	8	4.1	4299.2	51.5%		Test-SU	39,532	2,865	14	13.8	2823.7	98.6%
	Test-SS	32,869	8,233	35	4.0	939.1	11.4%		Test-SS	30,895	2,865	56	10.8	551.7	19.3%
ATP W106	Full	336,967	10,155	51	33.2	6607.2	65.1%	ATP W107	Full	179,427	5,074	55	35.4	3262.3	64.3%
	Train	168,674	8,124	41	20.8	4114.0	50.6%		Train	94,015	4,060	44	23.2	2136.7	52.6%
	Test-UU	14,755	2,031	10	7.3	1475.5	72.6%		Test-UU	6,518	1,012	11	6.4	592.5	58.6%
	Test-US	52,622	2,031	41	25.9	1283.5	63.2%		Test-US	29,231	1,014	44	28.8	664.3	65.5%
	Test-SU	58,748	8,124	10	7.2	5874.8	72.3%		Test-SU	26,160	4,060	11	6.4	2378.2	58.6%
	Test-SS	42,168	8,062	41	5.2	1028.5	12.8%		Test-SS	23,503	4,054	44	5.8	534.2	13.2%
ATP W108	Full	190,137	10,282	24	18.5	7922.4	77.1%	ATP W109	Full	225,910	10,188	38	22.2	5945.0	58.4%
	Train	109,977	8,226	20	13.4	5498.9	66.8%		Train	116,861	8,151	31	14.3	3769.7	46.2%
	Test-UU	3,703	2,046	4	1.8	925.8	45.2%		Test-UU	8,684	2,037	7	4.3	1240.6	60.9%
	Test-US	34,190	2,056	20	16.6	1709.5	83.1%		Test-US	36,461	2,037	31	17.9	1176.2	57.7%
	Test-SU	14,773	8,159	4	1.8	3693.2	45.3%		Test-SU	34,689	8,142	7	4.3	4955.6	60.9%
	Test-SS	27,494	8,006	20	3.4	1374.7	17.2%		Test-SS	29,215	7,969	31	3.7	942.4	11.8%
ATP W11	Full	18,451	3,037	7	6.1	2635.9	86.8%	ATP W110	Full	291,301	6,174	67	47.2	4347.8	70.4%
	Train	10,212	2,423	6	4.2	1702.0	70.2%		Train	152,940	4,940	54	31.0	2832.2	57.3%
	Test-UU	504	504	1	1.0	504.0	100.0%		Test-UU	10,473	1,232	13	8.5	805.6	65.4%
	Test-US	3,180	607	6	5.2	530.0	87.3%		Test-US	47,607	1,234	54	38.6	881.6	71.4%
	Test-SU	2,002	2,002	1	1.0	2002.0	100.0%		Test-SU	42,047	4,938	13	8.5	3234.4	65.5%
	Test-SS	2,553	1,663	6	1.5	425.5	25.6%		Test-SS	38,234	4,938	54	7.7	708.0	14.3%
ATP W111	Full	144,623	6,034	35	24.0	4132.1	68.5%	ATP W112	Full	223,340	12,147	27	18.4	8271.9	68.1%
	Train	73,016	4,828	28	15.1	2607.7	54.0%		Train	110,832	9,718	22	11.4	5037.8	51.8%
	Test-UU	6,077	1,206	7	5.0	868.1	72.0%		Test-UU	10,043	2,429	5	4.1	2008.6	82.7%
	Test-US	22,959	1,206	28	19.0	820.0	68.0%		Test-US	34,329	2,429	22	14.1	1560.4	64.2%
	Test-SU	24,317	4,827	7	5.0	3473.9	72.0%		Test-SU	40,428	9,716	5	4.2	8085.6	83.2%
	Test-SS	18,254	4,721	28	3.9	651.9	13.8%		Test-SS	27,708	9,214	22	3.0	1259.5	13.7%
ATP W113	Full	211,986	7,646	51	27.7	4156.6	54.4%	ATP W114	Full	537,700	10,588	57	50.8	9433.3	89.1%
	Train	105,404	6,110	41	17.3	2570.8	42.1%		Train	275,763	8,471	46	32.6	5994.8	70.8%
	Test-UU	9,386	1,528	10	6.1	938.6	61.4%		Test-UU	21,396	2,117	11	10.1	1945.1	91.9%
	Test-US	33,817	1,529	41	22.1	824.8	53.9%		Test-US	85,699	2,117	46	40.5	1863.0	88.0%
	Test-SU	37,029	6,111	10	6.1	3702.9	60.6%		Test-SU	85,902	8,471	11	10.1	7809.3	92.2%
	Test-SS	26,350	5,848	41	4.5	642.7	11.0%		Test-SS	68,940	8,470	46	8.1	1498.7	17.7%

Table 5: Full-pair benchmark split statistics (Part 2/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W115	Full	191,921	3,757	60	51.1	3198.7	85.1%	ATP W116	Full	694,224	5,098	177	136.2	3922.2	76.9%
	Train	98,296	3,006	48	32.7	2047.8	68.1%		Train	358,779	4,079	142	88.0	2526.6	61.9%
	Test-UU	7,630	751	12	10.2	635.8	84.7%		Test-UU	26,718	1,019	35	26.2	763.4	74.9%
	Test-US	30,794	751	48	41.0	641.5	85.4%		Test-US	111,962	1,019	142	109.9	788.5	77.4%
	Test-SU	30,628	3,005	12	10.2	2552.3	84.9%		Test-SU	107,071	4,079	35	26.2	3059.2	75.0%
	Test-SS	24,573	3,004	48	8.2	511.9	17.0%		Test-SS	89,694	4,079	142	22.0	631.6	15.5%
ATP W117	Full	430,993	11,377	74	37.9	5824.2	51.2%	ATP W118	Full	245,067	5,132	81	47.8	3025.5	59.0%
	Train	214,794	9,102	60	23.6	3579.9	39.3%		Train	125,542	4,106	65	30.6	1931.4	47.0%
	Test-UU	19,100	2,275	14	8.4	1364.3	60.0%		Test-UU	9,842	1,026	16	9.6	615.1	60.0%
	Test-US	66,644	2,275	60	29.3	1110.7	48.8%		Test-US	37,476	1,026	65	36.5	576.6	56.2%
	Test-SU	76,757	9,102	14	8.4	5482.6	60.2%		Test-SU	40,822	4,106	16	9.9	2551.4	62.1%
	Test-SS	53,698	9,069	60	5.9	895.0	9.9%		Test-SS	31,385	4,095	65	7.7	482.8	11.8%
ATP W119	Full	392,813	11,004	49	35.7	8016.6	72.9%	ATP W12	Full	27,687	3,094	9	8.9	3076.3	99.4%
	Train	205,328	8,803	40	23.3	5133.2	58.3%		Train	15,751	2,476	8	6.4	1968.9	79.5%
	Test-UU	14,437	2,197	9	6.6	1604.1	73.0%		Test-UU	616	616	1	1.0	616.0	100.0%
	Test-US	64,120	2,200	40	29.1	1603.0	72.9%		Test-US	4,923	618	8	8.0	615.4	99.6%
	Test-SU	57,596	8,798	9	6.5	6399.6	72.7%		Test-SU	2,460	2,460	1	1.0	2460.0	100.0%
	Test-SS	51,332	8,680	40	5.9	1283.3	14.8%		Test-SS	3,937	2,058	8	1.9	492.1	23.9%
ATP W120	Full	247,830	5,152	71	48.1	3490.6	67.8%	ATP W121	Full	638,333	5,902	143	108.2	4463.9	75.6%
	Train	124,292	4,122	57	30.2	2180.6	52.9%		Train	324,642	4,722	115	68.8	2823.0	59.8%
	Test-UU	10,711	1,030	14	10.4	765.1	74.3%		Test-UU	26,067	1,180	28	22.1	931.0	78.9%
	Test-US	38,813	1,030	57	37.7	680.9	66.1%		Test-US	101,711	1,180	115	86.2	884.4	75.0%
	Test-SU	42,942	4,122	14	10.4	3067.3	74.4%		Test-SU	104,753	4,722	28	22.2	3741.2	79.2%
	Test-SS	31,072	4,118	57	7.5	545.1	13.2%		Test-SS	81,160	4,722	115	17.2	705.7	14.9%
ATP W122	Full	248,540	4,742	59	52.4	4212.5	88.8%	ATP W123	Full	1,345,997	10,701	154	125.8	8740.2	81.7%
	Train	127,602	3,794	48	33.6	2658.4	70.1%		Train	701,272	8,561	124	81.9	5655.4	66.1%
	Test-UU	9,831	947	11	10.4	893.7	94.4%		Test-UU	50,025	2,140	30	23.4	1667.5	77.9%
	Test-US	39,848	948	48	42.0	830.2	87.6%		Test-US	219,545	2,140	124	102.6	1770.5	82.7%
	Test-SU	39,359	3,792	11	10.4	3578.1	94.4%		Test-SU	199,838	8,561	30	23.3	6661.3	77.8%
	Test-SS	31,900	3,791	48	8.4	664.6	17.5%		Test-SS	175,317	8,561	124	20.5	1413.8	16.5%
ATP W124	Full	559,498	3,576	198	156.5	2825.7	79.0%	ATP W125	Full	823,487	5,079	188	162.1	4380.2	86.2%
	Train	290,708	2,861	159	101.6	1828.4	63.9%		Train	419,168	4,064	151	103.1	2775.9	68.3%
	Test-UU	21,049	715	39	29.4	539.7	75.5%		Test-UU	33,696	1,015	37	33.2	910.7	89.7%
	Test-US	90,395	715	159	126.4	568.5	79.5%		Test-US	131,099	1,015	151	129.2	868.2	85.5%
	Test-SU	84,670	2,861	39	29.6	2171.0	75.9%		Test-SU	134,733	4,064	37	33.2	3641.4	89.6%
	Test-SS	72,676	2,861	159	25.4	457.1	16.0%		Test-SS	104,791	4,064	151	25.8	694.0	17.1%
ATP W126	Full	765,191	5,073	162	150.8	4723.4	93.1%	ATP W127	Full	859,681	5,101	187	168.5	4597.2	90.1%
	Train	392,603	4,059	130	96.7	3020.0	74.4%		Train	444,410	4,081	150	108.9	2962.7	72.6%
	Test-UU	30,329	1,014	32	29.9	947.8	93.5%		Test-UU	32,952	1,020	37	32.3	890.6	87.3%
	Test-US	122,312	1,014	130	120.6	940.9	92.8%		Test-US	138,951	1,020	150	136.2	926.3	90.8%
	Test-SU	121,797	4,059	32	30.0	3806.2	93.8%		Test-SU	132,266	4,081	37	32.4	3574.8	87.6%
	Test-SS	98,150	4,059	130	24.2	755.0	18.6%		Test-SS	111,102	4,081	150	27.2	740.7	18.1%
ATP W128	Full	1,534,966	10,329	173	148.6	8872.6	85.9%	ATP W129	Full	626,142	5,115	163	122.4	3841.4	75.1%
	Train	792,951	8,264	139	96.0	5704.7	69.0%		Train	322,428	4,092	131	78.8	2461.3	60.1%
	Test-UU	59,252	2,065	34	28.7	1742.7	84.4%		Test-UU	24,463	1,023	32	23.9	764.5	74.7%
	Test-US	247,395	2,065	139	119.8	1779.8	86.2%		Test-US	100,891	1,023	131	98.6	770.2	75.3%
	Test-SU	237,131	8,264	34	28.7	6974.4	84.4%		Test-SU	97,754	4,092	32	23.9	3054.8	74.7%
	Test-SS	198,237	8,264	139	24.0	1426.2	17.3%		Test-SS	80,606	4,092	131	19.7	615.3	15.0%
ATP W13	Full	313,428	4,787	97	65.5	3231.2	67.5%	ATP W130	Full	1,049,662	8,480	154	123.8	6816.0	80.4%
	Train	158,404	3,830	78	41.4	2030.8	53.0%		Train	544,184	6,784	124	80.2	4388.6	64.7%
	Test-UU	13,210	957	19	13.8	695.3	72.7%		Test-UU	39,919	1,696	30	23.5	1330.6	78.5%
	Test-US	49,859	957	78	52.1	639.2	66.8%		Test-US	170,448	1,696	124	100.5	1374.6	81.0%
	Test-SU	52,355	3,830	19	13.7	2755.5	71.9%		Test-SU	159,066	6,784	30	23.4	5302.2	78.2%
	Test-SS	39,600	3,829	78	10.3	507.7	13.3%		Test-SS	136,045	6,784	124	20.1	1097.1	16.2%
ATP W131	Full	847,422	5,057	207	167.6	4093.8	81.0%	ATP W132	Full	1,513,482	11,201	161	135.1	9400.5	83.9%
	Train	436,425	4,046	166	107.9	2629.1	65.0%		Train	766,948	8,961	129	85.6	5945.3	66.3%
	Test-UU	33,056	1,011	41	32.7	806.2	79.7%		Test-UU	63,149	2,240	32	28.2	1973.4	88.1%
	Test-US	135,401	1,011	166	133.9	815.7	80.7%		Test-US	239,624	2,240	129	107.0	1857.6	82.9%
	Test-SU	133,434	4,046	41	33.0	3254.5	80.4%		Test-SU	252,025	8,961	32	28.1	7875.8	87.9%
	Test-SS	109,106	4,046	166	27.0	657.3	16.2%		Test-SS	191,736	8,961	129	21.4	1486.3	16.6%

Table 6: Full-pair benchmark split statistics (Part 3/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W133	Full	949,757	11,945	103	79.5	9220.9	77.2%	ATP W134	Full	1,329,880	7,470	227	178.0	5858.5	78.4%
	Train	498,898	9,556	83	52.2	6010.8	62.9%		Train	693,597	5,976	182	116.1	3811.0	63.8%
	Test-UU	33,789	2,389	20	14.1	1689.5	70.7%		Test-UU	49,291	1,494	45	33.0	1095.4	73.3%
	Test-US	156,464	2,389	83	65.5	1885.1	78.9%		Test-US	217,539	1,494	182	145.6	1195.3	80.0%
	Test-SU	135,882	9,556	20	14.2	6794.1	71.1%		Test-SU	196,054	5,976	45	32.8	4356.8	72.9%
	Test-SS	124,724	9,556	83	13.1	1502.7	15.7%		Test-SS	173,399	5,976	182	29.0	952.7	15.9%
ATP W135	Full	1,179,571	8,842	160	133.4	7372.3	83.4%	ATP W137	Full	400,246	4,512	169	88.7	2368.3	52.5%
	Train	607,605	7,074	128	85.9	4746.9	67.1%		Train	206,606	3,610	136	57.2	1519.2	42.1%
	Test-UU	46,102	1,768	32	26.1	1440.7	81.5%		Test-UU	15,502	902	33	17.2	469.8	52.1%
	Test-US	189,868	1,768	128	107.4	1483.3	83.9%		Test-US	64,783	902	136	71.8	476.3	52.8%
	Test-SU	184,095	7,074	32	26.0	5753.0	81.3%		Test-SU	61,704	3,610	33	17.1	1869.8	51.8%
	Test-SS	151,901	7,074	128	21.5	1186.7	16.8%		Test-SS	51,651	3,610	136	14.3	379.8	10.5%
ATP W138	Full	920,680	5,078	194	181.3	4745.8	93.5%	ATP W139	Full	451,008	5,203	106	86.7	4254.8	81.8%
	Train	479,876	4,063	156	118.1	3076.1	75.7%		Train	237,469	4,163	85	57.0	2793.8	67.1%
	Test-UU	34,193	1,015	38	33.7	899.8	88.7%		Test-UU	16,011	1,040	21	15.4	762.4	73.3%
	Test-US	149,558	1,015	156	147.3	958.7	94.5%		Test-US	74,517	1,040	85	71.7	876.7	84.3%
	Test-SU	137,084	4,063	38	33.7	3607.5	88.8%		Test-SU	63,644	4,163	21	15.3	3030.7	72.8%
	Test-SS	119,969	4,063	156	29.5	769.0	18.9%		Test-SS	59,367	4,163	85	14.3	698.4	16.8%
ATP W14	Full	86,135	4,643	39	18.6	2208.6	47.6%	ATP W140	Full	725,984	5,140	222	141.2	3270.2	63.6%
	Train	47,243	3,668	32	12.9	1476.3	40.2%		Train	372,995	4,112	178	90.7	2095.5	51.0%
	Test-UU	2,513	847	7	3.0	359.0	42.4%		Test-UU	28,590	1,028	44	27.8	649.8	63.2%
	Test-US	14,308	928	32	15.4	447.1	48.2%		Test-US	116,573	1,028	178	113.4	654.9	63.7%
	Test-SU	10,261	3,420	7	3.0	1465.9	42.9%		Test-SU	114,578	4,112	44	27.9	2604.0	63.3%
	Test-SS	11,810	3,296	32	3.6	369.1	11.2%		Test-SS	93,248	4,112	178	22.7	523.9	12.7%
ATP W141	Full	1,012,160	5,146	220	196.7	4600.7	89.4%	ATP W142	Full	1,021,232	10,133	146	100.8	6994.7	69.0%
	Train	520,820	4,117	176	126.5	2959.2	71.9%		Train	514,349	8,107	117	63.4	4396.1	54.2%
	Test-UU	39,703	1,029	44	38.6	902.3	87.7%		Test-UU	43,514	2,026	29	21.5	1500.5	74.1%
	Test-US	162,243	1,029	176	157.7	921.8	89.6%		Test-US	160,824	2,026	117	79.4	1374.6	67.8%
	Test-SU	159,189	4,117	44	38.7	3617.9	87.9%		Test-SU	173,958	8,107	29	21.5	5998.6	74.0%
	Test-SS	130,205	4,117	176	31.6	739.8	18.0%		Test-SS	128,587	8,107	117	15.9	1099.0	13.6%
ATP W143	Full	2,062,340	12,693	188	162.5	10969.9	86.4%	ATP W144	Full	1,046,171	10,454	285	100.1	3670.8	35.1%
	Train	1,071,919	10,155	151	105.6	7098.8	69.9%		Train	543,311	8,364	228	65.0	2382.9	28.5%
	Test-UU	77,524	2,538	37	30.5	2095.2	82.6%		Test-UU	39,631	2,090	57	19.0	695.3	33.3%
	Test-US	335,652	2,538	151	132.3	2222.9	87.6%		Test-US	170,154	2,090	228	81.4	746.3	35.7%
	Test-SU	309,266	10,155	37	30.5	8358.5	82.3%		Test-SU	157,248	8,363	57	18.8	2758.7	33.0%
	Test-SS	267,979	10,155	151	26.4	1774.7	17.5%		Test-SS	135,827	8,361	228	16.2	595.7	7.1%
ATP W145	Full	593,333	3,600	195	164.8	3042.7	84.5%	ATP W146	Full	1,154,540	8,709	210	132.6	5497.8	63.1%
	Train	298,712	2,880	156	103.7	1914.8	66.5%		Train	600,585	6,968	168	86.2	3574.9	51.3%
	Test-UU	25,386	720	39	35.3	650.9	90.4%		Test-UU	43,009	1,741	42	24.7	1024.0	58.8%
	Test-US	93,685	720	156	130.1	600.5	83.4%		Test-US	187,772	1,741	168	107.9	1117.7	64.2%
	Test-SU	100,873	2,880	39	35.0	2586.5	89.8%		Test-SU	173,028	6,968	42	24.8	4119.7	59.1%
	Test-SS	74,677	2,880	156	25.9	478.7	16.6%		Test-SS	150,146	6,968	168	21.5	893.7	12.8%
ATP W147	Full	314,696	3,312	138	95.0	2280.4	68.9%	ATP W148	Full	1,001,318	8,638	171	115.9	5855.7	67.8%
	Train	159,482	2,650	111	60.2	1436.8	54.2%		Train	523,056	6,911	137	75.7	3817.9	55.2%
	Test-UU	13,070	662	27	19.7	484.1	73.1%		Test-UU	36,809	1,727	34	21.3	1082.6	62.7%
	Test-US	50,109	662	111	75.7	451.4	68.2%		Test-US	164,024	1,727	137	95.0	1197.3	69.3%
	Test-SU	52,165	2,650	27	19.7	1932.0	72.9%		Test-SU	146,666	6,911	34	21.2	4313.7	62.4%
	Test-SS	39,870	2,650	111	15.0	359.2	13.6%		Test-SS	130,763	6,911	137	18.9	954.5	13.8%
ATP W149	Full	1,343,077	9,424	201	142.5	6682.0	70.9%	ATP W15	Full	429,136	4,726	97	90.8	4424.1	93.6%
	Train	682,953	7,540	161	90.6	4241.9	56.3%		Train	219,883	3,781	78	58.2	2819.0	74.6%
	Test-UU	55,383	1,884	40	29.4	1384.6	73.5%		Test-UU	17,111	945	19	18.1	900.6	95.3%
	Test-US	212,649	1,884	161	112.9	1320.8	70.1%		Test-US	68,693	945	78	72.7	880.7	93.2%
	Test-SU	221,354	7,540	40	29.4	5533.9	73.4%		Test-SU	68,479	3,781	19	18.1	3604.2	95.3%
	Test-SS	170,738	7,540	161	22.6	1060.5	14.1%		Test-SS	54,970	3,780	78	14.5	704.7	18.6%
ATP W150	Full	634,415	10,658	85	59.5	7463.7	70.0%	ATP W151	Full	1,341,750	9,201	156	145.8	8601.0	93.5%
	Train	324,383	8,527	68	38.0	4770.3	55.9%		Train	684,756	7,361	125	93.0	5478.0	74.4%
	Test-UU	25,502	2,131	17	12.0	1500.1	70.4%		Test-UU	54,357	1,840	31	29.5	1753.5	95.3%
	Test-US	101,891	2,131	68	47.8	1498.4	70.3%		Test-US	213,936	1,840	125	116.3	1711.5	93.0%
	Test-SU	101,544	8,527	17	11.9	5973.2	70.1%		Test-SU	217,512	7,361	31	29.5	7016.5	95.3%
	Test-SS	81,095	8,526	68	9.5	1192.6	14.0%		Test-SS	171,189	7,361	125	23.3	1369.5	18.6%

Table 7: Full-pair benchmark split statistics (Part 4/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W152	Full	518,143	5,410	128	95.8	4048.0	74.8%	ATP W153	Full	1,545,341	9,720	226	159.0	6837.8	70.3%
	Train	268,273	4,328	103	62.0	2604.6	60.2%		Train	792,135	7,776	181	101.9	4376.4	56.3%
	Test-UU	19,845	1,082	25	18.3	793.8	73.4%		Test-UU	61,434	1,944	45	31.6	1365.2	70.2%
	Test-US	84,344	1,082	103	78.0	818.9	75.7%		Test-US	247,747	1,944	181	127.4	1368.8	70.4%
	Test-SU	78,613	4,328	25	18.2	3144.5	72.7%		Test-SU	245,992	7,776	45	31.6	5466.5	70.3%
	Test-SS	67,068	4,328	103	15.5	651.1	15.0%		Test-SS	198,033	7,776	181	25.5	1094.1	14.1%
ATP W154	Full	1,255,338	6,204	242	202.3	5187.3	83.6%	ATP W155	Full	1,002,494	9,680	106	103.6	9457.5	97.7%
	Train	646,820	4,964	194	130.3	3334.1	67.2%		Train	514,074	7,744	85	66.4	6047.9	78.1%
	Test-UU	49,035	1,240	48	39.5	1021.6	82.4%		Test-UU	39,862	1,936	21	20.6	1898.2	98.0%
	Test-US	202,038	1,240	194	162.9	1041.4	84.0%		Test-US	160,477	1,936	85	82.9	1888.0	97.5%
	Test-SU	195,741	4,964	48	39.4	4077.9	82.2%		Test-SU	159,563	7,744	21	20.6	7598.2	98.1%
	Test-SS	161,704	4,964	194	32.6	833.5	16.8%		Test-SS	128,518	7,744	85	16.6	1512.0	19.5%
ATP W156	Full	542,154	5,110	150	106.1	3614.4	70.7%	ATP W157	Full	767,749	5,395	199	142.3	3858.0	71.5%
	Train	277,492	4,088	120	67.9	2312.4	56.6%		Train	393,844	4,316	160	91.3	2461.5	57.0%
	Test-UU	21,697	1,022	30	21.2	723.2	70.8%		Test-UU	30,444	1,079	39	28.2	780.6	72.3%
	Test-US	86,582	1,022	120	84.7	721.5	70.6%		Test-US	123,562	1,079	160	114.5	772.3	71.6%
	Test-SU	87,011	4,088	30	21.3	2900.4	70.9%		Test-SU	121,439	4,316	39	28.1	3113.8	72.1%
	Test-SS	69,372	4,088	120	17.0	578.1	14.1%		Test-SS	98,460	4,316	160	22.8	615.4	14.3%
ATP W158	Full	1,466,321	9,593	192	152.9	7637.1	79.6%	ATP W159	Full	1,447,902	9,609	206	150.7	7028.7	73.1%
	Train	746,732	7,675	154	97.3	4848.9	63.2%		Train	722,112	7,688	165	93.9	4376.4	56.9%
	Test-UU	59,761	1,918	38	31.2	1572.7	82.0%		Test-UU	63,885	1,921	41	33.3	1558.2	81.1%
	Test-US	232,644	1,918	154	121.3	1510.7	78.8%		Test-US	225,307	1,921	165	117.3	1365.5	71.1%
	Test-SU	240,502	7,675	38	31.3	6329.0	82.5%		Test-SU	256,070	7,688	41	33.3	6245.6	81.2%
	Test-SS	186,682	7,675	154	24.3	1212.2	15.8%		Test-SS	180,528	7,688	165	23.5	1094.1	14.2%
ATP W16	Full	145,792	4,684	59	31.1	2471.1	52.8%	ATP W161	Full	571,869	5,086	148	112.4	3864.0	76.0%
	Train	74,968	3,747	48	20.0	1561.8	41.7%		Train	289,935	4,069	119	71.3	2436.4	59.9%
	Test-UU	5,708	936	11	6.1	518.9	55.4%		Test-UU	23,725	1,017	29	23.3	818.1	80.4%
	Test-US	23,438	936	48	25.0	488.3	52.2%		Test-US	90,704	1,017	119	89.2	762.2	74.9%
	Test-SU	22,936	3,747	11	6.1	2085.1	55.6%		Test-SU	95,022	4,069	29	23.4	3276.6	80.5%
	Test-SS	18,742	3,727	48	5.0	390.5	10.5%		Test-SS	72,483	4,069	119	17.8	609.1	15.0%
ATP W162	Full	792,864	9,544	135	83.1	5873.1	61.5%	ATP W17	Full	216,580	4,563	70	47.5	3094.0	67.8%
	Train	396,575	7,636	108	51.9	3672.0	48.1%		Train	108,731	3,651	56	29.8	1941.6	53.2%
	Test-UU	34,599	1,908	27	18.1	1281.4	67.2%		Test-UU	9,308	912	14	10.2	664.9	72.9%
	Test-US	124,069	1,908	108	65.0	1148.8	60.2%		Test-US	34,117	912	56	37.4	609.2	66.8%
	Test-SU	138,478	7,636	27	18.1	5128.8	67.2%		Test-SU	37,242	3,651	14	10.2	2660.1	72.9%
	Test-SS	99,143	7,636	108	13.0	918.0	12.0%		Test-SS	27,182	3,651	56	7.4	485.4	13.3%
ATP W18	Full	243,598	4,601	62	52.9	3929.0	85.4%	ATP W19	Full	247,490	4,579	76	54.0	3256.4	71.1%
	Train	126,867	3,681	50	34.5	2537.3	68.9%		Train	132,145	3,664	61	36.1	2166.3	59.1%
	Test-UU	9,073	920	12	9.9	756.1	82.2%		Test-UU	8,218	915	15	9.0	547.9	59.9%
	Test-US	39,698	920	50	43.1	794.0	86.3%		Test-US	41,023	915	61	44.8	672.5	73.5%
	Test-SU	36,244	3,681	12	9.8	3020.3	82.1%		Test-SU	33,068	3,664	15	9.0	2204.5	60.2%
	Test-SS	31,716	3,679	50	8.6	634.3	17.2%		Test-SS	33,036	3,663	61	9.0	541.6	14.8%
ATP W20	Full	329,439	4,538	103	72.6	3198.4	70.5%	ATP W21	Full	110,297	4,132	39	26.7	2828.1	68.4%
	Train	168,934	3,631	83	46.5	2035.3	56.1%		Train	58,672	3,306	32	17.7	1833.5	55.5%
	Test-UU	13,180	907	20	14.5	659.0	72.7%		Test-UU	3,728	826	7	4.5	532.6	64.5%
	Test-US	52,679	907	83	58.1	634.7	70.0%		Test-US	18,324	826	32	22.2	572.6	69.3%
	Test-SU	52,413	3,631	20	14.4	2620.7	72.2%		Test-SU	14,906	3,304	7	4.5	2129.4	64.5%
	Test-SS	42,233	3,631	83	11.6	508.8	14.0%		Test-SS	14,667	3,284	32	4.5	458.3	14.0%
ATP W22	Full	220,126	4,265	68	51.6	3237.1	75.9%	ATP W23	Full	402,888	4,183	141	96.3	2857.4	68.3%
	Train	111,359	3,412	55	32.6	2024.7	59.3%		Train	209,588	3,347	113	62.6	1854.8	55.4%
	Test-UU	9,252	853	13	10.8	711.7	83.4%		Test-UU	15,042	836	28	18.0	537.2	64.3%
	Test-US	34,636	853	55	40.6	629.7	73.8%		Test-US	65,283	836	113	78.1	577.7	69.1%
	Test-SU	37,040	3,412	13	10.9	2849.2	83.5%		Test-SU	60,579	3,347	28	18.1	2163.5	64.6%
	Test-SS	27,839	3,412	55	8.2	506.2	14.8%		Test-SS	52,396	3,347	113	15.7	463.7	13.9%
ATP W24	Full	5,780	4,222	3	1.4	1926.7	45.6%	ATP W25	Full	235,430	4,151	60	56.7	3923.8	94.5%
	Train	3,692	2,898	3	1.3	1230.7	42.5%		Train	119,170	3,321	48	35.9	2482.7	74.8%
	Test-UU	0	0	0	0.0	0.0	0.0%		Test-UU	9,851	830	12	11.9	820.9	98.9%
	Test-US	1,166	844	3	1.4	388.7	46.1%		Test-US	37,175	830	48	44.8	774.5	93.3%
	Test-SU	0	0	0	0.0	0.0	0.0%		Test-SU	39,442	3,321	12	11.9	3286.8	99.0%
	Test-SS	922	863	3	1.1	307.3	35.6%		Test-SS	29,792	3,320	48	9.0	620.7	18.7%

Table 8: Full-pair benchmark split statistics (Part 5/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W26	Full	425,191	4,168	151	102.0	2815.8	67.6%	ATP W27	Full	260,218	4,135	85	62.9	3061.4	74.0%
	Train	219,396	3,335	121	65.8	1813.2	54.4%		Train	133,796	3,308	68	40.4	1967.6	59.5%
	Test-UU	16,478	833	30	19.8	549.3	65.9%		Test-UU	10,225	827	17	12.4	601.5	72.7%
	Test-US	68,565	833	121	82.3	566.7	68.0%		Test-US	41,701	827	68	50.4	613.2	74.2%
	Test-SU	65,904	3,335	30	19.8	2196.8	65.9%		Test-SU	41,047	3,308	17	12.4	2414.5	73.0%
	Test-SS	54,848	3,335	121	16.4	453.3	13.6%		Test-SS	33,449	3,307	68	10.1	491.9	14.9%
ATP W28	Full	81,745	4,971	19	16.4	4302.4	86.5%	ATP W29	Full	252,228	4,867	70	51.8	3603.3	74.0%
	Train	42,876	3,977	16	10.8	2679.8	67.4%		Train	127,391	3,894	56	32.7	2274.8	58.4%
	Test-UU	2,952	993	3	3.0	984.0	99.1%		Test-UU	10,623	973	14	10.9	758.8	78.0%
	Test-US	13,398	994	16	13.5	837.4	84.2%		Test-US	39,952	973	56	41.1	713.4	73.3%
	Test-SU	11,801	3,974	3	3.0	3933.7	99.0%		Test-SU	42,415	3,894	14	10.9	3029.6	77.8%
	Test-SS	10,718	3,796	16	2.8	669.9	17.6%		Test-SS	31,847	3,892	56	8.2	568.7	14.6%
ATP W3	Full	213,485	3,243	70	65.8	3049.8	94.0%	ATP W30	Full	203,771	4,729	59	43.1	3453.7	73.0%
	Train	107,992	2,595	56	41.6	1928.4	74.3%		Train	104,452	3,784	48	27.6	2176.1	57.5%
	Test-UU	8,944	648	14	13.8	638.9	98.6%		Test-UU	8,118	945	11	8.6	738.0	78.1%
	Test-US	33,748	648	56	52.1	602.6	93.0%		Test-US	32,520	945	48	34.4	677.5	71.7%
	Test-SU	35,803	2,595	14	13.8	2557.4	98.5%		Test-SU	32,569	3,784	11	8.6	2960.8	78.2%
	Test-SS	26,998	2,595	56	10.4	482.1	18.6%		Test-SS	26,112	3,778	48	6.9	544.0	14.4%
ATP W31	Full	428,132	4,656	134	92.0	3195.0	68.6%	ATP W32	Full	502,735	6,251	97	80.4	5182.8	82.9%
	Train	224,235	3,725	108	60.2	2076.2	55.7%		Train	261,644	5,001	78	52.3	3354.4	67.1%
	Test-UU	15,594	931	26	16.7	599.8	64.4%		Test-UU	18,855	1,250	19	15.1	992.4	79.4%
	Test-US	70,000	931	108	75.2	648.1	69.6%		Test-US	81,618	1,250	78	65.3	1046.4	83.7%
	Test-SU	62,245	3,724	26	16.7	2394.0	64.3%		Test-SU	75,207	5,001	19	15.0	3958.3	79.1%
	Test-SS	56,058	3,725	108	15.0	519.1	13.9%		Test-SS	65,411	5,001	78	13.1	838.6	16.8%
ATP W33	Full	10,119	2,537	4	4.0	2529.8	99.7%	ATP W36	Full	248,546	4,587	91	54.2	2731.3	59.5%
	Train	6,479	2,027	4	3.2	1619.8	79.9%		Train	127,144	3,670	73	34.6	1741.7	47.5%
	Test-UU	0	0	0	0.0	0.0	0.0%		Test-UU	9,879	917	18	10.8	548.8	59.9%
	Test-US	2,021	507	4	4.0	505.2	99.7%		Test-US	40,110	917	73	43.7	549.5	59.9%
	Test-SU	0	0	0	0.0	0.0	0.0%		Test-SU	39,628	3,670	18	10.8	2201.6	60.0%
	Test-SS	1,619	1,192	4	1.4	404.8	34.0%		Test-SS	31,785	3,669	73	8.7	435.4	11.9%
ATP W37	Full	169,837	4,581	54	37.1	3145.1	68.7%	ATP W38	Full	1,327,362	10,682	193	124.3	6877.5	64.4%
	Train	85,084	3,665	44	23.2	1933.7	52.8%		Train	686,870	8,546	155	80.4	4431.4	51.9%
	Test-UU	7,358	916	10	8.0	735.8	80.3%		Test-UU	50,938	2,136	38	23.8	1340.5	62.8%
	Test-US	26,733	916	44	29.2	607.6	66.3%		Test-US	214,486	2,136	155	100.4	1383.8	64.8%
	Test-SU	29,391	3,665	10	8.0	2939.1	80.2%		Test-SU	203,351	8,546	38	23.8	5351.3	62.6%
	Test-SS	21,271	3,647	44	5.8	483.4	13.3%		Test-SS	171,717	8,546	155	20.1	1107.9	13.0%
ATP W39	Full	447,041	10,640	88	42.0	5080.0	47.7%	ATP W4	Full	193,197	3,217	73	60.1	2646.5	82.3%
	Train	236,102	8,512	71	27.7	3325.4	39.1%		Train	99,399	2,574	59	38.6	1684.7	65.5%
	Test-UU	15,610	2,128	17	7.3	918.2	43.2%		Test-UU	7,582	643	14	11.8	541.6	84.2%
	Test-US	73,897	2,128	71	34.7	1040.8	48.9%		Test-US	30,971	643	59	48.2	524.9	81.6%
	Test-SU	62,407	8,511	17	7.3	3671.0	43.1%		Test-SU	30,396	2,574	14	11.8	2171.1	84.3%
	Test-SS	59,025	8,507	71	6.9	831.3	9.8%		Test-SS	24,849	2,574	59	9.7	421.2	16.4%
ATP W40	Full	1,139,662	10,618	201	107.3	5670.0	53.4%	ATP W41	Full	94,792	2,524	44	37.6	2154.4	85.4%
	Train	574,279	8,495	161	67.6	3567.0	42.0%		Train	49,438	2,020	36	24.5	1373.3	68.0%
	Test-UU	48,389	2,123	40	22.8	1209.7	57.0%		Test-UU	3,513	504	8	7.0	439.1	87.1%
	Test-US	179,263	2,123	161	84.4	1113.4	52.4%		Test-US	15,377	504	36	30.5	427.1	84.7%
	Test-SU	194,162	8,495	40	22.9	4854.1	57.1%		Test-SU	14,105	2,020	8	7.0	1763.1	87.3%
	Test-SS	143,569	8,494	161	16.9	891.7	10.5%		Test-SS	12,359	2,019	36	6.1	343.3	17.0%
ATP W42	Full	238,331	4,464	99	53.4	2407.4	53.9%	ATP W43	Full	167,926	6,636	35	25.3	4797.9	72.3%
	Train	121,956	3,572	80	34.1	1524.5	42.7%		Train	86,548	5,309	28	16.3	3091.0	58.2%
	Test-UU	9,556	892	19	10.7	502.9	56.4%		Test-UU	6,519	1,325	7	4.9	931.3	70.3%
	Test-US	37,948	892	80	42.5	474.4	53.2%		Test-US	27,009	1,327	28	20.4	964.6	72.7%
	Test-SU	38,383	3,571	19	10.7	2020.2	56.6%		Test-SU	26,213	5,303	7	4.9	3744.7	70.6%
	Test-SS	30,488	3,570	80	8.5	381.1	10.7%		Test-SS	21,637	5,243	28	4.1	772.8	14.7%
ATP W44	Full	280,979	10,970	34	25.6	8264.1	75.3%	ATP W45	Full	81,047	6,126	19	13.2	4265.6	69.6%
	Train	152,183	8,773	28	17.3	5435.1	62.0%		Train	43,932	4,901	16	9.0	2745.8	56.0%
	Test-UU	8,631	2,148	6	4.0	1438.5	67.0%		Test-UU	2,486	1,185	3	2.1	828.7	69.9%
	Test-US	47,586	2,194	28	21.7	1699.5	77.5%		Test-US	13,757	1,225	16	11.2	859.8	70.2%
	Test-SU	34,534	8,596	6	4.0	5755.7	67.0%		Test-SU	9,890	4,714	3	2.1	3296.7	69.9%
	Test-SS	38,045	8,670	28	4.4	1358.8	15.7%		Test-SS	10,982	4,427	16	2.5	686.4	15.5%

Table 9: Full-pair benchmark split statistics (Part 6/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W46	Full	472,273	6,364	106	74.2	4455.4	70.0%	ATP W47	Full	350,026	10,523	41	33.3	8537.2	81.1%
	Train	245,160	5,092	85	48.1	2884.2	56.6%		Train	176,051	8,419	33	20.9	5334.9	63.4%
	Test-UU	17,846	1,272	21	14.0	849.8	66.8%		Test-UU	15,076	2,104	8	7.2	1884.5	89.6%
	Test-US	76,805	1,272	85	60.4	903.6	71.0%		Test-US	54,911	2,104	33	26.1	1664.0	79.1%
	Test-SU	71,173	5,092	21	14.0	3389.2	66.6%		Test-SU	59,976	8,419	8	7.1	7497.0	89.0%
	Test-SS	61,289	5,091	85	12.0	721.0	14.2%		Test-SS	44,012	8,390	33	5.2	1333.7	15.9%
ATP W48	Full	448,131	10,170	114	44.1	3931.0	38.7%	ATP W49	Full	226,574	4,272	79	53.0	2868.0	67.1%
	Train	222,056	8,136	92	27.3	2413.7	29.7%		Train	115,424	3,418	64	33.8	1803.5	52.8%
	Test-UU	20,402	2,033	22	10.0	927.4	45.6%		Test-UU	9,276	854	15	10.9	618.4	72.4%
	Test-US	69,207	2,033	92	34.0	752.2	37.0%		Test-US	35,982	854	64	42.1	562.2	65.8%
	Test-SU	80,952	8,135	22	10.0	3679.6	45.2%		Test-SU	37,036	3,418	15	10.8	2469.1	72.2%
	Test-SS	55,514	8,124	92	6.8	603.4	7.4%		Test-SS	28,856	3,417	64	8.4	450.9	13.2%
ATP W5	Full	42,190	3,351	13	12.6	3245.4	96.8%	ATP W50	Full	340,226	9,834	83	34.6	4099.1	41.7%
	Train	23,548	2,681	11	8.8	2140.7	79.8%		Train	183,309	7,868	67	23.3	2736.0	34.8%
	Test-UU	1,097	670	2	1.6	548.5	81.9%		Test-UU	10,760	1,966	16	5.5	672.5	34.2%
	Test-US	7,354	670	11	11.0	668.5	99.8%		Test-US	56,965	1,966	67	29.0	850.2	43.2%
	Test-SU	4,304	2,681	2	1.6	2152.0	80.3%		Test-SU	43,365	7,868	16	5.5	2710.3	34.4%
	Test-SS	5,887	2,454	11	2.4	535.2	21.8%		Test-SS	45,827	7,836	67	5.8	684.0	8.7%
ATP W51	Full	102,606	5,105	32	20.1	3206.4	62.8%	ATP W52	Full	133,008	4,175	43	31.9	3093.2	74.1%
	Train	54,312	4,082	26	13.3	2088.9	51.2%		Train	66,762	3,340	35	20.0	1907.5	57.1%
	Test-UU	3,539	1,020	6	3.5	589.8	57.8%		Test-UU	5,753	835	8	6.9	719.1	86.1%
	Test-US	16,799	1,021	26	16.5	646.1	63.3%		Test-US	20,969	835	35	25.1	599.1	71.8%
	Test-SU	14,378	4,077	6	3.5	2396.3	58.8%		Test-SU	22,834	3,340	8	6.8	2854.2	85.5%
	Test-SS	13,578	3,957	26	3.4	522.2	13.2%		Test-SS	16,690	3,320	35	5.0	476.9	14.4%
ATP W53	Full	331,795	9,893	56	33.5	5924.9	59.9%	ATP W54	Full	155,343	6,878	35	22.6	4438.4	64.5%
	Train	175,822	7,915	45	22.2	3907.2	49.4%		Train	77,756	5,503	28	14.1	2777.0	50.5%
	Test-UU	11,523	1,975	11	5.8	1047.5	53.0%		Test-UU	6,791	1,375	7	4.9	970.1	70.6%
	Test-US	54,715	1,978	45	27.7	1215.9	61.5%		Test-US	24,234	1,375	28	17.6	865.5	62.9%
	Test-SU	45,780	7,914	11	5.8	4161.8	52.6%		Test-SU	27,124	5,502	7	4.9	3874.9	70.4%
	Test-SS	43,955	7,888	45	5.6	976.8	12.4%		Test-SS	19,438	5,371	28	3.6	694.2	12.9%
ATP W55	Full	49,940	3,627	18	13.8	2774.4	76.5%	ATP W56	Full	216,828	4,860	86	44.6	2521.3	51.9%
	Train	27,607	2,902	15	9.5	1840.5	63.4%		Train	110,513	3,888	69	28.4	1601.6	41.2%
	Test-UU	1,360	457	3	3.0	453.3	99.2%		Test-UU	8,821	972	17	9.1	518.9	53.4%
	Test-US	8,631	725	15	11.9	575.4	79.4%		Test-US	34,730	972	69	35.7	503.3	51.8%
	Test-SU	5,441	1,824	3	3.0	1813.7	99.4%		Test-SU	35,136	3,888	17	9.0	2066.8	53.2%
	Test-SS	6,901	2,659	15	2.6	460.1	17.3%		Test-SS	27,628	3,885	69	7.1	400.4	10.3%
ATP W57	Full	215,786	12,041	19	17.9	11357.2	94.3%	ATP W58	Full	136,759	3,030	50	45.1	2735.2	90.3%
	Train	116,016	9,633	16	12.0	7251.0	75.3%		Train	68,907	2,424	40	28.4	1722.7	71.1%
	Test-UU	6,912	2,401	3	2.9	2304.0	96.0%		Test-UU	5,813	606	10	9.6	581.3	95.9%
	Test-US	36,251	2,408	16	15.1	2265.7	94.1%		Test-US	21,484	606	40	35.5	537.1	88.6%
	Test-SU	27,604	9,596	3	2.9	9201.3	95.9%		Test-SU	23,329	2,421	10	9.6	2332.9	96.4%
	Test-SS	29,003	9,295	16	3.1	1812.7	19.5%		Test-SS	17,226	2,419	40	7.1	430.6	17.8%
ATP W59	Full	391,321	12,638	58	31.0	6746.9	53.4%	ATP W6	Full	85,636	3,278	27	26.1	3171.7	96.8%
	Train	194,548	10,111	47	19.2	4139.3	40.9%		Train	44,371	2,623	22	16.9	2016.9	76.9%
	Test-UU	17,456	2,526	11	6.9	1586.9	62.8%		Test-UU	3,267	655	5	5.0	653.4	99.8%
	Test-US	60,906	2,527	47	24.1	1295.9	51.3%		Test-US	13,815	655	22	21.1	628.0	95.9%
	Test-SU	69,774	10,110	11	6.9	6343.1	62.7%		Test-SU	13,091	2,622	5	5.0	2618.2	99.9%
	Test-SS	48,637	10,029	47	4.8	1034.8	10.3%		Test-SS	11,092	2,598	22	4.3	504.2	19.4%
ATP W60	Full	746,357	8,660	99	86.2	7539.0	87.1%	ATP W61	Full	207,032	6,394	46	32.4	4500.7	70.4%
	Train	392,074	6,928	80	56.6	4900.9	70.7%		Train	105,336	5,116	37	20.6	2846.9	55.6%
	Test-UU	26,750	1,732	19	15.4	1407.9	81.3%		Test-UU	8,485	1,278	9	6.6	942.8	73.8%
	Test-US	122,179	1,732	80	70.5	1527.2	88.2%		Test-US	32,679	1,278	37	25.6	883.2	69.1%
	Test-SU	107,336	6,928	19	15.5	5649.3	81.5%		Test-SU	34,199	5,110	9	6.7	3799.9	74.4%
	Test-SS	98,018	6,927	80	14.2	1225.2	17.7%		Test-SS	26,333	5,080	37	5.2	711.7	14.0%
ATP W62	Full	835,134	10,300	92	81.1	9077.5	88.1%	ATP W63	Full	198,371	3,640	74	54.5	2680.7	73.6%
	Train	434,317	8,240	74	52.7	5869.1	71.2%		Train	103,063	2,912	60	35.4	1717.7	59.0%
	Test-UU	31,365	2,060	18	15.2	1742.5	84.6%		Test-UU	7,503	728	14	10.3	535.9	73.6%
	Test-US	135,990	2,060	74	66.0	1837.7	89.2%		Test-US	32,089	728	60	44.1	534.8	73.5%
	Test-SU	124,883	8,240	18	15.2	6937.9	84.2%		Test-SU	29,951	2,912	14	10.3	2139.4	73.5%
	Test-SS	108,579	8,240	74	13.2	1467.3	17.8%		Test-SS	25,765	2,910	60	8.9	429.4	14.8%

Table 10: Full-pair benchmark split statistics (Part 7/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP	Full	458,364	11,537	40	39.7	11459.1	99.3%	ATP	Full	175,017	4,917	44	35.6	3977.7	80.9%
W64	Train	234,700	9,230	32	25.4	7334.4	79.5%	W65	Train	91,667	3,934	36	23.3	2546.3	64.7%
	Test-UU	18,332	2,307	8	7.9	2291.5	99.3%		Test-UU	6,343	983	8	6.5	792.9	80.7%
	Test-US	73,306	2,307	32	31.8	2290.8	99.3%		Test-US	28,533	983	36	29.0	792.6	80.6%
	Test-SU	73,351	9,230	8	7.9	9168.9	99.3%		Test-SU	25,558	3,934	8	6.5	3194.8	81.2%
	Test-SS	58,675	9,215	32	6.4	1833.6	19.9%		Test-SS	22,916	3,930	36	5.8	636.6	16.2%
ATP	Full	760,583	10,139	83	75.0	9163.7	90.4%	ATP	Full	558,644	10,957	53	51.0	10540.5	96.2%
W66	Train	397,408	8,112	67	49.0	5931.5	73.1%	W67	Train	287,971	8,766	43	32.9	6697.0	76.4%
	Test-UU	27,932	2,027	16	13.8	1745.8	86.1%		Test-UU	21,730	2,191	10	9.9	2173.0	99.2%
	Test-US	124,255	2,027	67	61.3	1854.6	91.5%		Test-US	90,085	2,191	43	41.1	2095.0	95.6%
	Test-SU	111,637	8,112	16	13.8	6977.3	86.0%		Test-SU	86,866	8,766	10	9.9	8686.6	99.1%
	Test-SS	99,351	8,112	67	12.2	1482.9	18.3%		Test-SS	71,992	8,759	43	8.2	1674.2	19.1%
ATP	Full	310,785	9,654	42	32.2	7399.6	76.6%	ATP	Full	776,904	4,708	184	165.0	4222.3	89.7%
W68	Train	161,552	7,724	34	20.9	4751.5	61.5%	W69	Train	392,315	3,767	148	104.1	2650.8	70.4%
	Test-UU	11,683	1,930	8	6.1	1460.4	75.7%		Test-UU	32,749	941	36	34.8	909.7	96.7%
	Test-US	50,352	1,930	34	26.1	1480.9	76.7%		Test-US	122,438	941	148	130.1	827.3	87.9%
	Test-SU	46,811	7,723	8	6.1	5851.4	75.8%		Test-SU	131,324	3,767	36	34.9	3647.9	96.8%
	Test-SS	40,387	7,697	34	5.2	1187.9	15.4%		Test-SS	98,078	3,767	148	26.0	662.7	17.6%
ATP	Full	86,070	3,154	33	27.3	2608.2	82.7%	ATP	Full	261,001	10,211	37	25.6	7054.1	69.1%
W7	Train	48,248	2,524	27	19.1	1787.0	70.8%	W70	Train	133,369	8,169	30	16.3	4445.6	54.4%
	Test-UU	2,151	625	6	3.4	358.5	57.4%		Test-UU	10,515	2,042	7	5.1	1502.1	73.6%
	Test-US	14,999	630	27	23.8	555.5	88.2%		Test-US	41,706	2,042	30	20.4	1390.2	68.1%
	Test-SU	8,610	2,516	6	3.4	1435.0	57.0%		Test-SU	42,069	8,168	7	5.2	6009.9	73.6%
	Test-SS	12,062	2,510	27	4.8	446.7	17.8%		Test-SS	33,342	8,073	30	4.1	1111.4	13.8%
ATP	Full	1,554,961	11,001	196	141.3	7933.5	72.1%	ATP	Full	58,440	13,191	8	4.4	7305.0	55.4%
W71	Train	815,313	8,801	157	92.6	5193.1	59.0%	W72	Train	29,143	9,316	7	3.1	4163.3	44.7%
	Test-UU	56,129	2,200	39	25.5	1439.2	65.4%		Test-UU	2,622	2,622	1	1.0	2622.0	100.0%
	Test-US	254,626	2,200	157	115.7	1621.8	73.7%		Test-US	8,892	2,393	7	3.7	1270.3	53.1%
	Test-SU	225,065	8,801	39	25.6	5770.9	65.6%		Test-SU	10,498	10,498	1	1.0	10498.0	100.0%
	Test-SS	203,828	8,801	157	23.2	1298.3	14.8%		Test-SS	7,285	5,204	7	1.4	1040.7	20.0%
ATP	Full	271,452	9,220	39	29.4	6960.3	75.5%	ATP	Full	265,811	10,090	44	26.3	6041.2	59.9%
W73	Train	141,909	7,376	32	19.2	4434.7	60.1%	W74	Train	138,133	8,072	36	17.1	3837.0	47.5%
	Test-UU	9,975	1,844	7	5.4	1425.0	77.3%		Test-UU	10,033	2,016	8	5.0	1254.1	62.2%
	Test-US	44,298	1,844	32	24.0	1384.3	75.1%		Test-US	42,829	2,018	36	21.2	1189.7	59.0%
	Test-SU	39,793	7,375	7	5.4	5684.7	77.1%		Test-SU	40,283	8,072	8	5.0	5035.4	62.4%
	Test-SS	35,477	7,336	32	4.8	1108.7	15.1%		Test-SS	34,533	7,893	36	4.4	959.2	12.2%
ATP	Full	827,039	11,929	115	69.3	7191.6	60.3%	ATP	Full	147,304	10,059	15	14.6	9820.3	97.6%
W75	Train	432,577	9,544	92	45.3	4701.9	49.3%	W76	Train	75,076	8,048	12	9.3	6256.3	77.7%
	Test-UU	30,334	2,385	23	12.7	1318.9	55.3%		Test-UU	5,999	2,011	3	3.0	1999.7	99.4%
	Test-US	134,445	2,385	92	56.4	1461.4	61.3%		Test-US	23,499	2,011	12	11.7	1958.2	97.4%
	Test-SU	121,539	9,544	23	12.7	5284.3	55.4%		Test-SU	23,962	8,048	3	3.0	7987.3	99.2%
	Test-SS	108,144	9,544	92	11.3	1175.5	12.3%		Test-SS	18,768	7,430	12	2.5	1564.0	21.0%
ATP	Full	175,510	10,332	40	17.0	4387.8	42.5%	ATP	Full	1,557,031	11,818	153	131.8	10176.7	86.1%
W77	Train	88,376	8,266	32	10.7	2761.8	33.4%	W78	Train	801,274	9,455	123	84.7	6514.4	68.9%
	Test-UU	7,446	2,049	8	3.6	930.8	45.4%		Test-UU	61,013	2,363	30	25.8	2033.8	86.1%
	Test-US	27,948	2,066	32	13.5	873.4	42.3%		Test-US	250,876	2,363	123	106.2	2039.6	86.3%
	Test-SU	29,646	8,212	8	3.6	3705.8	45.1%		Test-SU	243,550	9,455	30	25.8	8118.3	85.9%
	Test-SS	22,094	7,283	32	3.0	690.4	9.5%		Test-SS	200,318	9,455	123	21.2	1628.6	17.2%
ATP	Full	360,086	12,648	42	28.5	8573.5	67.8%	ATP	Full	48,674	3,181	21	15.3	2317.8	72.9%
W79	Train	174,796	10,119	34	17.3	5141.1	50.8%	W8	Train	25,104	2,545	17	9.9	1476.7	58.0%
	Test-UU	17,383	2,529	8	6.9	2172.9	85.9%		Test-UU	1,901	636	4	3.0	475.2	74.7%
	Test-US	54,658	2,528	34	21.6	1607.6	63.6%		Test-US	7,801	636	17	12.3	458.9	72.2%
	Test-SU	69,551	10,119	8	6.9	8693.9	85.9%		Test-SU	7,593	2,543	4	3.0	1898.2	74.6%
	Test-SS	43,698	10,040	34	4.4	1285.2	12.8%		Test-SS	6,275	2,347	17	2.7	369.1	15.7%
ATP	Full	166,863	5,360	43	31.1	3880.5	72.4%	ATP	Full	635,187	10,334	99	61.5	6416.0	62.1%
W80	Train	84,668	4,288	35	19.7	2419.1	56.4%	W81	Train	340,387	8,268	80	41.2	4254.8	51.5%
	Test-UU	6,917	1,072	8	6.5	864.6	80.7%		Test-UU	20,700	2,066	19	10.0	1089.5	52.7%
	Test-US	26,598	1,072	35	24.8	759.9	70.9%		Test-US	105,839	2,066	80	51.2	1323.0	64.0%
	Test-SU	27,514	4,286	8	6.4	3439.2	80.2%		Test-SU	83,165	8,268	19	10.1	4377.1	52.9%
	Test-SS	21,166	4,267	35	5.0	604.7	14.2%		Test-SS	85,096	8,268	80	10.3	1063.7	12.9%

Table 11: Full-pair benchmark split statistics (Part 8/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ATP W82	Full	188,913	2,596	84	72.8	2249.0	86.6%	ATP W83	Full	216,474	10,121	25	21.4	8659.0	85.6%
	Train	98,835	2,077	68	47.6	1453.5	70.0%		Train	112,779	8,097	20	13.9	5638.9	69.6%
	Test-UU	6,927	519	16	13.3	432.9	83.4%		Test-UU	8,068	2,024	5	4.0	1613.6	79.7%
	Test-US	30,865	519	68	59.5	453.9	87.5%		Test-US	35,174	2,024	20	17.4	1758.7	86.9%
	Test-SU	27,578	2,077	16	13.3	1723.6	83.0%		Test-SU	32,259	8,097	5	4.0	6451.8	79.7%
	Test-SS	24,708	2,077	68	11.9	363.4	17.5%		Test-SS	28,194	7,908	20	3.6	1409.7	17.8%
ATP W84	Full	1,055,909	12,055	131	87.6	8060.4	66.9%	ATP W85	Full	590,303	12,044	52	49.0	11352.0	94.3%
	Train	548,283	9,644	105	56.9	5221.7	54.1%		Train	309,952	9,636	42	32.2	7379.8	76.6%
	Test-UU	39,913	2,411	26	16.6	1535.1	63.7%		Test-UU	21,276	2,408	10	8.8	2127.6	88.4%
	Test-US	171,135	2,411	105	71.0	1629.9	67.6%		Test-US	96,801	2,408	42	40.2	2304.8	95.7%
	Test-SU	159,508	9,644	26	16.5	6134.9	63.6%		Test-SU	84,787	9,634	10	8.8	8478.7	88.0%
	Test-SS	137,070	9,644	105	14.2	1305.4	13.5%		Test-SS	77,487	9,633	42	8.0	1844.9	19.2%
ATP W86	Full	141,554	3,375	61	41.9	2320.6	68.8%	ATP W87	Full	184,301	5,109	49	36.1	3761.2	73.6%
	Train	69,743	2,700	49	25.8	1423.3	52.7%		Train	99,099	4,088	40	24.2	2477.5	60.6%
	Test-UU	6,572	675	12	9.7	547.7	81.1%		Test-UU	5,844	1,021	9	5.7	649.3	63.6%
	Test-US	21,674	675	49	32.1	442.3	65.5%		Test-US	30,998	1,021	40	30.4	775.0	75.9%
	Test-SU	26,130	2,699	12	9.7	2177.5	80.7%		Test-SU	23,586	4,088	9	5.8	2620.7	64.1%
	Test-SS	17,435	2,700	49	6.5	355.8	13.2%		Test-SS	24,774	4,086	40	6.1	619.4	15.2%
ATP W88	Full	140,756	4,623	56	30.4	2513.5	54.4%	ATP W89	Full	998,026	13,749	87	72.6	11471.6	83.4%
	Train	72,677	3,699	45	19.6	1615.0	43.7%		Train	523,576	11,000	70	47.6	7479.7	68.0%
	Test-UU	5,411	924	11	5.9	491.9	53.2%		Test-UU	35,863	2,749	17	13.0	2109.6	76.7%
	Test-US	23,148	924	45	25.1	514.4	55.7%		Test-US	162,834	2,749	70	59.2	2326.2	84.6%
	Test-SU	21,351	3,698	11	5.8	1941.0	52.5%		Test-SU	144,859	11,000	17	13.2	8521.1	77.5%
	Test-SS	18,169	3,668	45	5.0	403.8	11.0%		Test-SS	130,894	10,999	70	11.9	1869.9	17.0%
ATP W9	Full	404,525	3,212	157	125.9	2576.6	80.2%	ATP W90	Full	138,456	2,548	57	54.3	2429.1	95.3%
	Train	207,429	2,570	126	80.7	1646.3	64.1%		Train	72,736	2,039	46	35.7	1581.2	77.5%
	Test-UU	16,069	642	31	25.0	518.4	80.7%		Test-UU	4,973	509	11	9.8	452.1	88.8%
	Test-US	65,073	642	126	101.4	516.5	80.4%		Test-US	22,581	509	46	44.4	490.9	96.4%
	Test-SU	64,097	2,570	31	24.9	2067.6	80.5%		Test-SU	19,982	2,038	11	9.8	1816.5	89.1%
	Test-SS	51,857	2,570	126	20.2	411.6	16.0%		Test-SS	18,184	2,039	46	8.9	395.3	19.4%
ATP W91	Full	836,483	10,606	86	78.9	9726.5	91.7%	ATP W92	Full	391,945	10,221	43	38.3	9115.0	89.2%
	Train	438,472	8,485	69	51.7	6354.7	74.9%		Train	205,653	8,177	35	25.2	5875.8	71.9%
	Test-UU	30,229	2,121	17	14.3	1778.2	83.8%		Test-UU	14,073	2,044	8	6.9	1759.1	86.1%
	Test-US	137,049	2,121	69	64.6	1986.2	93.6%		Test-US	64,384	2,044	35	31.5	1839.5	90.0%
	Test-SU	121,116	8,485	17	14.3	7124.5	84.0%		Test-SU	56,422	8,176	8	6.9	7052.8	86.3%
	Test-SS	109,617	8,485	69	12.9	1588.7	18.7%		Test-SS	51,413	8,167	35	6.3	1468.9	18.0%
ATP W93	Full	23,962	11,049	3	2.2	7987.3	72.3%	ATP W94	Full	370,354	10,348	59	35.8	6277.2	60.7%
	Train	15,351	8,328	3	1.8	5117.0	61.4%		Train	177,798	8,279	48	21.5	3704.1	44.7%
	Test-UU	0	0	0	0.0	0.0	0.0%		Test-UU	18,532	2,069	11	9.0	1684.7	81.4%
	Test-US	4,774	2,209	3	2.2	1591.3	72.0%		Test-US	55,190	2,069	48	26.7	1149.8	55.6%
	Test-SU	0	0	0	0.0	0.0	0.0%		Test-SU	74,385	8,279	11	9.0	6762.3	81.7%
	Test-SS	3,837	3,322	3	1.2	1279.0	38.5%		Test-SS	44,449	8,249	48	5.4	926.0	11.2%
ATP W95	Full	663,306	10,371	77	64.0	8614.4	83.1%	ATP W96	Full	783,367	6,485	159	120.8	4926.8	76.0%
	Train	342,056	8,297	62	41.2	5517.0	66.5%		Train	402,374	5,188	128	77.6	3143.5	60.6%
	Test-UU	25,731	2,074	15	12.4	1715.4	82.7%		Test-UU	30,889	1,297	31	23.8	996.4	76.8%
	Test-US	107,057	2,074	62	51.6	1726.7	83.3%		Test-US	125,644	1,297	128	96.9	981.6	75.7%
	Test-SU	102,949	8,297	15	12.4	6863.3	82.7%		Test-SU	123,867	5,188	31	23.9	3995.7	77.0%
	Test-SS	85,513	8,297	62	10.3	1379.2	16.6%		Test-SS	100,593	5,188	128	19.4	785.9	15.1%
ATP W97	Full	81,459	6,510	16	12.5	5091.2	78.2%	ATP W98	Full	221,961	9,676	42	22.9	5284.8	54.6%
	Train	41,492	5,207	13	8.0	3191.7	61.3%		Train	105,730	7,740	34	13.7	3109.7	40.2%
	Test-UU	3,319	1,302	3	2.5	1106.3	85.0%		Test-UU	11,376	1,935	8	5.9	1422.0	73.5%
	Test-US	12,974	1,302	13	10.0	998.0	76.7%		Test-US	32,752	1,935	34	16.9	963.3	49.8%
	Test-SU	13,301	5,203	3	2.6	4433.7	85.2%		Test-SU	45,671	7,740	8	5.9	5708.9	73.8%
	Test-SS	10,373	4,482	13	2.3	797.9	17.8%		Test-SS	26,432	7,478	34	3.5	777.4	10.4%
ATP W99	Full	554,962	10,260	101	54.1	5494.7	53.6%	AFROBAROMETER 2023	Full	10,328,498	53,444	212	193.3	48719.3	91.2%
	Train	285,802	8,208	81	34.8	3528.4	43.0%		Train	5,267,026	42,756	170	123.2	30982.5	72.5%
	Test-UU	21,708	2,051	20	10.6	1085.4	52.9%		Test-UU	420,064	10,688	42	39.3	10001.5	93.6%
	Test-US	89,099	2,052	81	43.4	1100.0	53.6%		Test-US	1,643,455	10,688	170	153.8	9667.4	90.5%
	Test-SU	86,903	8,206	20	10.6	4345.1	53.0%		Test-SU	1,681,197	42,756	42	39.3	40028.5	93.6%
	Test-SS	71,450	8,203	81	8.7	882.1	10.8%		Test-SS	1,316,756	42,756	170	30.8	7745.6	18.1%

Table 12: Full-pair benchmark split statistics (Part 9/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
AFROBAROMETER 2020	Full	2,417,971	45,823	56	52.8	43178.1	94.2%	AFROBAROMETER 2022	Full	1,140,900	48,084	32	23.7	35653.1	74.1%
	Train	1,243,636	36,659	45	33.9	27636.4	75.4%		Train	593,536	38,468	26	15.4	22828.3	59.3%
	Test-UU	95,029	9,164	11	10.4	8639.0	94.3%		Test-UU	42,657	9,577	6	4.5	7109.5	74.2%
	Test-US	388,885	9,164	45	42.4	8641.9	94.3%		Test-US	185,314	9,616	26	19.3	7127.5	74.1%
	Test-SU	379,512	36,658	11	10.4	34501.1	94.1%		Test-SU	171,010	38,277	6	4.5	28501.7	74.5%
	Test-SS	310,909	36,650	45	8.5	6909.1	18.9%		Test-SS	148,383	37,821	26	3.9	5707.0	15.1%
ARAB BAROMETER 2023	Full	2,436,306	15,627	292	155.9	8343.5	53.4%	ARAB BAROMETER 2016	Full	1,398,121	9,518	212	146.9	6594.9	69.3%
	Train	1,237,563	12,502	234	99.0	5288.7	42.3%		Train	721,220	7,615	170	94.7	4242.5	55.7%
	Test-UU	97,114	3,125	58	31.1	1674.4	53.6%		Test-UU	54,312	1,903	42	28.5	1293.1	68.0%
	Test-US	386,056	3,125	234	123.5	1649.8	52.8%		Test-US	225,066	1,903	170	118.3	1323.9	69.6%
	Test-SU	388,485	12,502	58	31.1	6698.0	53.6%		Test-SU	217,219	7,615	42	28.5	5171.9	67.9%
	Test-SS	309,390	12,502	234	24.7	1322.2	10.6%		Test-SS	180,304	7,615	170	23.7	1060.6	13.9%
ARAB BAROMETER 2018	Full	3,236,577	27,850	228	116.2	14195.5	51.0%	ARAB BAROMETER 2020	Full	220,302	5,729	54	38.5	4079.7	71.2%
	Train	1,639,296	22,280	183	73.6	8957.9	40.2%		Train	114,221	4,584	44	24.9	2595.9	56.6%
	Test-UU	134,960	5,570	45	24.2	2999.1	53.8%		Test-UU	8,370	1,145	10	7.3	837.0	73.1%
	Test-US	512,377	5,570	183	92.0	2799.9	50.3%		Test-US	35,641	1,145	44	31.1	810.0	70.7%
	Test-SU	540,121	22,280	45	24.2	12002.7	53.9%		Test-SU	33,515	4,584	10	7.3	3351.5	73.1%
	Test-SS	409,823	22,280	183	18.4	2239.5	10.1%		Test-SS	28,555	4,575	44	6.2	649.0	14.2%
ARAB BAROMETER 2021	Full	3,835,860	26,154	295	146.7	13002.9	49.7%	ASIAN BAROMETER Wave5	Full	1,306,986	26,951	60	48.5	21783.1	80.8%
	Train	2,002,302	20,924	236	95.7	8484.3	40.5%		Train	674,890	21,561	48	31.3	14060.2	65.2%
	Test-UU	141,355	5,230	59	27.0	2395.8	45.8%		Test-UU	50,495	5,390	12	9.4	4207.9	78.1%
	Test-US	627,769	5,230	236	120.0	2660.0	50.9%		Test-US	210,955	5,390	48	39.1	4394.9	81.5%
	Test-SU	563,859	20,924	59	26.9	9556.9	45.7%		Test-SU	201,924	21,561	12	9.4	16827.0	78.0%
	Test-SS	500,575	20,924	236	23.9	2121.1	10.1%		Test-SS	168,722	21,540	48	7.8	3515.0	16.3%
ASIAN BAROMETER Wave1	Full	1,228,626	12,217	148	100.6	8301.5	68.0%	ASIAN BAROMETER Wave2	Full	2,614,606	19,798	174	132.1	15026.5	75.9%
	Train	632,051	9,774	119	64.7	5311.4	54.3%		Train	1,349,410	15,839	140	85.2	9638.6	60.9%
	Test-UU	48,137	2,443	29	19.7	1659.9	67.9%		Test-UU	101,258	3,959	34	25.6	2978.2	75.2%
	Test-US	196,948	2,443	119	80.6	1655.0	67.7%		Test-US	421,100	3,959	140	106.4	3007.9	76.0%
	Test-SU	193,478	9,774	29	19.8	6671.7	68.3%		Test-SU	405,486	15,839	34	25.6	11926.1	75.3%
	Test-SS	158,012	9,772	119	16.2	1327.8	13.6%		Test-SS	337,352	15,839	140	21.3	2409.7	15.2%
ASIAN BAROMETER Wave3	Full	3,048,432	19,436	189	156.8	16129.3	83.0%	ASIAN BAROMETER Wave4	Full	3,623,339	20,667	223	175.3	16248.2	78.6%
	Train	1,572,064	15,549	152	101.1	10342.5	66.5%		Train	1,884,580	16,534	179	114.0	10528.4	63.7%
	Test-UU	118,324	3,887	37	30.4	3197.9	82.3%		Test-UU	135,554	4,133	44	32.8	3080.8	74.5%
	Test-US	491,477	3,887	152	126.4	3233.4	83.2%		Test-US	586,698	4,133	179	142.0	3277.6	79.3%
	Test-SU	473,551	15,549	37	30.5	12798.7	82.3%		Test-SU	545,362	16,534	44	33.0	12394.6	75.0%
	Test-SS	393,016	15,549	152	25.3	2585.6	16.6%		Test-SS	471,145	16,534	179	28.5	2632.1	15.9%
ASIAN BAROMETER Wave6	Full	38,506	1,003	42	38.4	916.8	91.4%	CGSS 2023	Full	452,206	11,326	68	39.9	6650.1	58.7%
	Train	20,364	803	34	25.4	598.9	74.6%		Train	235,563	9,061	55	26.0	4283.0	47.3%
	Test-UU	1,340	200	8	6.7	167.5	83.8%		Test-UU	16,799	2,265	13	7.4	1292.2	57.1%
	Test-US	6,348	200	34	31.7	186.7	93.4%		Test-US	74,224	2,265	55	32.8	1349.5	59.6%
	Test-SU	5,364	803	8	6.7	670.5	83.5%		Test-SU	66,730	9,061	13	7.4	5133.1	56.7%
	Test-SS	5,090	801	34	6.4	149.7	18.7%		Test-SS	58,890	8,967	55	6.6	1070.7	11.9%
CGSS 2003	Full	136,004	5,894	26	23.1	5230.9	88.7%	CGSS 2006	Full	212,070	3,207	94	66.1	2256.1	70.3%
	Train	71,619	4,716	21	15.2	3410.4	72.3%		Train	109,035	2,566	76	42.5	1434.7	55.9%
	Test-UU	4,831	1,178	5	4.1	966.2	82.0%		Test-UU	8,348	641	18	13.0	463.8	72.4%
	Test-US	22,505	1,178	21	19.1	1071.7	91.0%		Test-US	34,130	641	76	53.2	449.1	70.1%
	Test-SU	19,145	4,716	5	4.1	3829.0	81.2%		Test-SU	33,299	2,566	18	13.0	1849.9	72.1%
	Test-SS	17,904	4,629	21	3.9	852.6	18.4%		Test-SS	27,258	2,566	76	10.6	358.7	14.0%
CGSS 2010	Full	1,428,403	11,783	160	121.2	8927.5	75.8%	CGSS 2011	Full	515,806	5,620	111	91.8	4646.9	82.7%
	Train	729,062	9,426	128	77.3	5695.8	60.4%		Train	260,917	4,496	89	58.0	2931.7	65.2%
	Test-UU	57,773	2,356	32	24.5	1805.4	76.6%		Test-UU	21,601	1,124	22	19.2	981.9	87.4%
	Test-US	227,429	2,356	127	96.5	1790.8	76.0%		Test-US	81,430	1,124	89	72.4	914.9	81.4%
	Test-SU	231,874	9,425	32	24.6	7246.1	76.9%		Test-SU	86,629	4,496	22	19.3	3937.7	87.6%
	Test-SS	182,265	9,426	127	19.3	1435.2	15.2%		Test-SS	65,229	4,495	89	14.5	732.9	16.3%
CGSS 2012	Full	778,855	11,765	103	66.2	7561.7	64.3%	CGSS 2013	Full	954,094	11,438	107	83.4	8916.8	78.0%
	Train	392,662	9,412	83	41.7	4730.9	50.3%		Train	506,664	9,151	86	55.4	5891.4	64.4%
	Test-UU	33,058	2,353	20	14.0	1652.9	70.2%		Test-UU	32,638	2,287	21	14.3	1554.2	68.0%
	Test-US	122,914	2,353	83	52.2	1480.9	62.9%		Test-US	158,367	2,287	86	69.2	1841.5	80.5%
	Test-SU	132,056	9,412	20	14.0	6602.8	70.2%		Test-SU	129,759	9,151	21	14.2	6179.0	67.5%
	Test-SS	98,165	9,412	83	10.4	1182.7	12.6%		Test-SS	126,666	9,151	86	13.8	1472.9	16.1%

Table 13: Full-pair benchmark split statistics (Part 10/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
CGSS	Full	1,334,197	10,968	359	121.6	3716.4	33.9%	CGSS	Full	1,879,502	12,582	466	149.4	4033.3	32.1%
2015	Train	665,485	8,775	288	75.8	2310.7	26.3%	2017	Train	957,965	10,066	373	95.2	2568.3	25.5%
	Test-UU	58,710	2,193	71	26.8	826.9	37.7%		Test-UU	76,500	2,516	93	30.4	822.6	32.7%
	Test-US	205,605	2,193	288	93.8	713.9	32.6%		Test-US	297,066	2,516	373	118.1	796.4	31.7%
	Test-SU	238,026	8,775	71	27.1	3352.5	38.2%		Test-SU	308,480	10,066	93	30.6	3317.0	33.0%
	Test-SS	166,371	8,775	288	19.0	577.7	6.6%		Test-SS	239,491	10,066	373	23.8	642.1	6.4%
CGSS	Full	1,462,383	12,787	268	114.4	5456.7	42.7%	CGSS	Full	1,450,690	8,148	389	178.0	3729.3	45.8%
2018	Train	753,816	10,230	212	73.7	3555.7	34.8%	2021	Train	758,002	6,519	312	116.3	2429.5	37.3%
	Test-UU	56,885	2,557	53	22.2	1073.3	42.0%		Test-UU	53,428	1,629	77	32.8	693.9	42.6%
	Test-US	236,028	2,557	210	92.3	1123.9	44.0%		Test-US	235,373	1,629	308	144.5	764.2	46.9%
	Test-SU	227,200	10,230	53	22.2	4286.8	41.9%		Test-SU	214,387	6,519	77	32.9	2784.2	42.7%
	Test-SS	188,454	10,230	210	18.4	897.4	8.8%		Test-SS	189,500	6,519	312	29.1	607.4	9.3%
CSES	Full	1,932,820	33,871	79	57.1	24466.1	72.2%	CSES	Full	4,144,850	62,409	99	66.4	41867.2	67.1%
V6	Train	1,010,226	27,097	64	37.3	15784.8	58.3%	V1	Train	2,125,704	49,928	80	42.6	26571.3	53.2%
	Test-UU	70,769	6,774	15	10.4	4717.9	69.6%		Test-UU	164,721	12,481	19	13.2	8669.5	69.5%
	Test-US	315,423	6,774	64	46.6	4928.5	72.8%		Test-US	664,502	12,481	80	53.2	8306.3	66.6%
	Test-SU	283,846	27,097	15	10.5	18923.1	69.8%		Test-SU	658,497	49,928	19	13.2	34657.7	69.4%
	Test-SS	252,556	27,095	64	9.3	3946.2	14.6%		Test-SS	531,426	49,924	80	10.6	6642.8	13.3%
CSES	Full	5,094,333	64,256	124	79.3	41083.3	63.9%	CSES	Full	5,254,493	80,163	116	65.5	45297.4	56.5%
V2	Train	2,677,812	51,405	100	52.1	26778.1	52.1%	V3	Train	2,737,064	64,131	93	42.7	29430.8	45.9%
	Test-UU	181,937	12,851	24	14.2	7580.7	59.0%		Test-UU	195,807	16,032	23	12.2	8513.3	53.1%
	Test-US	839,104	12,851	100	65.3	8391.0	65.3%		Test-US	854,710	16,032	93	53.3	9190.4	57.3%
	Test-SU	726,027	51,405	24	14.1	30251.1	58.8%		Test-SU	782,647	64,131	23	12.2	34028.1	53.1%
	Test-SS	669,453	51,405	100	13.0	6694.5	13.0%		Test-SS	684,265	64,129	93	10.7	7357.7	11.5%
CSES	Full	5,342,611	75,558	111	70.7	48131.6	63.7%	CSES	Full	10,624,377	114,714	133	92.6	79882.5	69.6%
V4	Train	2,595,096	60,447	89	42.9	29158.4	48.2%	V5	Train	5,528,437	91,772	107	60.2	51667.6	56.3%
	Test-UU	257,562	15,111	22	17.0	11707.4	77.5%		Test-UU	396,957	22,942	26	17.3	15267.6	66.5%
	Test-US	810,809	15,111	89	53.7	9110.2	60.3%		Test-US	1,727,514	22,942	107	75.3	16145.0	70.4%
	Test-SU	1,030,370	60,447	22	17.0	46835.0	77.5%		Test-SU	1,589,360	91,772	26	17.3	61129.2	66.6%
	Test-SS	648,774	60,436	89	10.7	7289.6	12.1%		Test-SS	1,382,109	91,772	107	15.1	12916.9	14.1%
CSS	Full	1,261,459	10,136	188	124.5	6709.9	66.2%	CSS	Full	613,757	7,061	136	86.9	4512.9	63.9%
2021	Train	646,793	8,109	151	79.8	4283.4	52.8%	CSS2006	Train	311,664	5,649	109	55.2	2859.3	50.6%
	Test-UU	50,242	2,027	37	24.8	1357.9	67.0%		Test-UU	25,454	1,412	27	18.0	942.7	66.8%
	Test-US	200,807	2,027	151	99.1	1329.8	65.6%		Test-US	97,037	1,412	108	68.7	898.5	63.6%
	Test-SU	201,919	8,109	37	24.9	5457.3	67.3%		Test-SU	101,687	5,649	27	18.0	3766.2	66.7%
	Test-SS	161,698	8,109	151	19.9	1070.8	13.2%		Test-SS	77,915	5,649	108	13.8	721.4	12.8%
CSS	Full	1,287,348	7,036	416	183.0	3094.6	44.0%	CSS	Full	2,757,066	10,206	526	270.1	5241.6	51.4%
CSS2011	Train	657,963	5,629	332	116.9	1981.8	35.2%	CSS2013	Train	1,406,452	8,165	421	172.3	3340.7	40.9%
	Test-UU	51,826	1,407	83	36.8	624.4	44.4%		Test-UU	111,869	2,041	105	54.8	1065.4	52.2%
	Test-US	204,943	1,407	333	145.7	615.4	43.7%		Test-US	438,054	2,041	421	214.6	1040.5	51.0%
	Test-SU	208,126	5,629	83	37.0	2507.5	44.5%		Test-SU	449,078	8,165	105	55.0	4276.9	52.4%
	Test-SS	164,490	5,629	320	29.2	514.0	9.1%		Test-SS	351,613	8,165	409	43.1	859.7	10.5%
CSS	Full	2,201,157	10,243	434	214.9	5071.8	49.5%	CSS	Full	1,569,867	10,143	485	154.8	3236.8	31.9%
CSS2015	Train	1,105,900	8,195	339	134.9	3262.2	39.8%	CSS2017	Train	803,636	8,115	372	99.0	2160.3	26.6%
	Test-UU	94,586	2,048	82	46.2	1153.5	56.3%		Test-UU	62,769	2,028	81	31.0	774.9	38.2%
	Test-US	344,313	2,048	319	168.1	1079.4	52.7%		Test-US	248,772	2,028	308	122.7	807.7	39.8%
	Test-SU	379,883	8,195	85	46.4	4469.2	54.5%		Test-SU	253,782	8,115	94	31.3	2699.8	33.3%
	Test-SS	276,475	8,195	313	33.7	883.3	10.8%		Test-SS	200,908	8,115	318	24.8	631.8	7.8%
CSS	Full	1,002,106	10,283	317	97.5	3161.2	30.7%	CSS	Full	2,202,935	10,136	497	217.3	4432.5	43.7%
CSS2019	Train	496,064	8,227	246	60.3	2016.5	24.5%	CSS2021	Train	1,168,762	8,109	391	144.1	2989.2	36.9%
	Test-UU	45,530	2,056	50	22.1	910.6	44.3%		Test-UU	75,584	2,027	85	37.3	889.2	43.9%
	Test-US	155,130	2,056	191	75.5	812.2	39.5%		Test-US	364,015	2,027	340	179.6	1070.6	52.8%
	Test-SU	181,367	8,227	63	22.0	2878.8	35.0%		Test-SU	302,384	8,109	99	37.3	3054.4	37.7%
	Test-SS	124,015	8,227	222	15.1	558.6	6.8%		Test-SS	292,190	8,109	347	36.0	842.0	10.4%
ESS	Full	6,224,491	50,116	145	124.2	42927.5	85.7%	ESS	Full	5,727,611	37,611	260	152.3	22029.3	58.6%
11.04	Train	3,194,622	40,093	116	79.7	27539.8	68.7%	10.03	Train	2,868,978	30,089	208	95.3	13793.2	45.8%
	Test-UU	246,563	10,023	29	24.6	8502.2	84.8%		Test-UU	249,029	7,522	52	33.1	4789.0	63.7%
	Test-US	997,937	10,023	116	99.6	8602.9	85.8%		Test-US	895,583	7,522	206	119.1	4347.5	57.8%
	Test-SU	986,714	40,093	29	24.6	34024.6	84.9%		Test-SU	996,777	30,089	52	33.1	19168.8	63.7%
	Test-SS	798,655	40,092	116	19.9	6885.0	17.2%		Test-SS	717,244	30,089	206	23.8	3481.8	11.6%

Table 14: Full-pair benchmark split statistics (Part 11/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
ESS 1.06	Full	7,008,865	42,359	299	165.5	23441.0	55.3%	ESS 2.03	Full	9,964,734	47,537	384	209.6	25949.8	54.6%
	Train	3,689,551	33,888	237	108.9	15567.7	45.9%		Train	5,065,419	38,030	307	133.2	16499.7	43.4%
	Test-UU	248,315	8,471	59	29.3	4208.7	49.7%		Test-UU	409,660	9,507	76	43.1	5390.3	56.7%
	Test-US	1,153,720	8,471	240	136.2	4807.2	56.7%		Test-US	1,584,113	9,507	308	166.6	5143.2	54.1%
	Test-SU	994,892	33,888	56	29.4	17765.9	52.4%		Test-SU	1,639,188	38,030	76	43.1	21568.3	56.7%
	Test-SS	922,387	33,888	237	27.2	3891.9	11.5%		Test-SS	1,266,354	38,030	302	33.3	4193.2	11.0%
ESS 3.03	Full	7,842,017	43,000	300	182.4	26140.1	60.8%	ESS 4.04	Full	9,141,672	56,752	381	161.1	23993.9	42.3%
	Train	3,960,867	34,400	239	115.1	16572.7	48.2%		Train	4,646,604	45,402	305	102.3	15234.8	33.6%
	Test-UU	331,002	8,600	60	38.5	5516.7	64.1%		Test-UU	376,513	11,350	75	33.2	5020.2	44.2%
	Test-US	1,238,647	8,600	234	144.0	5293.4	61.6%		Test-US	1,449,777	11,350	304	127.7	4769.0	42.0%
	Test-SU	1,321,285	34,400	60	38.4	22021.4	64.0%		Test-SU	1,507,128	45,402	76	33.2	19830.6	43.7%
	Test-SS	990,216	34,400	239	28.8	4143.2	12.0%		Test-SS	1,161,650	45,402	304	25.6	3821.2	8.4%
ESS 5.03	Full	8,475,231	52,458	316	161.6	26820.4	51.1%	ESS 6.02	Full	8,419,577	54,673	291	154.0	28933.3	52.9%
	Train	4,200,656	41,967	253	100.1	16603.4	39.6%		Train	4,118,709	43,739	232	94.2	17753.1	40.6%
	Test-UU	382,374	10,491	63	36.4	6069.4	57.9%		Test-UU	397,023	10,934	58	36.3	6845.2	62.6%
	Test-US	1,310,902	10,491	249	125.0	5264.7	50.2%		Test-US	1,287,777	10,934	229	117.8	5623.5	51.4%
	Test-SU	1,531,136	41,967	63	36.5	24303.7	57.9%		Test-SU	1,586,391	43,739	58	36.3	27351.6	62.5%
	Test-SS	1,050,163	41,967	249	25.0	4217.5	10.0%		Test-SS	1,029,677	43,738	228	23.5	4516.1	10.3%
ESS 7.02	Full	5,095,059	40,185	223	126.8	22847.8	56.9%	ESS 8.02	Full	6,002,996	44,387	230	135.2	26100.0	58.8%
	Train	2,631,864	32,148	178	81.9	14785.8	46.0%		Train	3,050,217	35,510	184	85.9	16577.3	46.7%
	Test-UU	196,695	8,037	44	24.5	4470.3	55.6%		Test-UU	247,246	8,877	46	27.9	5374.9	60.5%
	Test-US	823,251	8,037	176	102.4	4677.6	58.2%		Test-US	951,890	8,877	184	107.2	5173.3	58.3%
	Test-SU	785,283	32,148	44	24.4	17847.3	55.5%		Test-SU	991,089	35,509	46	27.9	21545.4	60.7%
	Test-SS	657,966	32,148	176	20.5	3738.4	11.6%		Test-SS	762,554	35,510	184	21.5	4144.3	11.7%
ESS 9.03	Full	5,949,978	49,519	224	120.2	26562.4	53.6%	EVS wave5	Full	3,979,442	59,437	76	67.0	52361.1	88.1%
	Train	3,003,408	39,616	178	75.8	16873.1	42.6%		Train	2,020,170	47,550	61	42.5	33117.5	69.6%
	Test-UU	251,305	9,903	44	25.4	5711.5	57.7%		Test-UU	164,462	11,887	15	13.8	10964.1	92.2%
	Test-US	937,311	9,903	176	94.6	5325.6	53.8%		Test-US	631,116	11,887	61	53.1	10346.2	87.0%
	Test-SU	1,007,102	39,616	44	25.4	22888.7	57.8%		Test-SU	658,652	47,550	15	13.9	43910.1	92.3%
	Test-SS	750,852	39,616	176	19.0	4266.2	10.8%		Test-SS	505,042	47,546	61	10.6	8279.4	17.4%
GSS 2024	Full	750,747	3,986	375	188.3	2002.0	50.2%	GSS 2021	Full	942,705	4,032	416	233.8	2266.1	56.2%
	Train	379,292	3,189	300	118.9	1264.3	39.6%		Train	490,268	3,226	333	152.0	1472.3	45.6%
	Test-UU	31,793	797	75	39.9	423.9	53.2%		Test-UU	35,352	806	83	43.9	425.9	52.8%
	Test-US	117,829	797	300	147.8	392.8	49.3%		Test-US	153,215	806	332	190.1	461.5	57.3%
	Test-SU	127,011	3,189	75	39.8	1693.5	53.1%		Test-SU	141,304	3,226	83	43.8	1702.5	52.8%
	Test-SS	94,822	3,189	300	29.7	316.1	9.9%		Test-SS	122,566	3,226	330	38.0	371.4	11.5%
GSS 2022	Full	1,308,658	4,149	635	315.4	2060.9	49.7%	ISSP Social_Inequality	Full	6,430,184	157,435	71	40.8	90566.0	57.5%
	Train	664,722	3,320	508	200.2	1308.5	39.4%		Train	3,485,644	125,946	57	27.7	61151.6	48.6%
	Test-UU	53,983	829	127	65.1	425.1	51.3%		Test-UU	196,704	31,057	14	6.3	14050.3	45.2%
	Test-US	206,309	829	504	248.9	409.3	49.4%		Test-US	1,088,256	31,487	57	34.6	19092.2	60.6%
	Test-SU	217,464	3,320	127	65.5	1712.3	51.6%		Test-SU	788,170	124,292	14	6.3	56297.9	45.3%
	Test-SS	166,180	3,320	503	50.1	330.4	10.0%		Test-SS	871,410	125,499	57	6.9	15287.9	12.2%
ISSP Environment	Full	6,357,594	142,641	85	44.6	74795.2	52.4%	ISSP Government	Full	6,667,000	136,581	87	48.8	76632.2	56.1%
	Train	3,359,501	114,113	68	29.4	49404.4	43.3%		Train	3,477,510	109,265	70	31.8	49678.7	45.5%
	Test-UU	221,785	28,503	17	7.8	13046.2	45.8%		Test-UU	247,094	27,298	17	9.1	14534.9	53.2%
	Test-US	1,048,918	28,528	68	36.8	15425.3	54.1%		Test-US	1,087,256	27,316	70	39.8	15532.2	56.9%
	Test-SU	887,515	114,045	17	7.8	52206.8	45.8%		Test-SU	985,763	109,213	17	9.0	57986.1	53.1%
	Test-SS	839,875	114,010	68	7.4	12351.1	10.8%		Test-SS	869,377	109,145	70	8.0	12419.7	11.4%
ISSP Health	Full	3,214,268	73,450	61	43.8	52692.9	71.7%	ISSP Religion	Full	8,303,771	163,361	101	50.8	82215.6	50.3%
	Train	1,660,608	58,759	49	28.3	33890.0	57.7%		Train	4,135,539	130,686	81	31.6	51056.0	39.1%
	Test-UU	123,941	14,690	12	8.4	10328.4	70.3%		Test-UU	368,528	32,670	20	11.3	18426.4	56.4%
	Test-US	519,050	14,690	49	35.3	10592.9	72.1%		Test-US	1,293,522	32,672	81	39.6	15969.4	48.9%
	Test-SU	495,517	58,759	12	8.4	41293.1	70.3%		Test-SU	1,472,298	130,686	20	11.3	73614.9	56.3%
	Test-SS	415,152	58,698	49	7.1	8472.5	14.4%		Test-SS	1,033,884	130,602	81	7.9	12764.0	9.8%
ISSP Work	Full	5,189,421	121,945	78	42.6	66531.0	54.6%	AMERICAS BAR. 2023	Full	2,517,075	31,132	133	80.9	18925.4	60.8%
	Train	2,699,591	97,555	63	27.7	42850.7	43.9%		Train	1,259,240	24,905	107	50.6	11768.6	47.3%
	Test-UU	193,780	24,385	15	7.9	12918.7	53.0%		Test-UU	109,983	6,226	26	17.7	4230.1	67.9%
	Test-US	843,595	24,389	63	34.6	13390.4	54.9%		Test-US	393,303	6,226	107	63.2	3675.7	59.0%
	Test-SU	777,558	97,544	15	8.0	51837.2	53.1%		Test-SU	439,740	24,905	26	17.7	16913.1	67.9%
	Test-SS	674,897	97,051	63	7.0	10712.7	11.0%		Test-SS	314,809	24,897	107	12.6	2942.1	11.8%

Table 15: Full-pair benchmark split statistics (Part 12/12, continued from Table 4)

Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density	Dataset	Split	Pairs	Users	Qs	Q/U	U/Q	Density
AMERICAS BAR. 2018	Full	621,944	15,522	53	40.1	11734.8	75.6%	AMERICAS BAR. 2021	Full	781,265	39,498	35	19.8	22321.9	56.5%
	Train	322,659	12,418	43	26.0	7503.7	60.4%		Train	423,938	31,598	28	13.4	15140.6	47.9%
	Test-UU	23,632	3,104	10	7.6	2363.2	76.1%		Test-UU	23,746	7,884	7	3.0	3392.3	43.0%
	Test-US	100,275	3,104	43	32.3	2332.0	75.1%		Test-US	132,179	7,899	28	16.7	4720.7	59.8%
	Test-SU	94,714	12,418	10	7.6	9471.4	76.3%		Test-SU	95,418	31,522	7	3.0	13631.1	43.2%
	Test-SS	80,664	12,407	43	6.5	1875.9	15.1%		Test-SS	105,984	30,600	28	3.5	3785.1	12.4%
WVS W7	Full	22,179,427	97,220	251	228.1	88364.3	90.9%	WVS W1	Full	1,311,275	13,586	111	96.5	11813.3	87.0%
	Train	11,392,340	77,776	201	146.5	56678.3	72.9%		Train	675,278	10,869	89	62.1	7587.4	69.8%
	Test-UU	875,835	19,442	50	45.0	17516.7	90.1%		Test-UU	51,207	2,717	22	18.8	2327.6	85.7%
	Test-US	3,561,568	19,444	201	183.2	17719.2	91.1%		Test-US	211,388	2,717	89	77.8	2375.1	87.4%
	Test-SU	3,501,599	77,774	50	45.0	70032.0	90.0%		Test-SU	204,583	10,869	22	18.8	9299.2	85.6%
	Test-SS	2,848,085	77,773	201	36.6	14169.6	18.2%		Test-SS	168,819	10,869	89	15.5	1896.8	17.5%
WVS W2	Full	5,056,141	24,558	260	205.9	19446.7	79.2%	WVS W3	Full	12,157,148	77,818	190	156.2	63985.0	82.2%
	Train	2,584,283	19,647	208	131.5	12424.4	63.2%		Train	6,336,974	62,255	152	101.8	41690.6	67.0%
	Test-UU	203,583	4,911	52	41.5	3915.1	79.7%		Test-UU	451,555	15,563	38	29.0	11883.0	76.4%
	Test-US	809,976	4,911	208	164.9	3894.1	79.3%		Test-US	1,974,622	15,563	152	126.9	12990.9	83.5%
	Test-SU	812,229	19,647	52	41.3	15619.8	79.5%		Test-SU	1,809,754	62,255	38	29.1	47625.1	76.5%
	Test-SS	646,070	19,647	208	32.9	3106.1	15.8%		Test-SS	1,584,243	62,254	152	25.4	10422.7	16.7%
WVS W4	Full	9,155,251	60,045	243	152.5	37675.9	62.7%	WVS W5	Full	11,740,808	83,975	226	139.8	51950.5	61.9%
	Train	4,745,477	48,036	195	98.8	24335.8	50.7%		Train	5,978,120	67,180	181	89.0	33028.3	49.2%
	Test-UU	348,091	12,009	48	29.0	7251.9	60.4%		Test-UU	480,096	16,795	45	28.6	10668.8	63.5%
	Test-US	1,488,412	12,009	195	123.9	7632.9	63.6%		Test-US	1,865,568	16,795	181	111.1	10307.0	61.4%
	Test-SU	1,386,902	48,036	48	28.9	28893.8	60.2%		Test-SU	1,922,495	67,180	45	28.6	42722.1	63.6%
	Test-SS	1,186,369	48,035	195	24.7	6083.9	12.7%		Test-SS	1,494,529	67,180	181	22.2	8257.1	12.3%
WVS W6	Full	18,132,467	89,565	306	202.5	59256.4	66.2%	BSA 2024	Full	424,017	4,120	190	102.9	2231.7	54.2%
	Train	9,259,529	71,652	245	129.2	37794.0	52.7%		Train	213,463	3,296	152	64.8	1404.4	42.6%
	Test-UU	732,588	17,913	61	40.9	12009.6	67.0%		Test-UU	18,104	824	38	22.0	476.4	57.8%
	Test-US	2,893,744	17,913	245	161.5	11811.2	65.9%		Test-US	67,081	824	152	81.4	441.3	53.6%
	Test-SU	2,931,724	71,652	61	40.9	48061.0	67.1%		Test-SU	72,004	3,296	38	21.8	1894.8	57.5%
	Test-SS	2,314,882	71,652	245	32.3	9448.5	13.2%		Test-SS	53,365	3,295	152	16.2	351.1	10.7%
BSA 2021	Full	311,271	6,249	133	49.8	2340.4	37.5%	BSA 2022	Full	348,629	6,699	139	52.0	2508.1	37.4%
	Train	162,645	5,000	107	32.5	1520.0	30.4%		Train	178,331	5,360	112	33.3	1592.2	29.7%
	Test-UU	11,338	1,249	26	9.1	436.1	34.9%		Test-UU	14,146	1,339	27	10.6	523.9	39.1%
	Test-US	50,843	1,249	107	40.7	475.2	38.0%		Test-US	55,771	1,339	112	41.7	498.0	37.2%
	Test-SU	45,784	4,999	26	9.2	1760.9	35.2%		Test-SU	55,799	5,359	27	10.4	2066.6	38.6%
	Test-SS	40,661	4,996	107	8.1	380.0	7.6%		Test-SS	44,582	5,358	112	8.3	398.1	7.4%
BSA 2023	Full	626,426	5,578	206	112.3	3040.9	54.5%	BSA 2025	Full	61,525	47,292	4	1.3	15381.2	32.5%
	Train	332,136	4,462	165	74.4	2012.9	45.1%		Train	39,356	31,030	4	1.3	9839.0	31.7%
	Test-UU	21,514	1,115	41	19.3	524.7	47.1%		Test-UU	0	0	0	0.0	0.0	0.0%
	Test-US	103,446	1,115	165	92.8	626.9	56.2%		Test-US	12,330	9,458	4	1.3	3082.5	32.6%
	Test-SU	86,296	4,461	41	19.3	2104.8	47.2%		Test-SU	0	0	0	0.0	0.0	0.0%
	Test-SS	83,034	4,462	165	18.6	503.2	11.3%		Test-SS	9,839	9,058	4	1.1	2459.8	27.2%

F Evaluation Method Specifications

↩ Back to Experimental Setup

We evaluate two families of learned methods with progressively richer information sources: LLM prompting methods and training-based methods.

F.1 LLM Prompting Methods

LLM prompting methods query pretrained language models with different context configurations. All methods use greedy decoding (temperature=0) and constrain the model to output valid JSON with an integer answer index.

Zero-shot. Provides only the question text with no respondent information, instructing the model to answer as if it were completing a generic survey. This setting tests whether LLMs have learned general population priors during pretraining.

Persona. Augments the question with a structured respondent profile containing the respondent's observed persona facts, such as demographic attributes (age, gender, education, income, and region). The model is instructed to predict how a person matching that profile would answer, with the profile formatted as a structured list of attribute-value pairs preceding the target question.

HQA. Extends Persona by providing up to $k = 8$ observed question–answer pairs from the target respondent's training history before presenting the target question. This method is only applicable when the test respondent appears in the training set (Test-SS and Test-SU).

PRA. Provides the respondent profile plus sampled training responses to the same target question from other respondents. This method requires the target question to appear in training (Test-SS and Test-US).

DACA. Combines HQA and PRA: the prompt includes the target respondent's historical answers together with retrieved same-question responses from other training respondents. It is evaluated only when both the respondent and target question occur in training (Test-SS).

All prompting methods support optional shuffling of option order to mitigate position bias. For each respondent-question pair, we deterministically permute options based on a hash of the respondent and question IDs, then reverse the permutation to recover the true option index from the model's prediction.

F.2 Training-Based Methods

Training-based methods learn from the full training set to predict survey responses.

F.2.1 Traditional Training Methods

ModernBERT Classifier. Fine-tunes a pretrained ModernBERT-base encoder with a classification head. The input is the concatenation of the respondent profile and question text:

```
[Task description] [Target Question] [Person Profile]
```

The encoder's [CLS] token representation (or mean-pooled token embeddings if no [CLS] is available) is passed through a dropout layer and a linear classifier with $K_{\max}$ outputs. Invalid options are masked before loss computation. We use the ModernBERT-base checkpoint (22 transformer layers and hidden size 768), add the classification head, and fine-tune the full model with learning rate 2×10^{-5}, weight decay 0.01, warmup ratio 0.06, batch size 16, for 3 epochs with sequence length 512.

F.2.2 LLM Fine-Tuning Methods

SFT (Supervised Fine-Tuning). Fine-tunes a pretrained causal language model on survey response prediction via next-token prediction. Each training example is a chat-formatted conversation:

```
User: [Task + Person Profile + Target Question]
Assistant: {"answer": <index>}
```

The model is trained to generate the JSON response containing the correct answer index, with loss computed only on the assistant's tokens. We apply the model's native chat template when available, and shuffle option order deterministically per (respondent, question) pair to reduce position bias. We train with learning rate 2×10^{-5}, cosine schedule, warmup ratio 0.03, gradient accumulation to effective batch size 16, and 1 epoch. For efficiency, we support LoRA (Low-Rank Adaptation) with rank 16, alpha 32, dropout 0.05, targeting all attention and FFN projection matrices.

At inference, we generate up to 32 new tokens with greedy decoding and parse the output JSON to extract the predicted answer index. If parsing fails or returns an invalid index, we default to the ordinal center.

G LLM Evaluation Prompts

↩ Back to Experimental Setup

This section presents the actual prompts used to evaluate LLMs on survey response prediction, filled with real examples from the WVS Wave 7 dataset. All prompts require the model to output a JSON object with an integer answer index.

G.1 Zero-shot Prompt

The Zero-shot prompt provides only the target question with no respondent information.

Zero-shot Evaluation Prompt

```
You are answering a survey question. Select the most appropriate
option.

Output your answer as JSON in this exact format:
{"answer": <index>}
where <index> is the integer index (0, 1, 2, ...) of your chosen
option.

# Question
How important is family in your life? Would you say it is very
important, rather important, not very important or not important at
 all?
0.  Very important
1.  Rather important
2.  Not very important
3.  Not at all important
```

G.2 Persona Prompt

The Persona prompt augments the question with a structured respondent profile containing demographic and attitudinal attributes.

Persona Evaluation Prompt

```
You are predicting how a specific person would answer a survey
question.

Output your answer as JSON in this exact format:
{"answer": <index>}
where <index> is the integer index (0, 1, 2, ...) of your chosen
option.

# Task
Based on the person profile, predict this person's answer to the
target question.
```

```
# Person Profile
A woman in her fifties who lives in a three-person household, is
single, and has no children. She completed lower secondary
education, while both of her parents had only early childhood
education or no formal schooling. She is self-employed in the
service sector, working for a private business, and is not the
chief wage earner in her home. Her household managed to save money
in the past year, and she identifies as working class, placing her
household on the fourth step of a ten-point income scale. Although
she considers herself not a religious person, God is very important
 in her life, she believes in God, life after death, heaven, and
hell, and she prays several times a day while attending religious
services once a week. She is an inactive member of a church or
religious organization and is not a member of a labour union.

# Target Question
How important is family in your life? Would you say it is very
important, rather important, not very important or not important at
 all?
0.  Very important
1.  Rather important
2.  Not very important
3.  Not at all important
```

G.3 HQA PROMPT

The HQA prompt extends Persona by providing observed question–answer pairs from the target respondent's training history (up to $k = 8$ examples).

HQA Evaluation Prompt

```
You are predicting how a specific person would answer a survey
question.

Output your answer as JSON in this exact format:
{"answer": <index>}
where <index> is the integer index (0, 1, 2, ...) of your chosen
option.

# Task
Use the person's profile and prior answers to predict their answer
to the target question.

# Person Profile
A man in his forties who lives in a four-person household, is
married, and has two children. He completed upper secondary
education. He is employed full-time in the industrial sector and is
 the chief wage earner in his home. His household income is at the
fifth step of a ten-point scale. He considers himself a religious
person, attends religious services monthly, and is an active member
 of a church organization.

# Historical Answers From This Same Person (HQA)
These are this person's answers to other training questions:

Q: How important is family in your life? Would you say it is very
important, rather important, not very important or not important at
 all?
Options:
0.  Very important
1.  Rather important
```

```
2.  Not very important
3.  Not at all important
A: 0

Q: How important are friends in your life? Would you say they are
very important, rather important, not very important or not
important at all?
Options:
0.  Very important
1.  Rather important
2.  Not very important
3.  Not at all important
A: 0

# Target Question
How important is leisure time in your life? Would you say it is
very important, rather important, not very important or not
important at all?
0.  Very important
1.  Rather important
2.  Not very important
3.  Not at all important
```

G.4 PRA PROMPT

The PRA prompt provides the respondent profile plus retrieved evidence showing how other training respondents answered the same target question.

PRA Evaluation Prompt

```
You are predicting how a specific person would answer a survey
question.

Output your answer as JSON in this exact format:
{"answer": <index>}
where <index> is the integer index (0, 1, 2, ...) of your chosen
option.

# Task
Use the person's profile and the retrieved same-question evidence
to predict the target person's answer.

# Person Profile
A woman in her fifties who lives in a three-person household, is
single, and has no children. She completed lower secondary
education, while both of her parents had only early childhood
education or no formal schooling. She is self-employed in the
service sector, working for a private business, and is not the
chief wage earner in her home. Her household managed to save money
in the past year, and she identifies as working class, placing her
household on the fourth step of a ten-point income scale. Although
she considers herself not a religious person, God is very important
 in her life, she believes in God, life after death, heaven, and
hell, and she prays several times a day while attending religious
services once a week. She is an inactive member of a church or
religious organization and is not a member of a labour union.

# Population-Respondent Evidence (PRA)
The following evidence comes from training respondents who answered
 the same target question.

Answer distribution for this question (n=1247 training responses):
```

```
- Option 0 (Very important): 89.3%
- Option 1 (Rather important): 8.7%
- Option 2 (Not very important): 1.4%
- Option 3 (Not at all important): 0.6%

Sample responses from similar demographic groups:
- User 226 (working class, religious): answered 0
- User 1801 (self-employed, religious services weekly): answered 0
- User 3166 (lower secondary education): answered 0

# Target Question
How important is family in your life? Would you say it is very
important, rather important, not very important or not important at
 all?
0.  Very important
1.  Rather important
2.  Not very important
3.  Not at all important
```

G.5 DACA PROMPT

DACA combines HQA and PRA in one prompt. It is applicable when both the target respondent and target question are observed in training; the implementation uses the same JSON contract and option formatting as the other prompting methods.

DACA Evaluation Prompt

```
You are predicting how a specific person would answer a survey
question.

Output valid JSON only:
{"answer": <index>}
where <index> is the integer index (0, 1, 2, ...) of the chosen
option.

# Task
Use the person's profile, historical answers, and population
evidence for the target question.

# Person Profile
[PERSON_PROFILE]

# Historical Answers From This Same Person (HQA)
[HISTORICAL_ANSWERS]

# Population-Respondent Evidence (PRA)
[SAME-QUESTION_TRAINING_EVIDENCE]

# Target Question
[TARGET_QUESTION_AND_OPTIONS]
```

H DETAILED EXPERIMENTAL ANALYSES

↩ Back to Section 4

H.1 RELIABILITY OF THE CONSTRUCTION PIPELINE (RQ1)

Every instance in S^2Bench passes through the construction pipeline of Section 3.1, and five of its steps rely on an LLM. An error at any of these steps would silently corrupt the benchmark rather than cause a visible failure. A mis-extracted question would pair respondents' answers with the wrong

WVS Wave 7



3 independent human annotators

Pipeline step	Items	Appropriate	Pairwise agreement A-B	A-C	B-C
1 Question extraction	251	**100%**	100%	100%	100%
2 Question cleaning	251	**100%**	100%	100%	100%
3 Question rewriting	251	**100%**	100%	100%	100%
4 Persona identification	251	**100%**	100%	100%	100%
5 Profile writing	200	**100%**	100%	100%	100%

Figure 14: Human evaluation of the construction pipeline on WVS Wave 7 by three independent annotators. For each of the five LLM-based steps, listed in pipeline order, the table gives the number of evaluated items, the appropriateness rate, i.e., the fraction of items judged appropriate by all three annotators, and the pairwise agreement between annotators A–B, A–C, and B–C, i.e., the fraction of items on which the two annotators give the same label.

wording or options; an over-aggressive cleaning step would delete substantive options or usable items; a rewrite that alters the meaning would ask the simulator a different question from the one respondents actually answered; a misclassified question would either leak a target answer into the persona or withhold background information from it; and a profile that adds unobserved traits would give the simulator information the respondent never provided. Because all downstream results rest on these outputs, we treat the reliability of the pipeline as the first experiment and verify it through a human evaluation before evaluating any model.

Evaluation setup. We evaluate the pipeline on Wave 7 of the World Values Survey (WVS), which is processed by the same pipeline and the same processing LLM (DeepSeek-V4-Pro) as the rest of the benchmark. For the four question-level steps—question extraction, question cleaning, question rewriting, and persona identification—we evaluate all 251 questions retained after processing. For profile writing, we evaluate the first 200 persona profiles in processing order. For each step, annotators see the step's input next to its output and judge the output against the following criterion.

- *Question extraction.* Annotators compare a codebook or questionnaire passage with the extracted question object and check that the identifier, wording, question type, and options are complete and correctly matched.
- *Question cleaning.* Annotators check that only missing-value labels were removed from the option list and that the keep-or-remove decision and the single-choice filter were applied correctly.
- *Question rewriting.* Annotators compare t_q with $\tilde{t}_q$ and check that the rewrite is self-contained while preserving the original language, meaning, and option order.
- *Persona identification.* Annotators check that each question is correctly assigned to the persona set $\mathcal{Q}_{\text{per}}$ or the target set $\mathcal{Q}$.
- *Profile writing.* Annotators compare the persona facts $\mathcal{P}_r$ with the profile p_r and check that every fact is expressed correctly and that no attitude or unobserved trait is added.

Three annotators, all PhD candidates in relevant research fields, label every sampled item independently as *appropriate* or *inappropriate*. They are blind to one another's labels, the identity of the processing LLM, and the fact that the evaluated questions were drawn from the set retained by the pipeline. We report two quantities per step: the *appropriateness rate*, i.e., the fraction of items judged appropriate, and the pairwise *agreement* between annotators A–B, A–C, and B–C, i.e., the fraction of items on which the two annotators give the same label. We report raw agreement rather than Cohen's κ because κ is undefined when all labels fall into a single category, which is the case here.

Results. As Figure 14 shows, all three annotators judged every evaluated output of every step appropriate, so the appropriateness rate is 100% for all five steps and the pairwise agreement is 100% for all three annotator pairs. No step produced an error that any annotator flagged, including the two steps most exposed to subtle mistakes: rewriting, where a paraphrase could shift the meaning of a question, and profile writing, where a fluent paragraph could introduce unsupported traits. Because no item was judged inappropriate, the evaluation bounds rather than estimates the error rate: with zero

observed errors, the one-sided 95% upper confidence bound on a step's error rate is approximately $3/N$, corresponding to about 1.2% for each question-level step and 1.5% for profile writing.

Analysis. We attribute this result to the nature of the evaluated operations. Each step is a well-defined and heavily constrained text transformation rather than an open-ended generation task: extraction and cleaning copy and filter content that is already present in the source, rewriting must preserve the meaning and the option order, persona identification is a binary decision about whether a question describes the respondent, and profile writing may use only the given facts. These constraints, stated explicitly in the prompts of Appendix C, leave the processing LLM little room for the kinds of errors listed above, and such structured operations are well within the capabilities of current LLMs. The steps outside the evaluation, namely parsing, index normalization, and benchmark processing, do not involve an LLM and are therefore deterministic given their inputs. In particular, index normalization constructs g_q by mapping the lowest-indexed retained option to 0 and the remaining retained options to consecutive indices in option order.

H.2 Overall Performance (RQ2)

Before comparing individual design choices, we establish how well current LLMs and methods predict survey responses at all, and whether the four evaluation settings pose different levels of difficulty. This comparison also tests the premise of the two-axis split: if accuracy did not depend on whether the respondent and the question are seen, the four settings would add nothing beyond a single random split.

Evaluation setup. We evaluate the 13 LLMs of Section 4 with the five prompting methods and the three training-based methods on the most recent wave of each of the 15 survey families. Each prompting method is applied only in the settings where its context exists: HQA requires the respondent's own answers to other questions and therefore runs on Test-SU and Test-SS, PRA requires other respondents' answers to the target question and therefore runs on Test-US and Test-SS, and DACA requires both and therefore runs on Test-SS only. Figure 2 shows, for every model and setting, the accuracy of Zero-shot, Persona, and the applicable context-based method, together with the training-based methods, and Table 16 in Appendix I lists micro-averaged accuracy, MAE, KL divergence, and EMD for every model and method.

Results. Without any information about the respondent, all models perform similarly: Zero-shot prompting reaches an average accuracy of 0.382–0.385 in the four settings, with individual models between 0.346 (Qwen3-30B-A3B) and 0.414 (DeepSeek-V4-Pro). Adding the persona profile raises the average to 0.410–0.417, and adding context raises it further: HQA reaches 0.436 on Test-SU and 0.434 on Test-SS, PRA reaches 0.441 on both Test-US and Test-SS, and DACA, which combines both sources, reaches 0.452 on Test-SS. The strongest prompted models, DeepSeek-V4-Pro, Mimo-v2.5-Pro, and LongCat-2.0, reach 0.478–0.479 with DACA, whereas Gemma-3-27B-it and Qwen3-30B-A3B remain the weakest with context, with Qwen3-30B-A3B reaching only 0.385 even with DACA. Training-based methods are strongest when the question has been seen in training: BERT achieves the best accuracy on Test-SS (0.522) and Test-US (0.485), ahead of SFT on Qwen2.5-7B (0.492 and 0.479) and of the best prompted model (0.479 and 0.465). When the question is unseen, this advantage disappears: on Test-SU, the best HQA prompts (Mimo-v2.5-Pro, 0.460; DeepSeek-V4-Pro, 0.457) outperform both SFT (0.445) and BERT (0.426), and on Test-UU, Persona prompting with DeepSeek-V4-Flash (0.440) is on par with SFT on Qwen2.5-7B (0.437), while BERT drops to 0.392, below the average Persona model.

Analysis. Three observations follow. First, the difficulty of the task is not an artifact of the split: Zero-shot accuracy is nearly identical across the four settings because this method uses no training data, so the differences observed for the other methods reflect what each method can exploit rather than differences in the test sets themselves. Second, each source of information contributes a distinct signal. Persona information adds about 3 accuracy points but reduces the KL divergence between predicted and observed answer distributions from about 13 to about 3, indicating that its main effect is to make predictions demographically plausible rather than individually precise. Respondent history and population answers each add a further 2–3 points, and combining them in DACA adds about 4 points over Persona on Test-SS, so the two context types are complementary but not additive. Third, the ranking of methods depends on the setting, which is exactly what a single aggregate score would hide. Models trained on the benchmark exploit seen questions very effectively, but BERT loses 13.0

points from Test-SS to Test-UU and SFT on Qwen2.5-7B loses 5.5 points, whereas prompted LLMs lose little, because their knowledge of how people answer a question does not come from its training labels. Overall, even the best method answers only about half of the held-out questions correctly, leaving substantial room for improvement in every setting. Full results for all evaluated models and all datasets are provided in Appendix I (Table 17 through Table 31).

H.3 Effect of Model Scale (RQ3)

The models in RQ2 differ in size, but also in training data, architecture, and post-training, so their ranking alone cannot show whether scale itself matters. We therefore isolate model size within two model families and ask how accuracy grows with the number of parameters, and whether this growth depends on the kind of information the model receives.

Evaluation setup. We evaluate the Qwen2.5 family at seven sizes (0.5B to 72B) and the Qwen3 family at six sizes (0.6B to 32B), so that all models within a family share the same training recipe and differ mainly in size. Each model is prompted with Persona, HQA, PRA, and DACA on the Test-SS split, the only setting in which all four methods apply. We aggregate accuracy over all 15 datasets as $\sum_d n_d \,\mathrm{Acc}_d \,/\, \sum_d n_d$, where n_d is the number of Test-SS response instances in dataset d. Figure 3 plots accuracy against model size on a logarithmic axis.

Results. Accuracy rises steeply with size among small models and then levels off. With Persona prompting, Qwen2.5 improves from 0.276 at 0.5B to 0.407 at 14B, a gain of 13.1 points, but gains only 1.2 further points from 14B to 72B (0.419), with the 32B model (0.405) slightly below the 14B model. HQA shows the same shape, with Qwen2.5 improving from 0.283 to 0.422 between 0.5B and 14B and then only to 0.437 at 72B. Under PRA, Qwen2.5 peaks at 32B (0.457) and declines at 72B (0.447), while under DACA it still improves from 0.449 at 32B to 0.456 at 72B. The Qwen3 family follows a similar trajectory, from 0.284–0.305 at 0.6B to 0.415–0.440 at 32B, although small Qwen3 models are not strictly monotonic: under Persona, the 1.7B model (0.288) is below the 0.6B model (0.302). At comparable sizes the two families stay within a few points of each other, with Qwen3 slightly ahead at 4B–8B.

Analysis. The sharp rise at small scales indicates that a minimum level of language understanding is needed to use a persona or context at all: models up to 3B remain between 0.27 and 0.34 under every method, so additional information cannot compensate for limited capacity. Beyond about 14B, however, additional parameters yield little, and the largest gains no longer come from scale but from information: moving Qwen2.5-72B from Persona to DACA adds 3.7 points, about three times the gain from increasing the model size fivefold from 14B to 72B under Persona. The saturation is also consistent with RQ2, where SFT on Qwen2.5-7B (0.492) outperforms every prompted model on Test-SS, including models an order of magnitude larger, showing that task-specific training on survey data is a far more efficient route to accuracy than scaling a general model. In practice, a mid-sized model with rich context is therefore a more cost-effective survey simulator than a much larger model with a persona alone.

H.4 Impact of Reasoning Mode (RQ4)

Reasoning modes, in which a model generates an explicit chain of thought before answering, improve LLMs on mathematics and coding, and it is natural to ask whether they also help a model infer how a particular person would answer a survey question. Unlike those tasks, survey-response prediction has no verifiable derivation, so it is not obvious that explicit reasoning should help.

Evaluation setup. We compare each Qwen3 model with and without its thinking mode, since the same weights support both modes and the comparison is therefore not confounded by other differences between models. We use five sizes, the dense 4B, 8B, 14B, and 32B models and the mixture-of-experts Qwen3-30B-A3B, together with Persona, HQA, PRA, and DACA prompting, on the Test-SS split of WVS. Figure 4 shows the accuracy of each model–method pair in both modes, and we additionally record the end-to-end evaluation time.

Results. Thinking improves accuracy for almost every model–method pair, but the size of the gain varies widely. Gains are largest for the smaller and the mixture-of-experts models, 4B, 8B, and 30B-A3B, and for the context-rich methods PRA and DACA, where the thinking polygon in Figure 4 lies clearly outside the non-thinking one. For the strongest model, Qwen3-32B, the two modes are

almost indistinguishable under Persona and HQA, and the gain under DACA is modest: accuracy increases from 43.35% to 44.86%, i.e., by 1.51 percentage points. This gain is expensive: on the same $5{,}135$ WVS Test-SS requests, enabling thinking increases the evaluation time from about 6 minutes to 2 hours and 45 minutes, a slowdown of about $27\times$.

Analysis. The pattern suggests that reasoning helps mainly by making better use of the information in the prompt rather than by adding knowledge about respondents. With PRA and DACA, the prompt contains many other respondents' answers, and explicit reasoning lets the model aggregate them and relate them to the target persona; with a persona alone there is little to reason over, so the gain is small. Reasoning also acts as a partial substitute for scale: it narrows the gap between small and large models, but it leaves the largest model almost unchanged, where the remaining errors appear to reflect genuine uncertainty about individual answers rather than failures to process the context. Given a gain of about 1.5 points at a $27\times$ cost for the best model, reasoning mode is difficult to justify for large-scale survey simulation, where millions of responses may need to be generated, and the compute is better spent on a larger model or richer context.

H.5 General Capability and Simulation Performance (RQ5)

LLM progress is usually measured by general-purpose benchmarks, and a common assumption is that a model that scores higher on them will also be a better simulator of human respondents. If this assumption held, general benchmarks could be used to select survey simulators without a dedicated evaluation; if it does not, survey simulation must be measured directly.

Evaluation setup. We take each model's MMLU-Pro score from the public TIGER-Lab MMLU-Pro leaderboard[2] and relate it to the model's accuracy on Test-SS, the only setting in which HQA, PRA, and DACA can all be compared. For each method, we fit a linear regression across models and report Pearson's r and Spearman's ρ with their p-values, and we add a 1:1 reference line along which one point of MMLU-Pro would translate into one point of accuracy. Figure 5 shows the results, with models spanning MMLU-Pro scores from about 62 to 84.

Results. Accuracy increases with MMLU-Pro under all three methods, but the association is only moderate and statistically weak: Pearson's r is 0.55 for HQA ($p = 0.062$), 0.52 for PRA ($p = 0.081$), and 0.53 for DACA ($p = 0.073$), and the rank correlations are lower still, with Spearman's ρ between 0.45 and 0.52 and p-values between 0.081 and 0.138. The fitted slopes are far below the 1:1 line: across the 22-point MMLU-Pro range, the fitted accuracy rises by only about 4 points under each method. The scatter around the fits is large; for example, under DACA, models with MMLU-Pro scores between 66 and 69 reach up to about 48% accuracy, as high as the models scoring above 80.

Analysis. General capability is thus a weak predictor of simulation quality. The flat slopes indicate that the skills rewarded by MMLU-Pro, such as expert knowledge and multi-step problem solving, are only loosely related to what survey simulation requires, namely calibrated knowledge of how opinions are distributed across social groups and the ability to transfer patterns from context to a specific respondent. The large spread among models with similar MMLU-Pro scores further suggests that factors not captured by general benchmarks, such as the coverage of the pretraining data on social attitudes and the style of post-training, matter as much as general capability. Consequently, general benchmark scores should not be used as a proxy for simulation quality, and progress on survey simulation needs to be measured with dedicated benchmarks such as S^2Bench and pursued with targeted training rather than expected as a by-product of general improvements.

H.6 Transferability of Fine-Tuning Gains (RQ6)

RQ2 showed that training-based methods are strongest on seen questions, but the aggregate accuracies do not reveal whether their gains transfer equally along the two axes. This question matters for how fine-tuned simulators can be used: a simulator that transfers to new respondents can be applied to a new sample of an existing survey, whereas one that transfers to new questions can be used to pre-test a questionnaire before it is fielded.

Evaluation setup. We compare the three training-based methods, BERT and SFT on Qwen2.5-7B and Qwen3-8B, on Test-US, where respondents are unseen but questions are seen, and on Test-SU,

[2] https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

where questions are unseen but respondents are seen. Because datasets differ in intrinsic difficulty, we first subtract the mean accuracy of the Persona baseline in the same dataset and setting, which gives a persona-adjusted gain for each method. For each of the 15 datasets we then compute the difference between the gain on Test-US and the gain on Test-SU, shown in Figure 6.

Results. The difference is positive for most datasets and all three methods, i.e., fine-tuning helps more on new respondents than on new questions. BERT shows the largest asymmetry, with a positive difference on all 15 datasets and a mean of $+9.4$ percentage points, reaching $+33.0$ on CSES, $+16.0$ on CGSS, and $+13.8$ on CSS. The two SFT models show much smaller asymmetries, with means of $+2.3$ points for Qwen2.5-7B and $+1.8$ points for Qwen3-8B, and a few datasets reverse the sign: ATP (-3.6 and -2.5) and ISSP (-2.3 and -0.8) for both models, and AmericasBarometer (-0.8) for Qwen2.5-7B. CGSS and CSS are among the three datasets with the largest asymmetry for all three methods, whereas ISSP and AmericasBarometer are among the three with the smallest.

Analysis. Fine-tuning mainly learns how the population answers the questions it was trained on, which transfers readily to new respondents drawn from the same population but not to questions whose answer distributions were never observed. The size of the asymmetry follows how much a method relies on question-specific memorization: BERT, which learns a classifier from scratch on top of an encoder, has almost no prior knowledge of opinions and therefore gains little on new questions, whereas the SFT models start from LLMs that already encode general knowledge of public opinion and retain part of it when the question is new. The consistent ordering of datasets across methods further suggests that thematic coherence within a questionnaire affects how well knowledge about seen questions carries over to unseen ones. A plausible explanation is that each wave of ISSP and ATP centers on a relatively focused theme, creating greater overlap between seen and unseen questions, whereas broad, multi-topic surveys offer less related information to transfer. For survey simulation, this implies that fine-tuning on historical data is well suited to extending an existing survey to new samples, while forecasting responses to genuinely new questions remains an open problem for which LLM priors are currently more valuable than training labels.

I Detailed Result Tables

↩ Back to RQ2

Table 16 reports micro-averaged Acc., MAE, KL, and EMD for every evaluated prompting and training-based method on each of the four evaluation settings (Test-UU, Test-US, Test-SU, Test-SS), extending the summary in Figure 2 with the full method list and all four metrics side by side. The remaining tables in this section (Table 17 through Table 31) break these same results down by individual survey, allowing readers to inspect whether a method's aggregate performance is representative or is instead driven by a small number of datasets.

Table 16: Overall micro-averaged results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.414	1.015	12.408	0.310	0.402	1.069	12.667	0.323	0.405	1.063	12.574	0.316	0.399	1.045	12.747	0.320
DeepSeek-V4-Flash	0.393	1.090	12.874	0.329	0.401	1.111	12.708	0.330	0.401	1.102	12.662	0.330	0.398	1.089	12.785	0.331
Mimo-v2.5	0.394	1.094	12.857	0.330	0.391	1.084	12.936	0.332	0.395	1.067	12.790	0.323	0.393	1.076	12.900	0.331
Mimo-v2.5-Pro	0.379	1.109	13.204	0.348	0.387	1.114	13.017	0.341	0.385	1.128	13.026	0.340	0.392	1.092	12.918	0.336
LongCat-2.0	0.389	1.102	12.968	0.330	0.395	1.086	12.842	0.330	0.404	1.069	12.586	0.322	0.391	1.080	12.930	0.327
GPT-4o-mini	0.362	1.128	13.602	0.353	0.364	1.160	13.545	0.355	0.362	1.135	13.555	0.348	0.359	1.150	13.682	0.354
GPT-5-mini	0.390	1.089	12.952	0.325	0.392	1.077	12.916	0.326	0.370	1.145	13.365	0.339	0.378	1.104	13.246	0.336
Qwen2.5-72B-Instruct	0.387	1.091	13.020	0.327	0.378	1.092	13.228	0.333	0.390	1.086	12.912	0.326	0.378	1.073	13.250	0.329
Qwen3-32B	0.383	1.168	13.114	0.344	0.381	1.142	13.167	0.344	0.386	1.135	13.006	0.340	0.383	1.133	13.125	0.342
Qwen3-30B-A3B	0.351	1.161	13.846	0.359	0.346	1.168	13.956	0.362	0.348	1.175	13.878	0.362	0.349	1.152	13.913	0.358
Llama-3.3-70B-Instruct	0.386	1.093	13.036	0.337	0.390	1.109	12.963	0.339	0.389	1.095	12.943	0.337	0.379	1.131	13.213	0.350
Gemma-3-27B-it	0.374	1.087	13.317	0.330	0.384	1.081	13.100	0.329	0.378	1.094	13.191	0.333	0.376	1.086	13.290	0.333
Mistral-Small-3.1-24B-Instruct	0.393	1.061	12.892	0.316	0.389	1.061	12.976	0.322	0.384	1.056	13.047	0.324	0.385	1.058	13.074	0.322
Avg. (Zero-shot)	*0.384*	*1.099*	*13.084*	*0.334*	*0.385*	*1.104*	*13.079*	*0.336*	*0.384*	*1.104*	*13.041*	*0.334*	*0.382*	*1.098*	*13.159*	*0.336*
Persona																
DeepSeek-V4-Pro	0.434	0.956	3.534	0.179	0.429	0.960	3.464	0.175	0.432	0.964	2.267	0.168	0.428	0.955	3.696	0.179
DeepSeek-V4-Flash	**0.440**	**0.946**	3.117	**0.171**	0.437	0.966	3.186	0.174	0.436	0.956	2.022	0.162	0.436	0.959	3.515	0.176
Mimo-v2.5	0.432	0.958	3.263	0.177	0.430	0.968	2.937	0.173	0.425	0.970	2.022	0.170	0.422	0.977	3.419	0.180
Mimo-v2.5-Pro	0.431	0.968	2.945	0.173	0.430	0.973	2.886	0.168	0.433	0.967	1.770	0.157	0.430	0.961	3.194	0.176
LongCat-2.0	0.433	**0.945**	3.171	**0.172**	0.431	0.948	3.001	0.172	0.433	0.943	2.214	0.164	0.426	0.951	3.523	0.178
GPT-4o-mini	0.399	1.052	**2.149**	0.178	0.389	1.072	1.950	0.177	0.394	1.058	1.243	0.169	0.385	1.070	2.392	0.184
GPT-5-mini	0.411	0.962	2.978	0.174	0.417	0.955	3.023	0.173	0.411	0.963	2.041	0.168	0.414	0.961	3.375	0.181
Qwen2.5-72B-Instruct	0.409	0.974	2.944	0.184	0.403	0.990	2.830	0.182	0.407	0.980	1.978	0.176	0.399	0.992	3.118	0.185
Qwen3-32B	0.413	1.012	3.308	0.189	0.406	1.026	3.226	0.191	0.413	1.008	2.256	0.179	0.402	1.032	3.507	0.197
Qwen3-30B-A3B	0.389	1.063	**2.850**	0.194	0.381	1.056	2.617	0.191	0.390	1.060	1.719	0.184	0.380	1.056	2.943	0.196
Llama-3.3-70B-Instruct	0.433	0.947	3.620	0.181	0.427	0.956	3.509	0.178	0.427	0.952	2.514	0.175	0.420	0.966	3.985	0.190
Gemma-3-27B-it	0.383	1.029	3.109	0.193	0.382	1.035	2.661	0.190	0.386	1.029	1.935	0.185	0.383	1.029	2.976	0.195
Mistral-Small-3.1-24B-Instruct	0.418	0.983	3.145	0.189	0.405	1.027	3.124	0.196	0.416	0.992	2.292	0.181	0.401	1.029	3.442	0.198
Avg. (Persona)	*0.417*	*0.984*	*3.087*	*0.181*	*0.413*	*0.995*	*2.955*	*0.180*	*0.415*	*0.988*	*2.021*	*0.172*	*0.410*	*0.995*	*3.314*	*0.186*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	**0.457**	0.922	1.689	0.151	0.456	0.925	2.956	0.168
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.452	0.935	1.503	**0.149**	0.455	0.944	2.853	0.169
Mimo-v2.5	–	–	–	–	–	–	–	–	0.453	0.917	1.691	0.155	0.455	0.927	2.970	0.172
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	**0.460**	0.917	1.347	**0.146**	0.458	0.922	2.772	0.165
LongCat-2.0	–	–	–	–	–	–	–	–	0.456	**0.912**	1.811	0.157	0.455	0.917	3.103	0.174
GPT-4o-mini	–	–	–	–	–	–	–	–	0.411	1.051	**1.224**	0.177	0.410	1.050	2.622	0.198
GPT-5-mini	–	–	–	–	–	–	–	–	0.426	0.941	1.486	0.158	0.433	0.932	2.916	0.173
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.436	0.939	1.808	0.166	0.427	0.948	2.939	0.179
Qwen3-32B	–	–	–	–	–	–	–	–	0.432	0.972	1.855	0.170	0.427	0.985	3.028	0.187
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.403	1.041	**1.184**	0.173	0.395	1.050	2.569	0.200
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.447	**0.908**	1.937	0.161	0.444	0.926	3.204	0.176
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.414	0.992	1.528	0.172	0.405	1.006	2.680	0.190
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.427	1.005	1.828	0.179	0.420	1.030	3.065	0.197
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.436*	*0.958*	*1.607*	*0.163*	*0.434*	*0.966*	*2.898*	*0.181*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.462	0.923	3.818	0.181	–	–	–	–	0.461	0.919	4.178	0.180
DeepSeek-V4-Flash	–	–	–	–	0.451	0.969	3.092	0.171	–	–	–	–	0.451	0.960	3.218	0.173
Mimo-v2.5	–	–	–	–	0.451	0.955	2.800	**0.164**	–	–	–	–	0.449	0.950	2.997	0.167
Mimo-v2.5-Pro	–	–	–	–	0.457	0.945	2.982	**0.163**	–	–	–	–	0.458	0.933	3.272	0.165
LongCat-2.0	–	–	–	–	0.465	0.926	4.121	0.183	–	–	–	–	0.469	0.919	4.542	0.186
GPT-4o-mini	–	–	–	–	0.443	0.977	2.851	0.171	–	–	–	–	0.441	0.972	3.058	0.172
GPT-5-mini	–	–	–	–	0.451	0.926	3.314	0.177	–	–	–	–	0.453	0.932	3.743	0.178
Qwen2.5-72B-Instruct	–	–	–	–	0.443	0.952	3.659	0.178	–	–	–	–	0.441	0.947	4.039	0.182
Qwen3-32B	–	–	–	–	0.437	1.002	3.146	0.185	–	–	–	–	0.432	1.003	3.522	0.190
Qwen3-30B-A3B	–	–	–	–	0.377	1.143	**1.608**	0.169	–	–	–	–	0.379	1.134	**1.725**	0.169
Llama-3.3-70B-Instruct	–	–	–	–	0.450	0.946	3.699	0.184	–	–	–	–	0.447	0.949	4.062	0.190
Gemma-3-27B-it	–	–	–	–	0.395	1.092	**1.916**	0.172	–	–	–	–	0.396	1.083	2.180	0.174
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.447	0.990	4.079	0.197	–	–	–	–	0.452	0.979	4.387	0.197
Avg. (PRA)	–	–	–	–	*0.441*	*0.980*	*3.160*	*0.177*	–	–	–	–	*0.441*	*0.975*	*3.456*	*0.179*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.479	0.871	3.703	0.173
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.467	0.934	3.189	0.168
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.468	0.910	2.865	0.168
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.478	0.893	2.991	**0.162**
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.478	0.895	3.605	0.173
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.440	0.998	2.701	0.179
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.452	0.932	3.065	0.167
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.458	0.920	3.635	0.180
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.442	0.985	2.854	0.178
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.385	1.120	**1.632**	0.173
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.459	0.911	3.334	0.174
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.425	1.019	2.507	0.182
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.449	0.991	3.697	0.195
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.452*	*0.952*	*3.060*	*0.175*
Training																
BERT	0.392	1.091	7.195	0.261	**0.485**	**0.876**	5.197	0.192	0.426	1.020	3.704	0.216	**0.522**	**0.797**	4.159	0.166
SFT (Qwen2.5-7B-Instruct)	**0.437**	0.971	3.864	0.193	**0.479**	**0.879**	3.744	0.167	0.445	0.949	2.794	0.177	**0.492**	**0.840**	3.912	**0.164**
SFT (Qwen3-8B)	0.431	0.984	3.970	0.188	0.466	0.904	3.566	0.168	0.444	0.949	2.824	0.172	0.475	0.885	3.916	0.171

Table 17: Afrobarometer results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.366	1.009	13.406	0.310	0.347	1.073	13.838	0.334	0.388	0.968	12.881	0.299	0.361	1.010	13.535	0.313
DeepSeek-V4-Flash	0.333	1.077	14.152	0.339	0.345	1.101	13.883	0.343	0.372	1.095	13.253	0.333	0.341	1.096	13.984	0.342
Mimo-v2.5	0.343	1.062	13.936	0.326	0.324	1.129	14.368	0.353	0.372	0.998	13.261	0.308	0.356	1.072	13.653	0.337
Mimo-v2.5-Pro	0.335	1.179	14.102	0.366	0.332	1.117	14.189	0.350	0.369	1.094	13.320	0.336	0.349	1.091	13.793	0.343
LongCat-2.0	0.384	0.993	12.975	0.295	0.359	1.079	13.556	0.332	0.418	0.931	12.192	0.279	0.362	1.043	13.509	0.321
GPT-4o-mini	0.309	1.154	14.715	0.375	0.324	1.123	14.356	0.352	0.345	1.050	13.884	0.339	0.327	1.110	14.304	0.345
GPT-5-mini	0.352	1.073	13.721	0.330	0.359	1.080	13.548	0.334	0.372	1.017	13.261	0.311	0.365	1.061	13.426	0.331
Qwen2.5-72B-Instruct	0.353	1.058	13.687	0.311	0.305	1.105	14.805	0.347	0.370	1.028	13.311	0.310	0.341	1.044	13.999	0.326
Qwen3-32B	0.374	1.148	13.223	0.337	0.330	1.160	14.230	0.362	0.380	1.078	13.082	0.324	0.345	1.115	13.901	0.349
Qwen3-30B-A3B	0.311	1.118	14.666	0.360	0.308	1.118	14.723	0.349	0.308	1.180	14.744	0.385	0.314	1.078	14.608	0.338
Llama-3.3-70B-Instruct	0.353	1.084	13.687	0.337	0.350	1.104	13.764	0.343	0.359	1.037	13.554	0.322	0.363	1.079	13.483	0.338
Gemma-3-27B-it	0.356	1.053	13.621	0.323	0.331	1.083	14.209	0.336	0.377	1.021	13.144	0.313	0.347	1.054	13.849	0.327
Mistral-Small-3.1-24B-Instruct	0.354	1.025	13.671	0.312	0.329	1.080	14.246	0.336	0.379	0.959	13.103	0.293	0.335	1.022	14.118	0.319
Avg. (Zero-shot)	*0.348*	*1.079*	*13.812*	*0.332*	*0.334*	*1.104*	*14.132*	*0.344*	*0.370*	*1.035*	*13.307*	*0.319*	*0.347*	*1.067*	*13.859*	*0.333*
Persona																
DeepSeek-V4-Pro	0.379	0.978	3.803	0.177	0.357	0.990	3.897	0.186	0.392	0.939	2.782	0.167	0.363	0.981	4.548	0.189
DeepSeek-V4-Flash	0.389	0.976	4.530	**0.175**	0.358	1.059	4.240	0.206	0.402	0.958	2.665	0.169	0.366	1.046	4.149	0.200
Mimo-v2.5	0.370	1.022	3.867	0.200	0.359	1.033	3.174	0.182	0.404	0.963	2.264	0.179	0.371	1.028	3.515	0.174
Mimo-v2.5-Pro	0.376	1.011	**3.439**	0.193	0.349	1.031	3.152	0.185	0.402	0.957	1.846	0.159	0.360	1.010	3.722	0.184
LongCat-2.0	0.385	**0.956**	4.391	0.195	0.372	1.001	3.771	0.184	0.416	0.904	3.322	0.185	0.372	0.989	4.265	0.183
GPT-4o-mini	0.366	1.052	**1.691**	**0.157**	0.324	1.125	**1.470**	0.171	0.377	1.026	**1.238**	0.141	0.349	1.085	1.945	**0.155**
GPT-5-mini	0.366	0.988	3.826	0.190	0.363	1.002	3.401	0.186	0.377	0.982	3.019	0.182	0.362	1.026	3.606	0.195
Qwen2.5-72B-Instruct	0.360	0.983	4.230	0.186	0.333	1.008	3.910	0.193	0.380	0.930	2.922	0.174	0.343	0.987	4.169	0.179
Qwen3-32B	0.367	1.056	4.059	0.200	0.353	1.041	3.597	0.201	0.402	0.980	2.655	0.179	0.364	1.015	4.355	0.192
Qwen3-30B-A3B	0.353	1.079	3.837	0.201	0.321	1.053	3.555	0.191	0.362	1.054	2.083	0.178	0.332	1.030	3.731	0.185
Llama-3.3-70B-Instruct	0.377	0.970	4.444	0.194	0.354	1.017	4.452	0.201	0.401	0.929	2.994	0.181	0.364	1.009	4.615	0.196
Gemma-3-27B-it	0.331	1.033	4.424	0.193	0.308	1.060	3.414	0.204	0.360	0.997	2.525	0.180	0.321	1.036	3.663	0.198
Mistral-Small-3.1-24B-Instruct	**0.402**	**0.946**	4.143	0.191	0.356	1.030	3.795	0.203	0.398	0.924	3.075	0.175	0.361	1.024	4.093	0.195
Avg. (Persona)	*0.371*	*1.004*	*3.899*	*0.189*	*0.347*	*1.035*	*3.525*	*0.192*	*0.390*	*0.965*	*2.569*	*0.173*	*0.356*	*1.021*	*3.875*	*0.187*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.433	0.915	1.947	0.138	0.395	0.975	3.124	0.170
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.425	0.957	1.488	**0.136**	0.385	1.065	3.570	0.200
Mimo-v2.5	–	–	–	–	–	–	–	–	0.434	0.911	1.666	0.148	0.397	0.990	3.334	0.170
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	**0.443**	0.905	1.301	**0.123**	0.398	0.980	2.394	0.158
LongCat-2.0	–	–	–	–	–	–	–	–	**0.443**	0.891	2.299	0.162	0.411	0.962	3.319	0.168
GPT-4o-mini	–	–	–	–	–	–	–	–	0.411	1.003	**1.046**	0.166	0.368	1.069	2.044	0.179
GPT-5-mini	–	–	–	–	–	–	–	–	0.399	0.959	1.832	0.154	0.382	1.001	2.857	0.171
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.422	**0.873**	2.681	0.160	0.377	0.963	3.343	0.177
Qwen3-32B	–	–	–	–	–	–	–	–	0.424	0.925	2.434	0.170	0.389	0.975	3.499	0.184
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.395	0.958	1.667	0.152	0.363	1.017	2.892	0.187
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.427	**0.880**	2.155	0.163	0.387	0.966	3.758	0.189
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.399	0.942	2.609	0.171	0.354	1.000	3.835	0.191
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.429	0.953	2.069	0.169	0.382	1.046	3.133	0.198
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.422*	*0.929*	*1.938*	*0.155*	*0.384*	*1.001*	*3.162*	*0.180*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.413	0.979	4.088	0.194	–	–	–	–	0.410	0.976	4.560	0.188
DeepSeek-V4-Flash	–	–	–	–	0.400	1.037	3.356	0.190	–	–	–	–	0.403	1.021	3.479	0.184
Mimo-v2.5	–	–	–	–	0.394	1.013	2.390	**0.158**	–	–	–	–	0.393	1.023	2.729	0.160
Mimo-v2.5-Pro	–	–	–	–	0.401	1.022	2.928	0.181	–	–	–	–	0.398	1.018	3.415	0.177
LongCat-2.0	–	–	–	–	0.415	1.018	5.100	0.215	–	–	–	–	0.428	0.983	5.955	0.213
GPT-4o-mini	–	–	–	–	0.393	1.047	3.181	0.191	–	–	–	–	0.390	1.047	3.140	0.177
GPT-5-mini	–	–	–	–	0.413	0.988	4.081	0.201	–	–	–	–	0.405	0.994	4.499	0.201
Qwen2.5-72B-Instruct	–	–	–	–	0.389	0.992	4.778	0.208	–	–	–	–	0.391	0.982	4.896	0.195
Qwen3-32B	–	–	–	–	0.399	1.019	3.388	0.196	–	–	–	–	0.400	1.016	4.036	0.193
Qwen3-30B-A3B	–	–	–	–	0.328	1.170	**1.625**	**0.165**	–	–	–	–	0.334	1.144	**1.568**	**0.152**
Llama-3.3-70B-Instruct	–	–	–	–	0.408	1.000	4.607	0.208	–	–	–	–	0.404	1.002	4.780	0.203
Gemma-3-27B-it	–	–	–	–	0.362	1.136	2.930	0.240	–	–	–	–	0.367	1.108	3.272	0.224
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.398	1.050	4.608	0.231	–	–	–	–	0.411	1.020	4.735	0.211
Avg. (PRA)	–	–	–	–	*0.393*	*1.036*	*3.620*	*0.198*	–	–	–	–	*0.395*	*1.026*	*3.928*	*0.191*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.427	0.958	4.316	0.182
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.408	1.025	3.955	0.187
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.414	0.982	2.937	0.169
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.423	0.987	3.168	0.181
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.429	0.968	4.634	0.192
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.389	1.067	2.956	0.198
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.410	0.993	3.594	0.185
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.410	0.963	4.496	0.197
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.415	0.984	3.256	0.194
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.346	1.119	**1.672**	0.157
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.418	0.955	3.844	0.186
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.395	1.040	3.045	0.202
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.403	1.063	4.288	0.224
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.407*	*1.008*	*3.551*	*0.189*
Training																
BERT	0.381	1.128	6.933	0.257	0.430	**0.955**	5.107	0.201	**0.444**	0.997	3.530	0.195	**0.488**	**0.858**	3.381	0.165
SFT (Qwen2.5-7B-Instruct)	**0.405**	1.007	5.453	0.231	**0.431**	**0.963**	6.122	0.218	0.424	0.978	4.421	0.207	**0.444**	**0.939**	6.144	0.213
SFT (Qwen3-8B)	0.399	1.042	6.726	0.248	**0.433**	0.966	6.012	0.231	0.428	0.981	4.960	0.219	0.442	0.951	6.492	0.222

Table 18: AmericasBarometer results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.556	1.487	9.417	0.347	0.519	0.969	10.273	0.346	0.518	1.587	10.274	0.386	0.527	0.890	10.081	0.329
DeepSeek-V4-Flash	0.523	1.718	10.185	0.391	0.522	0.960	10.212	0.340	0.518	1.570	10.258	0.383	0.461	0.999	11.605	0.398
Mimo-v2.5	0.510	1.667	10.480	0.376	0.504	1.029	10.613	0.362	0.497	1.650	10.749	0.396	0.497	0.977	10.774	0.366
Mimo-v2.5-Pro	0.538	1.667	9.830	0.376	0.469	1.073	11.428	0.401	0.486	1.723	11.009	0.420	0.487	0.966	11.005	0.370
LongCat-2.0	0.556	1.559	9.417	0.359	0.515	0.956	10.358	0.342	0.530	1.556	9.983	0.380	0.519	0.886	10.266	0.324
GPT-4o-mini	0.538	1.462	9.830	0.335	0.469	1.018	11.415	0.376	0.426	1.668	12.388	0.435	0.451	0.953	11.851	0.371
GPT-5-mini	**0.582**	1.436	8.827	0.330	0.523	0.912	10.175	0.330	0.467	1.602	11.438	0.414	0.476	0.945	11.267	0.360
Qwen2.5-72B-Instruct	0.528	1.633	10.067	0.369	0.496	1.004	10.795	0.358	0.506	1.599	10.549	0.388	0.469	0.961	11.436	0.365
Qwen3-32B	0.536	1.610	9.889	0.378	0.495	1.026	10.820	0.359	0.540	1.442	9.768	0.361	0.489	0.947	10.959	0.354
Qwen3-30B-A3B	0.487	1.667	11.011	0.406	0.467	1.050	11.476	0.379	0.480	1.611	11.147	0.421	0.455	1.010	11.744	0.386
Llama-3.3-70B-Instruct	0.515	1.605	10.362	0.388	0.504	1.002	10.613	0.366	0.456	1.705	11.698	0.441	0.492	0.946	10.897	0.367
Gemma-3-27B-it	0.569	1.513	9.122	0.343	0.530	0.929	10.017	0.334	0.534	1.564	9.891	0.376	0.502	0.920	10.666	0.346
Mistral-Small-3.1-24B-Instruct	0.564	1.492	9.240	0.339	0.526	0.906	10.115	0.311	0.542	1.372	9.722	0.343	0.487	0.866	11.005	0.349
Avg. (Zero-shot)	*0.539*	*1.578*	*9.821*	*0.364*	*0.503*	*0.987*	*10.639*	*0.354*	*0.500*	*1.588*	*10.683*	*0.396*	*0.486*	*0.943*	*11.043*	*0.360*
Persona																
DeepSeek-V4-Pro	0.544	1.546	3.237	**0.208**	0.570	0.751	1.945	**0.170**	0.510	1.589	1.416	0.222	0.545	0.745	2.348	0.175
DeepSeek-V4-Flash	0.559	1.464	3.087	0.218	**0.579**	0.761	1.805	0.182	0.513	1.546	1.523	0.200	0.535	0.779	2.825	0.181
Mimo-v2.5	0.541	1.438	3.063	0.219	0.548	0.790	2.043	0.182	0.498	1.645	1.422	0.223	0.526	0.787	1.761	0.185
Mimo-v2.5-Pro	**0.574**	1.377	2.428	0.236	0.567	0.770	1.852	0.171	0.508	1.554	1.500	0.185	0.539	0.767	2.046	0.182
LongCat-2.0	0.556	**1.374**	2.543	0.210	0.559	0.772	**1.700**	0.172	0.480	1.554	1.140	0.194	0.528	0.805	1.940	0.197
GPT-4o-mini	0.526	1.451	**1.766**	**0.204**	0.541	0.806	**1.223**	**0.154**	0.464	1.653	1.448	0.222	0.497	0.824	1.565	0.181
GPT-5-mini	0.436	1.644	4.266	0.247	0.527	0.814	2.141	0.196	0.441	1.613	2.683	0.253	0.509	0.793	2.455	0.191
Qwen2.5-72B-Instruct	0.492	1.556	3.229	0.233	0.533	0.787	1.737	0.182	0.443	1.643	2.246	0.251	0.493	0.808	2.221	0.207
Qwen3-32B	0.564	**1.364**	3.709	0.221	0.541	0.813	2.419	0.202	0.514	1.460	2.906	0.216	0.517	0.796	2.312	0.204
Qwen3-30B-A3B	0.449	1.579	3.215	0.221	0.527	0.827	1.744	0.176	0.418	1.632	1.199	0.243	0.488	0.884	2.391	0.218
Llama-3.3-70B-Instruct	0.487	1.597	3.831	0.239	0.546	0.814	1.908	0.181	0.486	1.576	2.236	0.229	0.517	0.809	2.498	0.197
Gemma-3-27B-it	0.495	1.695	**2.023**	0.245	0.530	0.790	1.773	0.172	0.466	1.597	1.415	0.244	0.505	0.806	1.654	0.200
Mistral-Small-3.1-24B-Instruct	0.546	1.462	4.088	0.246	0.567	**0.740**	2.084	0.179	0.510	1.540	2.017	0.269	0.527	0.779	2.181	0.198
Avg. (Persona)	*0.521*	*1.504*	*3.114*	*0.227*	*0.549*	*0.787*	*1.875*	*0.178*	*0.481*	*1.585*	*1.781*	*0.227*	*0.517*	*0.799*	*2.169*	*0.194*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	**0.580**	1.191	**0.448**	0.127	0.566	0.739	1.793	0.183
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.552	1.214	0.726	**0.121**	0.569	0.722	1.703	0.175
Mimo-v2.5	–	–	–	–	–	–	–	–	0.564	1.210	0.815	0.161	0.542	0.764	1.721	0.191
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.573	**1.157**	0.562	**0.116**	0.547	0.745	**1.463**	0.174
LongCat-2.0	–	–	–	–	–	–	–	–	0.534	1.235	**0.404**	0.127	0.545	0.746	1.516	0.183
GPT-4o-mini	–	–	–	–	–	–	–	–	0.509	1.492	1.487	0.243	0.526	0.805	2.170	0.227
GPT-5-mini	–	–	–	–	–	–	–	–	0.496	1.296	1.349	0.200	0.531	0.759	2.193	0.179
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.524	1.456	0.773	0.205	0.517	0.771	1.809	0.205
Qwen3-32B	–	–	–	–	–	–	–	–	0.532	1.271	1.023	0.159	0.527	0.789	1.706	0.200
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.458	1.460	1.393	0.176	0.499	0.815	1.895	0.224
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	**0.578**	**1.153**	1.116	0.164	0.535	0.767	1.785	0.188
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.480	1.476	0.496	0.182	0.519	0.771	1.612	0.216
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.542	1.317	1.509	0.179	0.517	0.842	1.744	0.191
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.532*	*1.302*	*0.931*	*0.166*	*0.534*	*0.772*	*1.778*	*0.195*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.573	0.780	2.923	0.182	–	–	–	–	0.547	0.773	3.455	0.191
DeepSeek-V4-Flash	–	–	–	–	0.549	0.835	2.394	0.175	–	–	–	–	0.531	0.846	3.117	0.191
Mimo-v2.5	–	–	–	–	0.573	0.775	2.230	0.178	–	–	–	–	0.541	0.761	2.109	0.203
Mimo-v2.5-Pro	–	–	–	–	0.575	0.767	2.247	0.178	–	–	–	–	0.566	0.736	2.567	0.187
LongCat-2.0	–	–	–	–	**0.587**	0.764	2.931	0.194	–	–	–	–	0.555	0.777	3.388	0.221
GPT-4o-mini	–	–	–	–	0.551	0.792	3.091	0.198	–	–	–	–	0.539	0.785	3.253	0.210
GPT-5-mini	–	–	–	–	0.521	0.833	2.587	0.176	–	–	–	–	0.506	0.854	3.210	0.200
Qwen2.5-72B-Instruct	–	–	–	–	0.561	**0.749**	2.469	0.181	–	–	–	–	0.531	0.800	3.099	0.219
Qwen3-32B	–	–	–	–	0.534	0.833	2.250	0.192	–	–	–	–	0.519	0.823	2.633	0.204
Qwen3-30B-A3B	–	–	–	–	0.476	0.946	2.122	0.191	–	–	–	–	0.473	0.910	1.956	0.197
Llama-3.3-70B-Instruct	–	–	–	–	0.553	0.776	2.129	0.176	–	–	–	–	0.523	0.827	2.908	0.199
Gemma-3-27B-it	–	–	–	–	0.524	0.852	1.994	0.192	–	–	–	–	0.485	0.888	1.809	0.208
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.567	0.781	2.719	0.191	–	–	–	–	0.551	0.780	3.184	0.214
Avg. (PRA)	–	–	–	–	*0.550*	*0.806*	*2.468*	*0.185*	–	–	–	–	*0.528*	*0.812*	*2.822*	*0.204*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.557	**0.710**	2.375	0.206
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.559	0.741	1.860	0.180
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.555	0.725	1.934	0.190
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	**0.579**	**0.702**	1.715	**0.171**
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.549	0.761	2.152	0.206
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.535	0.809	2.288	0.222
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.511	0.826	2.708	0.199
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.542	0.775	2.558	0.209
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.528	0.801	**1.313**	0.182
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.471	0.877	1.691	0.201
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.558	0.716	1.937	**0.167**
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.521	0.802	1.884	0.238
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.537	0.805	2.217	0.206
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.539*	*0.773*	*2.048*	*0.198*
Training																
BERT	0.526	1.813	5.114	0.297	**0.587**	0.796	5.250	0.260	0.514	1.506	1.801	0.218	**0.576**	0.727	4.299	0.219
SFT (Qwen2.5-7B-Instruct)	0.533	1.756	2.861	0.252	**0.579**	0.770	3.172	0.182	0.522	1.613	3.067	0.276	0.552	0.769	3.804	0.204
SFT (Qwen3-8B)	0.551	1.459	2.584	0.221	0.559	0.814	1.964	0.172	0.526	1.403	2.600	0.217	0.517	0.852	2.946	0.207

Table 19: ANES results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.411	0.916	12.527	0.344	0.453	0.891	11.599	0.326	0.420	0.935	12.279	0.339	0.448	0.899	11.721	0.329
DeepSeek-V4-Flash	0.420	0.935	12.319	0.335	0.458	0.917	11.485	0.326	0.418	0.965	12.323	0.347	0.441	0.952	11.876	0.343
Mimo-v2.5	0.377	1.016	13.326	0.379	0.425	0.955	12.262	0.348	0.393	1.030	12.909	0.376	0.435	0.939	12.017	0.338
Mimo-v2.5-Pro	0.372	1.018	13.430	0.383	0.420	0.972	12.378	0.359	0.387	1.040	13.033	0.380	0.428	0.979	12.174	0.356
LongCat-2.0	0.376	1.021	13.350	0.379	0.438	0.944	11.960	0.341	0.392	1.045	12.926	0.380	0.432	0.949	12.086	0.340
GPT-4o-mini	0.406	0.939	12.655	0.348	0.407	0.994	12.677	0.365	0.393	1.021	12.909	0.361	0.405	1.022	12.717	0.370
GPT-5-mini	0.400	0.963	12.791	0.360	0.437	0.921	11.968	0.335	0.391	0.994	12.948	0.365	0.427	0.952	12.203	0.342
Qwen2.5-72B-Instruct	0.387	0.984	13.094	0.368	0.421	0.936	12.344	0.342	0.383	1.000	13.130	0.372	0.417	0.950	12.438	0.344
Qwen3-32B	0.394	1.015	12.934	0.369	0.433	0.938	12.064	0.340	0.389	1.073	12.992	0.382	0.431	0.950	12.118	0.341
Qwen3-30B-A3B	0.365	1.007	13.597	0.375	0.377	0.986	13.358	0.378	0.366	1.049	13.532	0.378	0.381	1.013	13.258	0.376
Llama-3.3-70B-Instruct	0.400	0.973	12.783	0.361	0.440	0.945	11.918	0.340	0.394	0.994	12.873	0.371	0.415	0.994	12.488	0.365
Gemma-3-27B-it	0.374	0.968	13.390	0.362	0.423	0.927	12.304	0.339	0.388	0.965	13.024	0.357	0.413	0.945	12.518	0.346
Mistral-Small-3.1-24B-Instruct	0.402	0.957	12.743	0.352	0.431	0.926	12.122	0.338	0.397	1.000	12.805	0.362	0.429	0.920	12.166	0.336
Avg. (Zero-shot)	*0.391*	*0.978*	*12.995*	*0.363*	*0.428*	*0.942*	*12.188*	*0.344*	*0.393*	*1.008*	*12.899*	*0.367*	*0.423*	*0.959*	*12.291*	*0.348*
Persona																
DeepSeek-V4-Pro	**0.470**	0.801	2.659	0.180	0.487	0.812	2.335	0.166	0.472	0.841	1.888	0.170	0.491	0.818	2.454	0.165
DeepSeek-V4-Flash	0.468	0.802	1.790	**0.169**	0.493	0.818	1.924	0.159	0.468	0.830	1.202	0.156	0.495	0.822	2.116	0.162
Mimo-v2.5	**0.470**	**0.795**	2.247	0.175	0.486	0.808	1.889	0.164	0.456	0.849	1.543	0.170	0.486	0.808	2.290	0.166
Mimo-v2.5-Pro	0.468	0.800	1.894	0.174	0.487	0.805	1.803	0.164	0.466	0.835	1.059	0.160	0.495	0.799	2.149	0.169
LongCat-2.0	0.465	0.800	2.093	**0.166**	0.484	0.807	1.968	0.162	0.468	0.827	1.702	0.158	0.491	0.797	2.314	0.164
GPT-4o-mini	0.439	0.869	**1.175**	0.173	0.456	0.873	**1.045**	0.153	0.429	0.895	**0.612**	0.152	0.452	0.873	**1.296**	0.158
GPT-5-mini	0.439	0.820	1.895	0.174	0.470	0.821	2.041	0.165	0.443	0.842	1.134	0.163	0.472	0.822	2.308	0.168
Qwen2.5-72B-Instruct	0.438	0.826	1.641	0.188	0.466	0.818	1.653	0.169	0.438	0.853	1.032	0.173	0.467	0.819	1.773	0.164
Qwen3-32B	0.451	0.867	2.160	0.195	0.470	0.857	2.200	0.166	0.445	0.887	1.790	0.183	0.461	0.887	2.326	0.175
Qwen3-30B-A3B	0.414	0.930	1.477	0.207	0.435	0.922	1.864	0.192	0.411	0.947	1.171	0.195	0.438	0.912	1.908	0.186
Llama-3.3-70B-Instruct	**0.470**	0.797	2.867	0.193	0.493	0.785	2.529	0.162	0.458	0.835	2.213	0.186	0.486	0.802	2.746	0.172
Gemma-3-27B-it	0.417	0.854	**1.332**	0.173	0.436	0.882	1.443	0.177	0.422	0.875	1.068	0.168	0.445	0.863	1.661	0.176
Mistral-Small-3.1-24B-Instruct	0.434	0.862	1.930	0.198	0.460	0.875	1.633	0.185	0.433	0.901	1.329	0.182	0.460	0.882	1.861	0.179
Avg. (Persona)	*0.449*	*0.832*	*1.935*	*0.182*	*0.471*	*0.837*	*1.871*	*0.168*	*0.447*	*0.863*	*1.365*	*0.171*	*0.472*	*0.839*	*2.092*	*0.169*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.493	0.802	1.650	0.157	0.518	0.771	2.324	0.152
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.481	0.826	1.007	0.154	0.502	0.810	2.070	0.166
Mimo-v2.5	–	–	–	–	–	–	–	–	0.485	0.806	1.209	0.158	0.520	0.760	2.220	0.156
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.492	0.793	0.814	0.154	0.515	0.766	2.066	0.163
LongCat-2.0	–	–	–	–	–	–	–	–	0.492	0.807	1.627	0.163	0.512	0.771	2.598	0.169
GPT-4o-mini	–	–	–	–	–	–	–	–	0.434	0.919	**0.428**	0.164	0.474	0.863	1.672	0.174
GPT-5-mini	–	–	–	–	–	–	–	–	0.454	0.833	0.895	0.162	0.489	0.798	2.146	0.166
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.466	0.812	1.050	0.155	0.497	0.774	1.889	0.161
Qwen3-32B	–	–	–	–	–	–	–	–	0.457	0.871	1.556	0.166	0.484	0.836	2.228	0.172
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.423	0.920	0.637	0.191	0.449	0.890	1.807	0.194
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.479	0.804	1.612	0.165	0.508	0.776	2.134	0.157
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.451	0.836	0.855	0.160	0.469	0.831	1.336	0.173
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.449	0.892	1.219	0.182	0.481	0.860	1.966	0.182
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.466*	*0.840*	*1.120*	*0.164*	*0.494*	*0.808*	*2.035*	*0.168*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.517	0.778	3.089	0.175	–	–	–	–	0.513	0.790	3.564	0.178
DeepSeek-V4-Flash	–	–	–	–	0.505	0.814	2.778	0.168	–	–	–	–	0.521	0.790	2.900	0.173
Mimo-v2.5	–	–	–	–	0.515	0.782	2.202	0.162	–	–	–	–	0.521	0.786	2.556	0.163
Mimo-v2.5-Pro	–	–	–	–	0.513	0.790	2.468	0.155	–	–	–	–	0.521	0.777	2.651	0.159
LongCat-2.0	–	–	–	–	0.518	0.777	3.235	0.169	–	–	–	–	0.522	0.772	3.534	0.171
GPT-4o-mini	–	–	–	–	0.506	0.805	2.445	0.168	–	–	–	–	0.510	0.797	2.628	0.174
GPT-5-mini	–	–	–	–	0.510	0.786	2.944	0.184	–	–	–	–	0.522	0.781	3.310	0.176
Qwen2.5-72B-Instruct	–	–	–	–	0.504	0.788	2.894	0.166	–	–	–	–	0.509	0.775	3.179	0.169
Qwen3-32B	–	–	–	–	0.497	0.837	2.603	0.176	–	–	–	–	0.491	0.851	2.879	0.181
Qwen3-30B-A3B	–	–	–	–	0.437	0.961	1.309	0.160	–	–	–	–	0.440	0.956	1.503	0.167
Llama-3.3-70B-Instruct	–	–	–	–	0.512	0.786	3.039	0.182	–	–	–	–	0.509	0.804	3.429	0.191
Gemma-3-27B-it	–	–	–	–	0.448	0.903	**1.255**	0.151	–	–	–	–	0.450	0.914	1.438	0.158
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.508	0.814	3.565	0.194	–	–	–	–	0.516	0.815	3.957	0.196
Avg. (PRA)	–	–	–	–	*0.499*	*0.817*	*2.602*	*0.170*	–	–	–	–	*0.503*	*0.816*	*2.887*	*0.174*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.541	0.729	2.949	0.154
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.526	0.783	2.617	0.161
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.538	0.739	2.459	0.159
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.543	0.746	2.388	0.154
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.538	0.742	3.096	0.161
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.508	0.812	1.869	0.160
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.526	0.760	2.947	0.165
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.525	0.747	2.651	0.163
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.504	0.828	2.138	0.165
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.447	0.941	**1.140**	0.165
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.515	0.779	2.790	0.174
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.482	0.845	1.597	0.163
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.514	0.811	3.217	0.186
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.516*	*0.790*	*2.451*	*0.164*
Training																
BERT	0.391	0.963	4.102	0.237	0.529	0.754	4.963	0.191	0.423	0.896	2.234	0.218	0.582	0.643	3.089	0.142
SFT (Qwen2.5-7B-Instruct)	**0.496**	**0.757**	2.665	0.175	**0.547**	**0.705**	2.620	**0.138**	**0.518**	**0.715**	1.498	**0.135**	**0.593**	**0.607**	1.805	**0.115**
SFT (Qwen3-8B)	0.469	0.809	3.330	0.194	**0.539**	**0.710**	2.163	**0.125**	**0.505**	**0.737**	1.518	**0.150**	**0.584**	**0.625**	2.189	**0.122**

Table 20: Arab Barometer results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.372	1.129	13.294	0.337	0.362	1.067	13.528	0.342	0.344	1.142	13.876	0.344	0.363	1.026	13.553	0.327
DeepSeek-V4-Flash	0.370	1.020	13.337	0.331	0.363	1.081	13.511	0.346	0.359	1.166	13.537	0.351	0.363	1.049	13.560	0.336
Mimo-v2.5	0.345	1.255	13.916	0.385	0.348	1.077	13.861	0.348	0.316	1.225	14.528	0.369	0.331	1.090	14.300	0.355
Mimo-v2.5-Pro	0.321	1.184	14.452	0.359	0.347	1.141	13.883	0.370	0.303	1.302	14.814	0.396	0.339	1.078	14.101	0.354
LongCat-2.0	0.362	1.082	13.508	0.325	0.343	1.094	13.972	0.357	0.338	1.146	14.024	0.343	0.339	1.070	14.115	0.346
GPT-4o-mini	0.322	1.176	14.430	0.351	0.326	1.157	14.350	0.376	0.338	1.143	14.019	0.340	0.321	1.136	14.526	0.364
GPT-5-mini	0.366	1.079	13.423	0.320	0.345	1.053	13.928	0.339	0.349	1.119	13.775	0.342	0.354	1.030	13.772	0.330
Qwen2.5-72B-Instruct	0.291	1.225	15.138	0.372	0.350	1.093	13.817	0.344	0.344	1.140	13.871	0.341	0.347	1.034	13.930	0.327
Qwen3-32B	0.314	1.192	14.623	0.358	0.344	1.128	13.950	0.362	0.319	1.246	14.464	0.378	0.351	1.109	13.834	0.352
Qwen3-30B-A3B	0.321	1.172	14.452	0.349	0.327	1.125	14.344	0.366	0.295	1.187	15.005	0.368	0.331	1.079	14.286	0.352
Llama-3.3-70B-Instruct	0.339	1.248	14.044	0.375	0.348	1.109	13.850	0.363	0.350	1.167	13.733	0.351	0.340	1.102	14.087	0.361
Gemma-3-27B-it	0.300	1.153	14.945	0.350	0.336	1.075	14.139	0.343	0.339	1.159	13.993	0.352	0.354	1.010	13.772	0.323
Mistral-Small-3.1-24B-Instruct	0.291	1.126	15.159	0.335	0.340	1.090	14.039	0.345	0.301	1.191	14.867	0.358	0.351	1.007	13.834	0.325
Avg. (Zero-shot)	*0.332*	*1.157*	*14.209*	*0.350*	*0.344*	*1.099*	*13.936*	*0.354*	*0.330*	*1.179*	*14.193*	*0.356*	*0.345*	*1.063*	*13.974*	*0.342*
Persona																
DeepSeek-V4-Pro	0.350	1.042	4.824	0.229	0.372	1.019	4.519	0.203	0.377	0.974	2.761	0.190	0.383	0.979	4.672	0.195
DeepSeek-V4-Flash	**0.376**	**0.985**	3.278	0.203	0.372	1.021	3.785	0.193	0.374	0.966	1.974	0.188	0.385	0.959	4.444	0.202
Mimo-v2.5	0.325	1.051	3.988	0.224	0.371	1.017	3.963	0.197	0.346	1.009	2.156	0.198	0.377	0.987	4.350	0.198
Mimo-v2.5-Pro	0.337	1.036	4.161	0.209	0.362	1.023	4.041	0.196	0.363	0.994	2.492	0.193	0.361	0.987	4.304	0.198
LongCat-2.0	0.334	1.034	3.763	0.205	0.366	1.022	4.022	0.204	0.369	0.973	2.538	0.184	0.381	0.968	4.599	0.203
GPT-4o-mini	0.318	1.085	**2.838**	0.194	0.318	1.141	3.217	0.207	0.347	1.036	1.612	0.175	0.344	1.068	3.638	0.206
GPT-5-mini	0.357	**0.996**	4.047	0.195	0.361	1.042	4.311	0.211	0.370	0.974	2.615	0.183	0.376	0.979	4.618	0.202
Qwen2.5-72B-Instruct	0.336	1.027	3.656	0.200	0.333	1.050	3.281	0.191	0.344	1.004	1.981	0.172	0.355	0.982	3.252	0.181
Qwen3-32B	0.323	1.047	4.119	0.211	0.349	1.068	4.401	0.216	0.346	1.027	2.573	0.200	0.366	1.025	4.247	0.213
Qwen3-30B-A3B	0.311	1.064	3.960	0.199	0.319	1.086	3.260	0.193	0.336	1.052	2.247	0.186	0.336	1.027	3.809	0.202
Llama-3.3-70B-Instruct	0.344	1.036	4.043	**0.188**	0.366	1.019	4.638	0.206	0.349	1.017	2.661	0.174	0.374	0.984	4.744	0.218
Gemma-3-27B-it	0.307	1.058	4.916	0.251	0.314	1.082	4.022	0.212	0.330	1.026	2.852	0.225	0.329	1.032	4.339	0.216
Mistral-Small-3.1-24B-Instruct	0.322	1.084	4.257	0.215	0.330	1.088	4.335	0.210	0.344	1.026	2.483	0.202	0.355	1.023	4.420	0.210
Avg. (Persona)	*0.334*	*1.042*	*3.989*	*0.210*	*0.349*	*1.052*	*3.984*	*0.203*	*0.353*	*1.006*	*2.380*	*0.190*	*0.363*	*1.000*	*4.264*	*0.203*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.407	0.944	2.142	0.188	0.405	0.994	3.593	0.196
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.405	0.934	**1.607**	**0.159**	0.407	0.962	3.596	0.182
Mimo-v2.5	–	–	–	–	–	–	–	–	0.388	0.956	2.113	0.196	0.399	0.970	3.688	0.199
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.405	**0.931**	2.152	0.174	0.401	0.970	3.985	0.190
LongCat-2.0	–	–	–	–	–	–	–	–	**0.410**	**0.917**	1.983	**0.164**	0.422	0.929	3.689	0.194
GPT-4o-mini	–	–	–	–	–	–	–	–	0.361	0.992	**1.399**	0.187	0.364	1.029	3.251	0.215
GPT-5-mini	–	–	–	–	–	–	–	–	0.383	0.940	2.195	0.183	0.386	0.953	4.223	0.202
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.374	0.959	2.104	0.185	0.379	0.936	3.461	0.185
Qwen3-32B	–	–	–	–	–	–	–	–	0.364	1.001	2.359	0.211	0.390	0.994	3.649	0.208
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.339	1.069	1.947	0.191	0.340	1.078	3.347	0.229
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.377	0.950	2.321	0.183	0.390	0.974	4.550	0.212
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.348	0.997	2.277	0.203	0.354	1.020	3.865	0.209
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.352	1.090	2.794	0.205	0.370	1.048	4.302	0.222
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.378*	*0.975*	*2.107*	*0.187*	*0.385*	*0.989*	*3.785*	*0.203*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.407	0.997	5.368	0.226	–	–	–	–	0.408	0.973	5.301	0.217
DeepSeek-V4-Flash	–	–	–	–	0.404	1.008	4.090	0.200	–	–	–	–	0.404	0.979	4.042	0.192
Mimo-v2.5	–	–	–	–	0.395	1.023	3.708	**0.188**	–	–	–	–	0.397	0.992	3.751	0.186
Mimo-v2.5-Pro	–	–	–	–	0.396	0.995	4.045	0.193	–	–	–	–	0.405	0.947	4.463	0.193
LongCat-2.0	–	–	–	–	0.403	1.014	5.616	0.235	–	–	–	–	0.419	0.951	5.966	0.227
GPT-4o-mini	–	–	–	–	0.382	1.029	3.577	**0.188**	–	–	–	–	0.401	0.967	4.232	0.187
GPT-5-mini	–	–	–	–	0.399	1.000	5.317	0.228	–	–	–	–	0.415	0.946	5.256	0.213
Qwen2.5-72B-Instruct	–	–	–	–	0.374	1.013	4.269	**0.188**	–	–	–	–	0.390	0.949	4.503	0.188
Qwen3-32B	–	–	–	–	0.385	1.046	4.200	0.217	–	–	–	–	0.382	1.023	4.615	0.216
Qwen3-30B-A3B	–	–	–	–	0.362	1.093	**1.586**	**0.153**	–	–	–	–	0.350	1.081	**1.987**	**0.160**
Llama-3.3-70B-Instruct	–	–	–	–	0.392	1.023	4.919	0.217	–	–	–	–	0.398	0.974	5.453	0.222
Gemma-3-27B-it	–	–	–	–	0.363	1.078	**2.791**	0.190	–	–	–	–	0.362	1.039	3.131	0.193
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.396	1.053	4.825	0.231	–	–	–	–	0.404	0.996	5.448	0.223
Avg. (PRA)	–	–	–	–	*0.389*	*1.029*	*4.178*	*0.204*	–	–	–	–	*0.395*	*0.986*	*4.473*	*0.201*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.427	0.920	4.720	0.204
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.414	0.960	4.291	0.194
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.418	0.965	3.642	0.202
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.419	0.927	4.156	0.191
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	**0.434**	0.941	4.699	0.210
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.401	0.968	3.499	0.193
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.409	0.956	4.717	0.201
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.409	**0.913**	4.007	0.185
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.394	1.019	3.656	0.204
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.341	1.105	**2.113**	0.189
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.406	0.954	4.609	0.213
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.369	1.013	3.586	0.213
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.402	1.018	4.885	0.223
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.403*	*0.974*	*4.044*	*0.202*
Training																
BERT	**0.378**	1.075	4.438	0.246	**0.440**	**0.957**	4.216	0.205	**0.416**	0.985	1.786	0.172	**0.491**	**0.820**	3.327	**0.170**
SFT (Qwen2.5-7B-Instruct)	0.356	1.105	4.389	0.193	**0.414**	**0.982**	4.508	0.200	0.379	1.114	3.021	0.199	0.425	0.918	4.564	0.192
SFT (Qwen3-8B)	0.342	1.128	**3.016**	**0.170**	0.408	0.985	4.461	0.205	0.377	1.061	2.601	0.174	0.415	0.940	4.944	0.203

Table 21: Asian Barometer results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.328	1.199	14.279	0.301	0.456	0.821	11.492	0.274	0.286	1.433	15.188	0.377	0.407	0.897	12.588	0.292
DeepSeek-V4-Flash	0.320	1.352	14.459	0.321	0.418	0.916	12.365	0.295	0.231	1.366	16.445	0.442	0.403	0.931	12.694	0.300
Mimo-v2.5	0.305	1.332	14.819	0.335	0.373	1.030	13.413	0.339	**0.348**	1.222	13.770	0.304	0.405	0.994	12.641	0.323
Mimo-v2.5-Pro	0.137	1.617	18.687	0.487	0.387	0.977	13.085	0.322	0.259	1.560	15.805	0.383	0.392	0.985	12.934	0.320
LongCat-2.0	0.188	1.508	17.517	0.462	0.430	0.924	12.082	0.303	0.269	1.356	15.577	0.350	0.427	0.865	12.135	0.283
GPT-4o-mini	0.137	1.637	18.687	0.498	0.367	0.961	13.544	0.319	0.264	1.311	15.691	0.393	0.373	0.985	13.387	0.322
GPT-5-mini	0.301	1.172	14.909	0.312	0.422	0.892	12.278	0.292	0.236	1.545	16.331	0.419	0.434	0.867	11.975	0.288
Qwen2.5-72B-Instruct	0.328	1.219	14.279	0.317	0.400	0.918	12.780	0.303	0.259	1.560	15.805	0.383	0.378	0.990	13.254	0.323
Qwen3-32B	0.320	1.375	14.459	0.329	0.407	0.933	12.627	0.303	0.311	1.322	14.616	0.332	0.409	0.941	12.561	0.302
Qwen3-30B-A3B	0.176	1.551	17.787	0.470	0.309	1.036	14.875	0.346	0.133	1.728	18.709	0.509	0.287	1.120	15.359	0.364
Llama-3.3-70B-Instruct	0.172	1.273	17.877	0.435	0.417	0.955	12.387	0.310	0.219	1.324	16.720	0.404	0.421	0.895	12.268	0.292
Gemma-3-27B-it	0.328	**0.949**	14.279	0.253	0.442	0.800	11.820	0.266	0.227	1.339	16.537	0.395	0.383	0.904	13.147	0.291
Mistral-Small-3.1-24B-Instruct	**0.359**	1.125	13.560	0.278	0.415	0.912	12.431	0.301	0.228	1.522	16.514	0.399	0.427	0.848	12.135	0.277
Avg. (Zero-shot)	*0.261*	*1.331*	*15.815*	*0.369*	*0.403*	*0.929*	*12.706*	*0.306*	*0.252*	*1.430*	*15.978*	*0.391*	*0.396*	*0.940*	*12.852*	*0.306*
Persona																
DeepSeek-V4-Pro	0.324	1.207	1.444	0.204	0.427	0.900	2.308	0.173	0.289	1.317	0.699	0.193	0.411	0.928	2.750	0.187
DeepSeek-V4-Flash	0.301	1.355	2.912	0.247	0.436	0.892	2.064	0.149	0.313	1.352	1.355	0.241	0.434	0.942	2.611	0.156
Mimo-v2.5	0.324	1.168	1.732	0.181	0.418	0.916	1.700	0.154	0.282	1.279	0.828	0.175	0.397	1.009	2.834	0.169
Mimo-v2.5-Pro	0.340	1.285	2.412	0.203	0.424	0.916	2.074	0.156	0.288	1.381	0.838	0.198	0.418	0.978	2.827	0.163
LongCat-2.0	0.336	1.184	1.412	0.182	0.438	0.875	1.795	0.147	0.290	1.255	1.540	0.174	0.397	0.934	2.261	0.154
GPT-4o-mini	0.293	1.297	1.042	**0.147**	0.387	0.954	1.011	0.150	0.284	1.304	0.324	**0.127**	0.356	1.056	1.150	0.173
GPT-5-mini	0.223	1.289	2.466	0.230	0.423	0.836	2.118	**0.129**	0.244	1.254	1.046	0.208	0.409	0.894	3.119	0.165
Qwen2.5-72B-Instruct	0.328	1.207	0.913	0.184	0.379	0.963	1.567	0.174	0.284	1.332	1.656	0.223	0.388	1.029	3.002	0.191
Qwen3-32B	0.305	1.172	**0.412**	0.151	0.363	0.991	2.463	0.189	0.310	1.225	0.340	**0.128**	0.360	1.062	2.349	0.187
Qwen3-30B-A3B	0.312	1.363	1.113	0.163	0.366	0.991	1.197	0.173	0.295	1.342	0.533	0.157	0.344	1.065	2.329	0.181
Llama-3.3-70B-Instruct	0.250	1.305	3.168	0.278	0.445	0.895	2.525	0.157	0.246	1.360	2.511	0.259	0.417	0.962	3.188	0.184
Gemma-3-27B-it	0.230	1.320	0.833	0.221	0.392	0.921	1.483	0.155	0.242	1.344	0.536	0.200	0.399	0.946	2.501	0.171
Mistral-Small-3.1-24B-Instruct	0.309	1.348	2.157	0.166	0.357	1.080	1.944	0.233	0.310	1.359	1.359	0.192	0.350	1.147	3.741	0.252
Avg. (Persona)	*0.298*	*1.269*	*1.694*	*0.197*	*0.404*	*0.933*	*1.865*	*0.165*	*0.283*	*1.316*	*1.044*	*0.190*	*0.391*	*0.996*	*2.666*	*0.180*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.308	1.244	0.706	0.216	0.475	0.846	1.384	0.164
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.300	1.346	1.622	0.264	0.477	0.851	1.140	0.151
Mimo-v2.5	–	–	–	–	–	–	–	–	0.319	1.215	0.555	0.171	0.480	0.824	1.016	**0.143**
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.294	1.299	0.510	0.214	0.494	0.811	1.505	0.148
LongCat-2.0	–	–	–	–	–	–	–	–	0.323	1.224	1.480	0.188	0.470	0.809	1.624	0.159
GPT-4o-mini	–	–	–	–	–	–	–	–	0.302	1.296	0.343	0.133	0.428	0.889	1.478	0.178
GPT-5-mini	–	–	–	–	–	–	–	–	0.293	1.215	0.615	0.201	0.459	0.775	1.496	**0.143**
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.288	1.284	1.129	0.227	0.434	0.858	1.776	0.155
Qwen3-32B	–	–	–	–	–	–	–	–	0.339	**1.131**	**0.253**	0.139	0.440	0.870	0.986	0.154
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.311	1.251	0.645	0.153	0.365	0.999	1.517	0.206
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.278	1.256	2.527	0.217	0.449	0.867	2.065	0.166
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.287	1.260	0.501	0.188	0.434	0.898	1.512	0.168
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.319	1.323	1.097	0.196	0.395	0.991	1.581	0.202
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.305*	*1.257*	*0.922*	*0.193*	*0.446*	*0.868*	*1.468*	*0.165*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.500	0.711	3.342	0.146	–	–	–	–	0.480	0.779	4.224	0.168
DeepSeek-V4-Flash	–	–	–	–	0.451	0.822	2.174	0.156	–	–	–	–	0.446	0.897	2.299	0.161
Mimo-v2.5	–	–	–	–	0.462	0.827	2.055	0.162	–	–	–	–	0.435	0.895	2.738	0.165
Mimo-v2.5-Pro	–	–	–	–	0.475	0.776	2.489	0.152	–	–	–	–	0.465	0.803	3.090	0.152
LongCat-2.0	–	–	–	–	**0.501**	**0.710**	4.521	0.176	–	–	–	–	0.497	0.749	5.383	0.179
GPT-4o-mini	–	–	–	–	0.484	0.792	2.916	0.158	–	–	–	–	0.446	0.904	2.863	0.164
GPT-5-mini	–	–	–	–	0.486	0.731	3.175	0.161	–	–	–	–	0.498	0.753	4.176	0.171
Qwen2.5-72B-Instruct	–	–	–	–	0.481	0.746	2.747	0.138	–	–	–	–	0.440	0.894	3.728	0.160
Qwen3-32B	–	–	–	–	0.430	0.811	2.092	**0.133**	–	–	–	–	0.404	0.928	2.223	0.154
Qwen3-30B-A3B	–	–	–	–	0.348	1.040	**0.661**	0.146	–	–	–	–	0.352	1.064	**0.446**	**0.136**
Llama-3.3-70B-Instruct	–	–	–	–	0.460	0.798	3.396	0.145	–	–	–	–	0.455	0.847	3.289	0.145
Gemma-3-27B-it	–	–	–	–	0.355	0.995	**0.685**	0.147	–	–	–	–	0.374	1.049	0.952	0.161
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.449	0.862	2.840	0.158	–	–	–	–	0.450	0.938	3.980	0.182
Avg. (PRA)	–	–	–	–	*0.452*	*0.817*	*2.546*	*0.152*	–	–	–	–	*0.442*	*0.885*	*3.030*	*0.161*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	**0.515**	**0.697**	3.274	0.164
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.473	0.812	2.140	0.155
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.481	0.782	1.200	0.147
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.505	0.738	1.930	0.149
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.514	0.740	3.044	0.168
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.468	0.858	1.794	0.157
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.473	0.791	2.540	0.155
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.472	0.795	3.283	0.168
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.450	0.851	1.646	0.148
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.363	1.028	**0.898**	0.153
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.468	0.796	2.442	0.148
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.414	0.948	1.143	0.170
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.455	0.885	2.228	0.173
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.466*	*0.825*	*2.120*	*0.158*
Training																
BERT	**0.418**	1.305	12.211	0.259	**0.536**	**0.666**	4.416	0.175	**0.413**	1.307	8.056	0.269	**0.520**	**0.708**	5.380	0.190
SFT (Qwen2.5-7B-Instruct)	0.273	1.262	1.795	0.196	0.497	0.719	3.513	0.159	0.272	1.276	1.043	0.213	0.464	0.765	4.704	0.158
SFT (Qwen3-8B)	0.324	**1.102**	**0.715**	**0.131**	0.430	0.811	2.934	0.148	0.324	**1.172**	**0.218**	0.131	0.429	0.807	3.639	0.149

Table 22: ATP results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.347	0.839	14.014	0.376	0.339	0.941	14.055	0.348	0.421	0.694	12.230	0.305	0.341	0.868	14.075	0.335
DeepSeek-V4-Flash	0.457	0.754	11.469	0.294	0.282	1.124	15.358	0.408	0.390	0.851	12.949	0.321	0.320	0.963	14.550	0.359
Mimo-v2.5	0.387	0.945	13.089	0.373	0.335	0.931	14.127	0.353	0.338	0.894	14.127	0.358	0.320	0.948	14.550	0.357
Mimo-v2.5-Pro	0.312	0.935	14.824	0.423	0.321	1.002	14.465	0.368	0.335	0.959	14.213	0.374	0.341	0.947	14.075	0.359
LongCat-2.0	0.347	0.839	14.014	0.376	0.332	0.987	14.199	0.366	0.338	0.894	14.127	0.358	0.395	0.835	12.827	0.294
GPT-4o-mini	0.271	1.407	15.750	0.487	0.285	1.114	15.285	0.414	0.243	1.231	16.312	0.453	0.319	1.010	14.580	0.381
GPT-5-mini	0.347	0.839	14.014	0.376	0.313	0.990	14.634	0.373	0.393	0.787	12.862	0.305	0.363	0.890	13.570	0.334
Qwen2.5-72B-Instruct	0.327	0.955	14.477	0.430	0.320	0.992	14.489	0.366	0.417	0.784	12.316	0.329	0.346	0.854	13.956	0.330
Qwen3-32B	0.236	1.201	16.560	0.476	0.301	1.198	14.923	0.408	0.356	0.980	13.725	0.364	0.316	1.052	14.639	0.374
Qwen3-30B-A3B	0.231	1.302	16.676	0.490	0.301	1.097	14.923	0.399	0.323	1.114	14.472	0.386	0.293	1.055	15.174	0.379
Llama-3.3-70B-Instruct	0.407	0.965	12.626	0.364	0.286	1.090	15.261	0.407	0.356	0.980	13.725	0.364	0.289	1.039	15.263	0.409
Gemma-3-27B-it	0.347	0.940	14.014	0.399	0.321	0.976	14.465	0.371	0.287	0.965	15.306	0.411	0.364	0.849	13.540	0.325
Mistral-Small-3.1-24B-Instruct	0.312	1.035	14.824	0.447	0.344	0.892	13.934	0.345	0.378	0.884	13.207	0.355	0.326	0.907	14.402	0.349
Avg. (Zero-shot)	*0.333*	*0.997*	*14.335*	*0.408*	*0.314*	*1.026*	*14.625*	*0.379*	*0.352*	*0.924*	*13.813*	*0.360*	*0.333*	*0.940*	*14.246*	*0.353*
Persona																
DeepSeek-V4-Pro	0.442	0.724	1.956	0.111	0.420	0.835	3.082	0.154	0.442	0.705	1.315	0.121	0.418	0.832	2.999	0.153
DeepSeek-V4-Flash	0.487	0.698	3.196	0.166	0.420	0.836	2.863	0.185	0.412	0.774	1.603	0.171	0.386	0.888	2.669	0.178
Mimo-v2.5	0.452	0.754	1.620	0.124	0.417	0.878	2.650	0.192	0.424	0.792	0.788	0.152	0.363	0.933	2.815	0.187
Mimo-v2.5-Pro	0.447	0.759	1.891	0.121	0.415	0.834	2.015	0.144	0.442	0.773	1.168	0.144	0.410	0.857	2.244	0.151
LongCat-2.0	0.487	0.719	3.384	0.165	0.435	0.824	2.794	0.168	0.458	0.725	2.189	0.141	0.414	0.845	3.791	0.175
GPT-4o-mini	0.362	0.955	3.904	0.246	0.374	0.956	2.177	0.197	0.351	0.938	2.868	0.209	0.368	0.943	2.946	0.195
GPT-5-mini	0.392	0.759	3.055	0.179	0.408	0.840	2.390	0.173	0.400	0.767	1.838	0.182	0.391	0.850	3.433	0.183
Qwen2.5-72B-Instruct	0.387	0.804	3.214	0.215	0.384	0.904	2.217	0.156	0.429	0.737	1.920	0.167	0.394	0.859	2.809	0.158
Qwen3-32B	0.437	0.839	5.253	0.218	0.348	1.058	3.097	0.236	0.365	0.904	2.535	0.238	0.355	1.015	3.297	0.251
Qwen3-30B-A3B	0.377	0.945	3.583	0.210	0.386	0.988	1.476	0.161	0.361	0.918	1.835	0.206	0.341	1.049	1.433	0.201
Llama-3.3-70B-Instruct	0.457	0.724	4.266	0.168	0.389	0.884	3.386	0.216	0.447	0.714	3.019	0.175	0.361	0.924	4.612	0.224
Gemma-3-27B-it	0.432	0.739	2.779	0.147	0.419	0.853	2.480	0.156	0.400	0.778	1.885	0.159	0.404	0.855	2.750	0.149
Mistral-Small-3.1-24B-Instruct	0.372	0.884	2.592	0.224	0.391	0.940	1.963	0.179	0.373	0.881	2.238	0.206	0.381	0.945	2.541	0.178
Avg. (Persona)	*0.426*	*0.792*	*3.130*	*0.176*	*0.400*	*0.895*	*2.507*	*0.178*	*0.408*	*0.800*	*1.939*	*0.175*	*0.383*	*0.907*	*2.949*	*0.183*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.507	0.664	0.656	0.090	0.446	0.757	2.359	0.149
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.488	0.733	0.309	0.108	0.467	0.778	1.825	0.146
Mimo-v2.5	–	–	–	–	–	–	–	–	0.484	0.735	0.350	0.121	0.445	0.766	1.382	0.165
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.507	0.718	1.006	0.098	0.462	0.764	2.419	0.158
LongCat-2.0	–	–	–	–	–	–	–	–	0.499	0.678	1.158	0.125	0.486	0.697	2.520	0.143
GPT-4o-mini	–	–	–	–	–	–	–	–	0.412	0.838	0.367	0.172	0.444	0.809	1.413	0.152
GPT-5-mini	–	–	–	–	–	–	–	–	0.444	0.738	0.964	0.129	0.421	0.792	2.624	0.183
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.502	0.672	0.988	0.126	0.444	0.760	3.295	0.148
Qwen3-32B	–	–	–	–	–	–	–	–	0.489	0.712	1.286	0.140	0.439	0.841	2.022	0.179
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.393	0.830	1.883	0.211	0.379	0.988	1.436	0.205
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.509	0.652	1.248	0.141	0.444	0.783	2.942	0.186
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.442	0.748	0.758	0.182	0.431	0.794	1.779	0.150
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.442	0.823	0.457	0.152	0.412	0.916	2.619	0.194
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.471*	*0.734*	*0.879*	*0.138*	*0.440*	*0.804*	*2.203*	*0.166*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.428	0.845	3.143	0.188	–	–	–	–	0.445	0.777	2.819	0.151
DeepSeek-V4-Flash	–	–	–	–	0.408	0.909	1.973	0.168	–	–	–	–	0.391	0.905	1.963	0.152
Mimo-v2.5	–	–	–	–	0.407	0.900	1.672	0.150	–	–	–	–	0.399	0.902	1.053	0.135
Mimo-v2.5-Pro	–	–	–	–	0.432	0.819	1.982	0.155	–	–	–	–	0.421	0.853	2.093	0.155
LongCat-2.0	–	–	–	–	0.434	0.831	2.983	0.169	–	–	–	–	0.454	0.782	4.004	0.164
GPT-4o-mini	–	–	–	–	0.430	0.846	2.295	0.156	–	–	–	–	0.423	0.832	2.411	0.159
GPT-5-mini	–	–	–	–	0.416	0.825	3.087	0.165	–	–	–	–	0.417	0.821	3.274	0.144
Qwen2.5-72B-Instruct	–	–	–	–	0.431	0.821	2.566	0.164	–	–	–	–	0.405	0.815	2.859	0.139
Qwen3-32B	–	–	–	–	0.365	1.026	2.436	0.193	–	–	–	–	0.370	0.977	3.100	0.187
Qwen3-30B-A3B	–	–	–	–	0.350	1.047	1.192	0.141	–	–	–	–	0.345	1.045	1.345	0.167
Llama-3.3-70B-Instruct	–	–	–	–	0.390	0.922	3.623	0.200	–	–	–	–	0.382	0.905	3.645	0.194
Gemma-3-27B-it	–	–	–	–	0.350	0.980	1.214	0.140	–	–	–	–	0.360	0.952	0.818	0.148
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.403	0.904	2.460	0.195	–	–	–	–	0.419	0.884	2.819	0.173
Avg. (PRA)	–	–	–	–	*0.403*	*0.898*	*2.356*	*0.168*	–	–	–	–	*0.402*	*0.881*	*2.477*	*0.159*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.476	0.714	2.684	0.140
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.470	0.766	1.249	0.122
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.462	0.755	1.408	0.127
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.465	0.763	1.335	0.120
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.499	0.681	2.373	0.139
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.468	0.742	1.904	0.138
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.448	0.761	2.035	0.145
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.457	0.719	2.708	0.138
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.435	0.828	1.441	0.138
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.370	1.023	1.209	0.193
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.452	0.768	3.107	0.161
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.427	0.791	1.597	0.151
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.450	0.848	2.463	0.186
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.452*	*0.781*	*1.963*	*0.146*
Training																
BERT	0.412	0.693	8.195	0.293	0.435	0.788	7.031	0.197	0.419	0.693	10.920	0.304	0.455	0.742	5.775	0.175
SFT (Qwen2.5-7B-Instruct)	0.407	0.894	3.353	0.222	0.411	0.838	2.177	0.142	0.393	0.797	3.356	0.199	0.426	0.781	2.710	0.128
SFT (Qwen3-8B)	0.312	0.945	0.812	0.198	0.390	0.922	1.648	0.143	0.346	0.853	0.281	0.151	0.365	0.923	2.047	0.147

Table 23: BSA results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.283	1.076	15.270	0.282	0.380	1.054	13.125	0.294	0.305	1.090	14.656	0.290	0.355	1.044	13.700	0.303
DeepSeek-V4-Flash	0.257	1.250	15.881	0.327	0.379	1.114	13.166	0.303	0.298	1.197	14.823	0.322	0.358	1.105	13.635	0.314
Mimo-v2.5	0.283	1.143	15.270	0.297	0.371	1.035	13.343	0.287	0.295	1.047	14.897	0.282	0.385	0.981	13.011	0.281
Mimo-v2.5-Pro	**0.350**	1.076	13.721	0.279	0.366	1.101	13.446	0.314	0.310	1.199	14.551	0.306	0.348	1.120	13.869	0.319
LongCat-2.0	0.292	1.133	15.066	0.294	0.335	1.122	14.162	0.308	0.320	1.104	14.331	0.282	0.351	1.024	13.791	0.302
GPT-4o-mini	0.299	1.273	14.903	0.330	0.363	1.113	13.529	0.311	0.236	1.415	16.247	0.373	0.325	1.198	14.403	0.337
GPT-5-mini	0.244	1.237	16.166	0.324	0.380	1.023	13.135	0.283	0.260	1.175	15.692	0.305	0.349	1.072	13.843	0.302
Qwen2.5-72B-Instruct	0.269	1.340	15.596	0.334	0.332	1.183	14.235	0.323	0.250	1.384	15.922	0.356	0.331	1.143	14.259	0.321
Qwen3-32B	0.306	1.267	14.740	0.315	0.355	1.114	13.706	0.310	0.221	1.510	16.603	0.398	0.343	1.155	13.986	0.330
Qwen3-30B-A3B	0.273	1.322	15.514	0.341	0.297	1.266	15.034	0.357	0.276	1.330	15.326	0.347	0.278	1.262	15.482	0.360
Llama-3.3-70B-Instruct	0.288	1.191	15.148	0.310	0.373	1.175	13.291	0.319	0.319	1.215	14.342	0.306	0.325	1.229	14.390	0.346
Gemma-3-27B-it	0.202	1.428	17.145	0.363	0.353	1.059	13.758	0.298	0.248	1.278	15.985	0.329	0.338	1.124	14.090	0.318
Mistral-Small-3.1-24B-Instruct	0.296	1.209	14.985	0.307	0.354	1.055	13.726	0.291	0.311	1.071	14.530	0.278	0.308	1.134	14.793	0.322
Avg. (Zero-shot)	*0.280*	*1.227*	*15.339*	*0.316*	*0.357*	*1.109*	*13.666*	*0.307*	*0.281*	*1.232*	*15.224*	*0.321*	*0.338*	*1.122*	*14.096*	*0.320*
Persona																
DeepSeek-V4-Pro	0.329	1.042	5.596	0.190	0.399	0.957	4.272	0.164	0.325	1.044	2.602	0.161	0.414	0.936	4.686	0.168
DeepSeek-V4-Flash	0.301	1.080	4.708	0.177	0.374	1.000	3.551	0.171	0.312	1.063	3.261	0.168	0.391	0.986	4.102	0.175
Mimo-v2.5	0.313	**1.035**	5.253	**0.170**	0.368	1.005	3.801	0.166	0.309	1.060	3.384	0.157	0.369	0.998	4.681	0.178
Mimo-v2.5-Pro	0.319	1.081	4.344	0.180	0.376	1.012	3.758	0.167	0.328	1.062	2.057	0.153	0.388	0.973	3.855	0.176
LongCat-2.0	**0.333**	**1.037**	5.095	0.177	0.384	0.956	3.607	0.163	0.340	1.003	2.503	0.158	0.392	0.933	4.706	0.164
GPT-4o-mini	0.265	1.313	**2.113**	0.212	0.338	1.115	**1.974**	**0.155**	0.269	1.273	0.885	0.192	0.347	1.102	2.825	0.174
GPT-5-mini	0.276	1.140	3.168	**0.174**	0.366	1.022	3.725	0.164	0.290	1.116	2.625	0.178	0.380	0.994	4.510	0.184
Qwen2.5-72B-Instruct	0.322	1.039	4.055	0.175	0.345	1.032	3.610	0.179	0.307	1.049	2.217	0.171	0.363	1.005	4.165	0.186
Qwen3-32B	0.228	1.340	5.934	0.221	0.351	1.074	3.924	0.190	0.277	1.222	3.188	0.189	0.364	1.099	3.967	0.205
Qwen3-30B-A3B	0.251	1.329	6.044	0.248	0.308	1.166	3.574	0.206	0.256	1.291	2.298	0.240	0.309	1.169	3.782	0.213
Llama-3.3-70B-Instruct	0.301	1.071	3.538	0.180	0.377	0.980	3.688	0.163	0.316	1.043	2.717	0.167	0.372	0.977	4.436	0.185
Gemma-3-27B-it	0.267	1.156	3.959	0.183	0.334	1.085	3.640	0.201	0.278	1.169	1.464	0.165	0.344	1.056	4.078	0.205
Mistral-Small-3.1-24B-Instruct	0.283	1.108	**2.904**	**0.170**	0.352	1.075	3.193	0.183	0.307	1.090	1.311	0.154	0.349	1.063	3.595	0.185
Avg. (Persona)	*0.291*	*1.136*	*4.362*	*0.189*	*0.360*	*1.037*	*3.563*	*0.175*	*0.301*	*1.114*	*2.347*	*0.173*	*0.368*	*1.022*	*4.107*	*0.184*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.361	0.958	1.562	**0.128**	**0.450**	**0.842**	3.096	0.155
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.362	0.973	1.222	0.130	0.427	0.924	3.068	0.157
Mimo-v2.5	–	–	–	–	–	–	–	–	0.361	**0.946**	1.854	0.139	0.419	0.912	3.534	0.172
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	**0.378**	**0.955**	**0.769**	**0.129**	0.434	**0.871**	2.860	**0.144**
LongCat-2.0	–	–	–	–	–	–	–	–	**0.367**	0.963	1.284	0.141	0.441	**0.871**	3.236	**0.151**
GPT-4o-mini	–	–	–	–	–	–	–	–	0.301	1.239	1.418	0.202	0.378	1.081	3.252	0.193
GPT-5-mini	–	–	–	–	–	–	–	–	0.320	1.056	1.318	0.170	0.413	0.919	3.112	0.159
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.325	1.035	1.610	0.174	0.380	0.989	3.529	0.184
Qwen3-32B	–	–	–	–	–	–	–	–	0.303	1.145	2.171	0.176	0.390	1.036	3.844	0.191
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.267	1.280	1.341	0.203	0.341	1.131	3.455	0.223
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.363	0.964	1.326	0.136	0.414	0.906	3.070	0.169
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.328	1.085	0.904	0.148	0.376	0.991	3.097	0.177
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.328	1.087	**0.648**	0.132	0.379	1.051	3.144	0.188
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.336*	*1.053*	*1.340*	*0.155*	*0.403*	*0.963*	*3.254*	*0.174*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.414	0.947	5.131	0.180	–	–	–	–	0.410	0.955	5.118	0.176
DeepSeek-V4-Flash	–	–	–	–	0.412	0.973	3.402	0.158	–	–	–	–	0.391	1.008	2.928	0.170
Mimo-v2.5	–	–	–	–	0.401	0.991	4.251	0.172	–	–	–	–	0.387	1.010	4.558	0.193
Mimo-v2.5-Pro	–	–	–	–	**0.420**	0.958	4.503	0.173	–	–	–	–	0.398	0.965	5.022	0.179
LongCat-2.0	–	–	–	–	0.418	0.942	4.280	0.170	–	–	–	–	0.413	0.943	5.597	0.188
GPT-4o-mini	–	–	–	–	0.403	0.998	3.788	0.165	–	–	–	–	0.398	0.982	4.020	0.176
GPT-5-mini	–	–	–	–	0.412	0.980	4.238	0.183	–	–	–	–	0.405	0.981	4.753	0.184
Qwen2.5-72B-Instruct	–	–	–	–	0.379	1.005	3.988	0.169	–	–	–	–	0.401	0.944	5.018	0.179
Qwen3-32B	–	–	–	–	0.401	1.055	3.791	0.182	–	–	–	–	0.382	1.079	4.423	0.197
Qwen3-30B-A3B	–	–	–	–	0.323	1.220	**1.465**	0.172	–	–	–	–	0.342	1.188	**1.655**	0.158
Llama-3.3-70B-Instruct	–	–	–	–	0.403	0.977	4.093	0.174	–	–	–	–	0.386	1.025	5.076	0.198
Gemma-3-27B-it	–	–	–	–	0.361	1.109	2.430	0.157	–	–	–	–	0.365	1.105	3.255	0.165
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.406	0.998	4.929	0.187	–	–	–	–	0.393	1.020	5.631	0.210
Avg. (PRA)	–	–	–	–	*0.396*	*1.012*	*3.868*	*0.173*	–	–	–	–	*0.390*	*1.016*	*4.389*	*0.183*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.432	0.908	4.477	0.176
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.438	0.920	3.248	0.157
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.411	0.979	3.868	0.181
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.424	0.918	3.226	0.162
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.441	0.892	3.933	0.164
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.408	0.992	3.328	0.176
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.414	0.968	3.951	0.172
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.410	0.941	4.342	0.185
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.402	1.028	4.108	0.188
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.349	1.159	**2.195**	0.181
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.428	0.916	3.805	0.168
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.393	1.017	3.218	0.176
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.406	1.013	4.173	0.186
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.412*	*0.973*	*3.682*	*0.175*
Training																
BERT	0.317	1.104	10.720	0.258	**0.425**	**0.933**	4.491	0.165	0.341	1.179	3.907	0.211	**0.457**	0.881	4.548	0.170
SFT (Qwen2.5-7B-Instruct)	0.310	1.113	6.042	0.227	0.416	**0.915**	3.961	**0.154**	0.302	1.098	5.613	0.204	0.402	0.926	4.492	0.163
SFT (Qwen3-8B)	0.292	1.175	5.191	0.200	0.387	0.972	3.748	0.166	0.300	1.134	3.313	0.164	0.371	0.994	4.132	0.189

Table 24: CGSS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	**0.427**	1.075	11.928	0.233	0.310	1.188	14.701	0.294	0.192	1.564	17.412	0.378	0.317	1.210	14.540	0.316
DeepSeek-V4-Flash	**0.427**	1.075	11.928	0.233	0.307	1.271	14.760	0.301	**0.463**	1.098	11.160	0.225	0.343	1.317	13.932	0.326
Mimo-v2.5	0.382	1.276	12.970	0.265	0.348	1.128	13.817	0.292	0.461	1.052	11.220	0.213	0.383	1.050	13.020	0.274
Mimo-v2.5-Pro	**0.427**	1.106	11.928	0.231	0.401	0.999	12.608	0.262	0.223	1.586	16.691	0.378	0.370	1.000	13.324	0.265
LongCat-2.0	0.402	1.106	12.507	0.237	0.321	1.232	14.436	0.305	0.458	1.206	11.280	0.246	0.302	1.302	14.882	0.339
GPT-4o-mini	0.327	1.397	14.242	0.304	0.250	1.457	16.087	0.366	0.454	1.099	11.370	0.219	0.226	1.512	16.630	0.376
GPT-5-mini	0.206	1.563	17.019	0.365	0.330	1.238	14.230	0.300	0.090	1.751	19.757	0.416	0.252	1.370	16.022	0.359
Qwen2.5-72B-Instruct	**0.427**	1.116	11.928	0.240	0.325	1.175	14.348	0.289	0.299	1.491	14.947	0.303	0.297	1.195	14.996	0.312
Qwen3-32B	0.291	1.427	15.052	0.312	0.289	1.315	15.173	0.337	0.446	1.025	11.551	0.209	0.274	1.538	15.528	0.398
Qwen3-30B-A3B	0.261	1.332	15.747	0.307	0.303	1.323	14.849	0.305	0.272	1.424	15.579	0.344	0.295	1.191	15.034	0.306
Llama-3.3-70B-Instruct	0.201	1.432	17.135	0.321	0.342	1.243	13.964	0.320	0.303	1.423	14.857	0.291	0.269	1.404	15.642	0.365
Gemma-3-27B-it	0.176	1.477	17.714	0.357	0.316	1.280	14.554	0.322	0.200	1.644	17.232	0.399	0.272	1.309	15.566	0.346
Mistral-Small-3.1-24B-Instruct	0.412	1.171	12.275	0.256	0.298	1.175	14.967	0.302	0.423	1.034	12.092	0.224	0.353	1.111	13.704	0.294
Avg. (Zero-shot)	*0.336*	*1.273*	*14.029*	*0.282*	*0.319*	*1.233*	*14.500*	*0.307*	*0.330*	*1.338*	*14.242*	*0.296*	*0.304*	*1.270*	*14.832*	*0.329*
Persona																
DeepSeek-V4-Pro	**0.427**	**0.975**	3.721	**0.110**	0.402	1.038	4.118	0.160	0.383	1.051	1.922	0.138	0.393	1.066	4.670	0.174
DeepSeek-V4-Flash	0.422	1.065	**2.300**	0.144	0.387	1.110	3.228	0.163	0.424	1.050	1.286	0.146	0.416	1.036	4.509	0.160
Mimo-v2.5	**0.427**	1.005	4.344	0.136	0.391	1.019	3.139	0.150	0.362	1.078	1.524	0.130	0.403	0.974	4.479	0.155
Mimo-v2.5-Pro	0.377	1.070	3.054	**0.100**	0.393	1.082	3.183	0.145	0.430	**0.950**	1.241	**0.092**	0.408	0.995	3.826	0.148
LongCat-2.0	0.372	1.085	3.102	0.168	0.379	1.060	3.240	0.159	0.393	1.069	2.749	0.159	0.380	1.076	5.207	0.180
GPT-4o-mini	0.342	1.196	**2.838**	0.162	0.337	1.238	3.344	0.172	0.351	1.154	**0.747**	0.142	0.340	1.219	3.826	0.185
GPT-5-mini	0.302	1.111	3.594	0.163	0.357	1.102	3.350	0.159	0.321	1.117	3.520	0.156	0.345	1.124	4.320	0.184
Qwen2.5-72B-Instruct	0.312	1.080	3.561	0.167	0.362	1.090	4.084	0.181	0.360	1.076	3.472	0.170	0.351	1.099	3.995	0.189
Qwen3-32B	0.327	1.156	2.845	0.190	0.365	1.141	4.082	0.171	0.349	1.151	4.618	0.187	0.368	1.157	4.547	0.192
Qwen3-30B-A3B	0.392	1.090	4.135	0.165	0.334	1.136	3.290	0.162	0.359	1.138	2.251	0.169	0.320	1.074	4.215	0.167
Llama-3.3-70B-Instruct	0.337	1.171	5.429	0.153	0.362	1.079	4.453	0.159	0.322	1.138	2.183	0.156	0.347	1.096	4.937	0.194
Gemma-3-27B-it	0.352	1.050	4.361	0.142	0.332	1.119	3.247	0.165	0.322	1.123	1.685	0.159	0.327	1.114	3.452	0.174
Mistral-Small-3.1-24B-Instruct	0.377	1.060	2.943	0.165	0.351	1.111	4.727	0.178	0.364	1.047	1.641	0.172	0.365	1.097	4.457	0.182
Avg. (Persona)	*0.366*	*1.086*	*3.556*	*0.151*	*0.366*	*1.102*	*3.653*	*0.163*	*0.365*	*1.088*	*2.218*	*0.152*	*0.366*	*1.087*	*4.342*	*0.176*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.433	1.039	1.642	0.131	0.413	0.964	2.678	0.136
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.449	1.077	1.027	0.134	0.411	1.031	2.892	0.130
Mimo-v2.5	–	–	–	–	–	–	–	–	0.428	1.027	1.650	**0.128**	0.399	0.985	**2.352**	0.137
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.445	**1.003**	0.754	0.131	0.409	0.977	3.559	**0.121**
LongCat-2.0	–	–	–	–	–	–	–	–	0.446	1.007	0.839	0.144	0.426	0.929	2.612	**0.124**
GPT-4o-mini	–	–	–	–	–	–	–	–	0.394	1.124	0.887	0.172	0.361	1.089	3.179	0.159
GPT-5-mini	–	–	–	–	–	–	–	–	0.332	1.074	2.604	0.154	0.366	1.074	3.787	0.172
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.394	1.023	1.531	0.161	0.365	1.021	4.011	0.161
Qwen3-32B	–	–	–	–	–	–	–	–	0.402	1.076	0.920	0.149	0.384	1.073	2.837	0.151
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.407	1.127	**0.701**	0.182	0.342	1.096	3.120	0.158
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.332	1.168	1.526	0.174	0.361	1.081	2.896	0.157
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.375	1.097	1.138	0.165	0.353	1.132	3.466	0.170
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.406	1.078	1.611	0.175	0.343	1.158	3.397	0.183
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.403*	*1.071*	*1.295*	*0.154*	*0.380*	*1.047*	*3.137*	*0.151*
PRA																
DeepSeek-V4-Pro	–	–	–	–	**0.440**	**0.997**	5.753	0.170	–	–	–	–	0.434	0.934	5.121	0.159
DeepSeek-V4-Flash	–	–	–	–	0.408	1.101	3.396	0.163	–	–	–	–	0.411	1.064	4.246	0.171
Mimo-v2.5	–	–	–	–	0.410	1.069	2.735	0.140	–	–	–	–	0.417	0.993	2.990	0.134
Mimo-v2.5-Pro	–	–	–	–	0.417	1.047	3.330	0.156	–	–	–	–	0.429	1.002	4.804	0.152
LongCat-2.0	–	–	–	–	0.429	1.009	5.707	0.177	–	–	–	–	0.427	0.959	6.185	0.172
GPT-4o-mini	–	–	–	–	0.403	1.114	3.133	0.160	–	–	–	–	0.358	1.149	3.711	0.162
GPT-5-mini	–	–	–	–	0.416	1.037	3.412	0.158	–	–	–	–	0.401	0.992	4.533	0.163
Qwen2.5-72B-Instruct	–	–	–	–	0.411	1.012	4.578	0.171	–	–	–	–	0.383	1.053	5.931	0.183
Qwen3-32B	–	–	–	–	0.392	1.115	3.251	0.156	–	–	–	–	0.404	1.068	4.173	0.174
Qwen3-30B-A3B	–	–	–	–	0.343	1.291	**2.309**	0.167	–	–	–	–	0.322	1.244	2.461	0.137
Llama-3.3-70B-Instruct	–	–	–	–	0.403	1.063	4.513	0.171	–	–	–	–	0.398	1.066	4.640	0.176
Gemma-3-27B-it	–	–	–	–	0.356	1.254	**2.329**	0.166	–	–	–	–	0.320	1.219	2.799	0.162
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.380	1.143	4.582	0.163	–	–	–	–	0.401	1.068	4.782	0.161
Avg. (PRA)	–	–	–	–	*0.401*	*1.096*	*3.771*	*0.163*	–	–	–	–	*0.393*	*1.062*	*4.337*	*0.162*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.436	**0.919**	3.990	0.154
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.429	1.023	3.897	0.171
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.422	0.975	3.092	0.141
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.437	0.942	2.910	0.132
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	**0.447**	**0.873**	3.719	0.132
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.370	1.117	3.179	0.159
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.404	0.990	3.660	0.153
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.389	1.002	4.521	0.151
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.394	1.054	3.004	0.131
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.320	1.241	**1.631**	0.141
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.381	1.078	2.637	0.160
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.335	1.191	2.680	0.175
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.380	1.122	4.041	0.173
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.396*	*1.041*	*3.305*	*0.152*
Training																
BERT	0.317	1.141	10.949	0.224	**0.457**	**0.926**	8.269	0.178	0.258	1.222	12.426	0.233	**0.439**	0.934	6.805	0.179
SFT (Qwen2.5-7B-Instruct)	0.286	1.191	3.380	0.183	0.416	1.013	3.275	**0.127**	0.316	1.195	1.728	0.184	0.426	0.936	3.864	0.140
SFT (Qwen3-8B)	0.307	1.251	4.673	0.187	0.403	1.064	2.950	0.135	0.343	1.214	2.170	0.190	0.373	1.092	2.567	0.139

Table 25: CSES results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	**0.610**	0.542	8.336	0.313	0.429	0.985	12.322	0.344	0.543	0.595	9.843	0.322	0.413	0.974	12.718	0.347
DeepSeek-V4-Flash	0.410	0.801	12.928	0.456	0.461	0.906	11.593	0.319	**0.614**	0.538	8.219	0.292	0.462	0.949	11.591	0.327
Mimo-v2.5	0.593	0.559	8.724	0.324	0.437	0.817	12.151	0.335	0.427	0.778	12.524	0.407	0.422	0.941	12.503	0.356
Mimo-v2.5-Pro	0.329	0.890	14.804	0.544	0.410	1.025	12.766	0.356	0.403	0.773	13.060	0.420	0.455	0.911	11.734	0.317
LongCat-2.0	0.565	0.604	9.371	0.353	0.447	0.947	11.908	0.338	0.560	**0.487**	9.449	0.306	0.419	1.179	12.575	0.374
GPT-4o-mini	0.337	0.899	14.610	0.535	0.406	0.991	12.852	0.380	0.539	0.681	9.938	0.366	0.399	1.085	13.040	0.374
GPT-5-mini	0.500	0.683	10.858	0.357	0.443	0.975	11.993	0.342	0.434	0.703	12.367	0.397	0.361	1.177	13.917	0.405
Qwen2.5-72B-Instruct	0.455	0.753	11.893	0.428	0.343	1.019	14.310	0.407	0.433	0.703	12.382	0.397	0.332	1.128	14.579	0.434
Qwen3-32B	0.461	0.815	11.764	0.437	0.433	1.051	12.237	0.357	0.405	0.726	13.013	0.448	0.431	1.049	12.289	0.353
Qwen3-30B-A3B	0.393	0.632	13.316	0.439	0.462	0.814	11.564	0.312	0.408	0.723	12.950	0.445	0.441	0.967	12.074	0.342
Llama-3.3-70B-Instruct	0.435	0.705	12.346	0.441	0.482	0.836	11.107	0.302	0.418	0.808	12.714	0.487	0.470	0.938	11.394	0.315
Gemma-3-27B-it	0.494	0.713	10.988	0.406	0.365	1.142	13.810	0.413	0.422	0.805	12.635	0.448	0.369	1.181	13.720	0.397
Mistral-Small-3.1-24B-Instruct	0.492	0.604	11.052	0.374	0.462	0.907	11.564	0.322	0.358	0.675	14.117	0.461	0.492	0.908	10.893	0.297
Avg. (Zero-shot)	*0.467*	*0.708*	*11.614*	*0.416*	*0.429*	*0.955*	*12.321*	*0.348*	*0.459*	*0.692*	*11.785*	*0.400*	*0.420*	*1.030*	*12.541*	*0.357*
Persona																
DeepSeek-V4-Pro	0.478	0.649	1.766	0.259	0.400	1.058	4.060	0.256	0.470	0.665	0.900	0.285	0.397	1.075	4.452	0.272
DeepSeek-V4-Flash	0.525	0.559	1.596	0.196	0.491	0.841	4.251	0.206	**0.582**	**0.525**	1.382	0.188	0.485	0.877	4.221	0.220
Mimo-v2.5	0.539	0.567	3.397	0.230	0.437	1.004	3.861	0.233	0.532	0.577	2.674	0.248	0.430	1.023	4.047	0.262
Mimo-v2.5-Pro	0.514	0.581	1.414	0.192	0.442	0.955	3.374	0.219	0.537	0.571	0.717	0.189	0.448	0.936	3.370	0.237
LongCat-2.0	0.494	0.562	1.683	**0.170**	0.454	0.973	2.856	0.230	0.527	0.545	0.656	**0.175**	0.451	0.991	2.662	0.244
GPT-4o-mini	0.469	0.694	**0.704**	0.241	0.388	1.116	2.031	0.219	0.489	0.662	**0.398**	0.241	0.382	1.117	2.224	0.246
GPT-5-mini	0.562	**0.534**	2.077	0.233	0.442	0.866	2.915	0.207	0.544	0.556	1.467	0.241	0.461	0.859	3.523	0.208
Qwen2.5-72B-Instruct	0.497	0.660	2.452	0.272	0.382	1.035	3.319	0.294	0.488	0.631	1.985	0.289	0.377	1.082	3.474	0.298
Qwen3-32B	0.494	0.671	2.702	0.265	0.406	1.017	4.005	0.268	0.512	0.660	1.614	0.226	0.421	1.000	4.201	0.276
Qwen3-30B-A3B	0.497	0.666	3.392	0.294	0.366	1.082	3.371	0.284	0.510	0.625	2.997	0.288	0.390	1.064	2.403	0.289
Llama-3.3-70B-Instruct	0.565	0.581	2.645	0.257	0.461	0.920	3.470	0.231	0.581	0.549	1.666	0.247	0.485	0.894	5.023	0.237
Gemma-3-27B-it	0.416	0.772	**0.872**	0.271	0.361	1.177	**1.594**	0.274	0.443	0.721	0.839	0.264	0.350	1.211	1.733	0.303
Mistral-Small-3.1-24B-Instruct	0.522	0.551	4.009	0.259	0.392	1.025	4.202	0.291	0.515	0.566	4.133	0.266	0.394	1.034	4.979	0.301
Avg. (Persona)	*0.506*	*0.619*	*2.208*	*0.242*	*0.417*	*1.005*	*3.332*	*0.247*	*0.518*	*0.604*	*1.648*	*0.242*	*0.421*	*1.012*	*3.562*	*0.261*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.508	0.620	0.774	0.256	0.448	0.992	3.332	0.248
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.562	0.582	1.250	0.202	0.507	0.862	4.120	0.204
Mimo-v2.5	–	–	–	–	–	–	–	–	0.556	0.550	1.477	0.230	0.451	1.035	3.298	0.238
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.564	0.558	**0.521**	0.203	0.490	0.869	2.693	0.209
LongCat-2.0	–	–	–	–	–	–	–	–	0.546	0.539	0.594	0.191	0.469	0.996	3.248	0.246
GPT-4o-mini	–	–	–	–	–	–	–	–	0.477	0.656	0.834	0.256	0.399	1.131	2.644	0.281
GPT-5-mini	–	–	–	–	–	–	–	–	0.568	**0.525**	1.753	0.206	0.479	0.824	2.422	0.200
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.515	0.581	0.904	0.235	0.404	1.039	2.452	0.283
Qwen3-32B	–	–	–	–	–	–	–	–	0.535	0.621	0.885	0.258	0.449	0.929	3.200	0.229
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.543	0.606	0.772	0.258	0.423	0.997	2.962	0.279
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.566	0.561	1.318	0.199	0.503	0.854	2.980	0.197
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.498	0.621	0.668	0.222	0.375	1.167	2.510	0.293
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.536	0.575	0.960	0.257	0.409	1.052	3.298	0.284
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.536*	*0.584*	*0.978*	*0.229*	*0.447*	*0.981*	*3.012*	*0.245*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.511	0.748	2.833	0.174	–	–	–	–	0.547	0.719	2.602	0.157
DeepSeek-V4-Flash	–	–	–	–	0.514	0.765	1.910	0.154	–	–	–	–	0.514	0.776	1.761	0.169
Mimo-v2.5	–	–	–	–	0.536	0.684	2.393	0.162	–	–	–	–	0.558	0.673	2.474	0.144
Mimo-v2.5-Pro	–	–	–	–	0.543	0.705	2.185	**0.149**	–	–	–	–	0.567	0.674	2.715	0.148
LongCat-2.0	–	–	–	–	0.555	0.697	3.696	0.169	–	–	–	–	0.566	0.663	2.986	0.156
GPT-4o-mini	–	–	–	–	0.557	0.694	2.283	0.153	–	–	–	–	0.577	0.673	2.388	**0.127**
GPT-5-mini	–	–	–	–	0.537	**0.676**	2.278	0.152	–	–	–	–	0.551	0.656	2.557	0.137
Qwen2.5-72B-Instruct	–	–	–	–	0.528	0.734	2.799	0.169	–	–	–	–	0.544	0.707	2.966	0.168
Qwen3-32B	–	–	–	–	0.511	0.791	2.787	0.187	–	–	–	–	0.546	0.724	2.772	0.192
Qwen3-30B-A3B	–	–	–	–	0.471	0.917	2.106	0.209	–	–	–	–	0.490	0.858	1.864	0.185
Llama-3.3-70B-Instruct	–	–	–	–	0.535	0.690	3.201	0.164	–	–	–	–	0.568	0.648	2.766	0.149
Gemma-3-27B-it	–	–	–	–	0.450	0.980	**1.238**	0.181	–	–	–	–	0.478	0.925	**1.398**	0.160
Mistral-Small-3.1-24B-Instruct	–	–	–	–	**0.558**	0.680	3.730	0.164	–	–	–	–	0.570	0.680	3.260	0.165
Avg. (PRA)	–	–	–	–	*0.523*	*0.751*	*2.572*	*0.168*	–	–	–	–	*0.544*	*0.721*	*2.501*	*0.158*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.543	0.709	1.728	0.145
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.560	0.722	1.853	0.136
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.559	0.703	1.560	0.128
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	**0.596**	**0.626**	2.657	**0.126**
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.562	0.740	1.834	0.129
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.537	0.799	2.415	0.177
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.567	0.661	2.203	0.128
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.533	0.729	2.574	0.185
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.541	0.749	2.904	0.172
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.481	0.873	**1.479**	0.169
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.579	0.648	2.203	0.131
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.491	0.907	1.542	0.162
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.559	0.721	2.405	0.159
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.547*	*0.738*	*2.105*	*0.150*
Training																
BERT	0.399	0.803	6.897	0.357	**0.628**	**0.528**	3.901	**0.151**	0.409	0.799	5.837	0.371	**0.655**	**0.514**	4.227	0.149
SFT (Qwen2.5-7B-Instruct)	0.551	0.556	2.705	0.257	0.522	0.713	2.379	0.181	0.523	0.578	1.480	0.264	0.563	0.665	3.125	0.176
SFT (Qwen3-8B)	0.590	**0.480**	2.468	**0.176**	0.486	0.843	2.622	0.201	0.546	0.544	2.654	**0.187**	0.499	0.824	3.468	0.195

Table 26: CSS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.404	0.930	12.626	0.257	0.413	0.816	12.518	0.277	**0.420**	0.795	12.235	0.229	0.386	0.847	13.143	0.284
DeepSeek-V4-Flash	**0.439**	0.903	11.822	0.255	0.414	0.794	12.497	0.271	0.372	0.875	13.356	0.246	0.401	0.867	12.803	0.282
Mimo-v2.5	0.406	0.897	12.584	0.256	0.429	0.767	12.148	0.262	0.382	0.873	13.118	0.253	0.393	0.860	12.986	0.290
Mimo-v2.5-Pro	**0.436**	0.888	11.906	0.254	0.431	0.770	12.105	0.263	0.379	0.853	13.197	0.246	0.447	0.737	11.745	0.249
LongCat-2.0	0.373	0.925	13.345	0.256	0.418	0.817	12.413	0.273	0.373	0.923	13.326	0.262	0.385	0.917	13.156	0.298
GPT-4o-mini	0.279	1.184	15.504	0.342	0.355	0.958	13.854	0.324	0.289	1.105	15.269	0.322	0.344	0.967	14.110	0.324
GPT-5-mini	0.292	0.998	15.208	0.281	0.361	0.921	13.716	0.304	0.273	1.096	15.626	0.317	0.296	1.023	15.221	0.336
Qwen2.5-72B-Instruct	0.375	0.923	13.303	0.255	0.364	0.910	13.663	0.302	0.361	0.941	13.603	0.266	0.358	0.897	13.784	0.301
Qwen3-32B	0.351	1.029	13.853	0.291	0.396	0.884	12.921	0.301	0.388	0.892	12.969	0.254	0.362	0.921	13.692	0.311
Qwen3-30B-A3B	0.278	1.143	15.546	0.324	0.332	0.956	14.383	0.318	0.245	1.115	16.261	0.326	0.382	0.940	13.235	0.306
Llama-3.3-70B-Instruct	0.362	0.982	13.599	0.273	0.364	0.934	13.652	0.308	0.317	1.054	14.615	0.308	0.344	0.972	14.097	0.328
Gemma-3-27B-it	0.340	1.068	14.107	0.304	0.388	0.850	13.101	0.288	0.337	0.966	14.159	0.279	0.385	0.940	13.169	0.308
Mistral-Small-3.1-24B-Instruct	0.404	**0.860**	12.626	0.239	0.369	0.908	13.536	0.299	0.321	1.080	14.526	0.314	0.384	0.838	13.196	0.280
Avg. (Zero-shot)	*0.365*	*0.979*	*13.541*	*0.276*	*0.387*	*0.868*	*13.116*	*0.292*	*0.343*	*0.967*	*14.020*	*0.278*	*0.374*	*0.902*	*13.411*	*0.300*
Persona																
DeepSeek-V4-Pro	0.408	0.866	4.734	0.161	0.431	0.727	3.494	0.147	0.400	0.851	2.740	0.159	0.390	0.758	3.987	0.171
DeepSeek-V4-Flash	0.417	**0.833**	3.323	**0.141**	0.472	0.708	3.588	0.136	0.393	0.842	2.257	0.145	0.439	0.710	3.764	0.149
Mimo-v2.5	0.412	0.877	4.615	**0.147**	0.454	0.724	3.588	0.152	0.381	0.890	2.945	0.160	0.394	0.798	3.938	0.186
Mimo-v2.5-Pro	0.382	0.877	4.648	0.152	0.441	0.756	3.565	0.149	0.385	0.862	2.909	0.148	0.409	0.767	3.202	0.176
LongCat-2.0	0.384	0.941	5.048	0.170	0.410	0.790	3.585	0.166	0.351	0.941	2.711	0.187	0.369	0.828	4.113	0.198
GPT-4o-mini	0.327	1.215	**3.122**	0.225	0.342	1.083	1.996	0.232	0.269	1.282	2.147	0.263	0.292	1.152	3.114	0.270
GPT-5-mini	0.357	0.947	5.028	0.181	0.407	0.804	3.325	0.161	0.314	0.969	3.669	0.202	0.386	0.804	3.962	0.181
Qwen2.5-72B-Instruct	0.355	1.018	4.362	0.202	0.377	0.866	3.316	0.193	0.301	1.065	3.047	0.217	0.334	0.896	3.686	0.216
Qwen3-32B	0.397	0.928	3.175	0.158	0.393	0.862	2.978	0.181	0.362	0.923	2.005	0.163	0.371	0.886	3.593	0.202
Qwen3-30B-A3B	0.349	0.982	**2.775**	0.157	0.389	0.886	2.628	0.177	0.338	0.976	2.415	0.166	0.357	0.900	3.327	0.199
Llama-3.3-70B-Instruct	0.390	0.903	4.846	0.167	0.421	0.794	3.610	0.168	0.354	0.934	3.995	0.182	0.370	0.841	3.977	0.203
Gemma-3-27B-it	0.322	1.033	3.726	0.195	0.391	0.853	3.496	0.191	0.309	1.034	3.267	0.205	0.363	0.866	3.604	0.204
Mistral-Small-3.1-24B-Instruct	0.355	1.002	3.614	0.164	0.397	0.847	3.382	0.185	0.332	0.962	2.800	0.160	0.360	0.863	3.208	0.202
Avg. (Persona)	*0.373*	*0.955*	*4.078*	*0.171*	*0.410*	*0.823*	*3.273*	*0.172*	*0.345*	*0.964*	*2.839*	*0.181*	*0.372*	*0.851*	*3.652*	*0.197*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	**0.428**	0.791	1.341	0.150	0.457	0.659	2.957	0.145
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.404	**0.783**	1.429	**0.144**	0.485	0.686	2.459	0.123
Mimo-v2.5	–	–	–	–	–	–	–	–	0.409	0.823	1.811	**0.143**	0.460	0.687	3.131	0.151
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.417	**0.774**	1.202	0.150	0.471	0.675	3.225	0.147
LongCat-2.0	–	–	–	–	–	–	–	–	0.389	0.842	3.689	0.175	0.430	0.728	3.534	0.173
GPT-4o-mini	–	–	–	–	–	–	–	–	0.303	1.120	1.341	0.225	0.367	0.943	3.173	0.220
GPT-5-mini	–	–	–	–	–	–	–	–	0.331	0.927	2.740	0.195	0.436	0.716	3.465	0.165
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.348	0.969	2.968	0.200	0.404	0.766	3.756	0.177
Qwen3-32B	–	–	–	–	–	–	–	–	0.387	0.882	1.741	0.160	0.430	0.775	2.359	0.164
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.338	1.013	**1.188**	0.186	0.382	0.884	2.155	0.182
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.364	0.917	3.039	0.181	0.418	0.757	3.183	0.169
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.337	1.012	1.503	0.203	0.389	0.850	3.206	0.188
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.382	0.890	**0.933**	0.151	0.399	0.812	2.889	0.174
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.372*	*0.903*	*1.917*	*0.174*	*0.425*	*0.764*	*3.038*	*0.168*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.515	0.617	2.817	0.125	–	–	–	–	0.495	0.626	3.068	0.122
DeepSeek-V4-Flash	–	–	–	–	0.493	0.671	1.878	0.121	–	–	–	–	0.486	0.667	2.009	0.119
Mimo-v2.5	–	–	–	–	0.453	0.762	1.326	0.109	–	–	–	–	0.443	0.749	1.336	0.120
Mimo-v2.5-Pro	–	–	–	–	0.498	0.636	1.949	**0.102**	–	–	–	–	0.495	0.632	2.069	0.115
LongCat-2.0	–	–	–	–	0.509	0.613	3.735	0.149	–	–	–	–	0.491	0.622	4.105	0.150
GPT-4o-mini	–	–	–	–	0.451	0.789	1.435	0.126	–	–	–	–	0.446	0.783	1.638	0.135
GPT-5-mini	–	–	–	–	0.474	0.668	1.532	**0.105**	–	–	–	–	0.486	0.648	1.850	**0.114**
Qwen2.5-72B-Instruct	–	–	–	–	0.437	0.750	2.955	0.150	–	–	–	–	0.404	0.774	3.339	0.169
Qwen3-32B	–	–	–	–	0.457	0.751	2.011	0.133	–	–	–	–	0.442	0.761	2.235	0.148
Qwen3-30B-A3B	–	–	–	–	0.353	1.026	**0.742**	0.156	–	–	–	–	0.337	1.009	0.972	0.163
Llama-3.3-70B-Instruct	–	–	–	–	0.497	0.654	1.666	0.120	–	–	–	–	0.467	0.701	1.962	0.127
Gemma-3-27B-it	–	–	–	–	0.412	0.873	**0.800**	0.131	–	–	–	–	0.386	0.885	**0.875**	0.147
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.486	0.711	2.630	0.147	–	–	–	–	0.458	0.724	2.563	0.153
Avg. (PRA)	–	–	–	–	*0.464*	*0.732*	*1.960*	*0.129*	–	–	–	–	*0.449*	*0.737*	*2.155*	*0.137*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	**0.516**	**0.578**	2.886	0.125
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.510	0.624	2.463	0.131
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.466	0.705	1.448	0.125
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.508	0.599	2.511	0.123
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.495	0.620	2.807	0.141
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.446	0.784	1.461	0.147
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.452	0.704	1.395	**0.106**
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.461	0.671	3.039	0.147
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.458	0.741	1.964	0.140
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.358	0.964	**0.521**	0.158
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.472	0.663	2.176	0.118
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.445	0.758	1.909	0.140
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.461	0.725	2.105	0.150
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.465*	*0.703*	*2.053*	*0.135*
Training																
BERT	0.388	1.000	6.862	0.204	**0.551**	**0.559**	5.284	0.152	0.391	1.008	5.370	0.226	**0.568**	**0.529**	4.737	0.143
SFT (Qwen2.5-7B-Instruct)	0.403	0.875	3.370	0.157	**0.525**	**0.590**	5.128	0.143	0.394	0.857	2.702	0.156	0.510	0.600	5.212	0.152
SFT (Qwen3-8B)	0.404	**0.833**	4.214	0.160	0.499	0.647	4.235	0.143	0.390	0.823	2.503	0.154	0.493	0.642	4.627	0.148

Table 27: ESS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.366	1.203	13.532	0.340	0.306	1.759	14.563	0.318	0.358	1.375	13.606	0.321	0.312	1.649	14.470	0.305
DeepSeek-V4-Flash	0.351	1.359	13.859	0.338	0.295	2.010	14.811	0.346	0.382	1.418	13.060	0.301	0.298	1.828	14.794	0.333
Mimo-v2.5	0.377	1.340	13.264	0.323	0.307	1.775	14.542	0.317	0.379	1.212	13.125	0.282	0.301	1.677	14.706	0.308
Mimo-v2.5-Pro	0.319	1.355	14.603	0.373	0.303	1.836	14.634	0.329	0.379	1.355	13.111	0.306	0.271	1.751	15.406	0.335
LongCat-2.0	0.357	1.554	13.740	0.347	0.292	1.812	14.875	0.332	0.369	1.377	13.358	0.322	0.281	1.747	15.179	0.322
GPT-4o-mini	0.298	1.329	15.079	0.385	0.286	1.811	15.024	0.334	0.293	1.416	15.093	0.356	0.267	1.739	15.494	0.342
GPT-5-mini	0.389	1.286	12.997	0.302	0.287	1.723	15.003	0.317	0.355	1.423	13.679	0.326	0.284	1.750	15.109	0.318
Qwen2.5-72B-Instruct	0.470	1.143	11.122	0.228	0.299	1.831	14.719	0.326	0.431	1.072	11.916	0.233	0.297	1.684	14.802	0.309
Qwen3-32B	0.389	1.652	12.997	0.346	0.286	1.882	15.010	0.340	0.374	1.356	13.227	0.313	0.294	1.721	14.881	0.315
Qwen3-30B-A3B	0.407	1.447	12.580	0.299	0.288	1.862	14.967	0.330	0.419	1.243	12.200	0.273	0.263	1.763	15.599	0.344
Llama-3.3-70B-Instruct	0.413	1.198	12.431	0.270	0.294	1.766	14.833	0.326	0.409	1.222	12.440	0.270	0.296	1.724	14.820	0.325
Gemma-3-27B-it	0.384	1.256	13.116	0.296	0.304	1.742	14.613	0.311	0.352	1.340	13.737	0.311	0.257	1.693	15.730	0.329
Mistral-Small-3.1-24B-Instruct	0.413	1.389	12.431	0.288	0.287	1.757	14.996	0.313	0.356	1.317	13.650	0.310	0.290	1.702	14.960	0.309
Avg. (Zero-shot)	*0.380*	*1.347*	*13.212*	*0.318*	*0.295*	*1.813*	*14.815*	*0.326*	*0.374*	*1.317*	*13.246*	*0.302*	*0.285*	*1.725*	*15.073*	*0.323*
Persona																
DeepSeek-V4-Pro	0.495	0.942	2.297	0.123	0.356	1.440	4.101	0.155	0.446	1.007	1.467	0.131	0.345	1.398	4.295	0.156
DeepSeek-V4-Flash	0.487	0.929	1.688	**0.100**	0.348	1.513	3.981	0.164	0.442	1.023	1.751	0.123	0.352	1.427	4.102	0.159
Mimo-v2.5	0.484	1.022	2.785	0.135	0.349	1.498	3.830	0.148	0.430	1.041	1.487	0.138	0.331	1.458	4.268	0.152
Mimo-v2.5-Pro	0.487	1.067	1.834	0.127	0.349	1.571	3.714	0.146	0.449	1.108	0.779	0.111	0.336	1.491	3.768	0.149
LongCat-2.0	0.491	0.935	2.466	**0.098**	0.349	1.436	3.555	0.156	0.445	0.975	1.835	0.105	0.332	1.407	4.275	0.160
GPT-4o-mini	0.469	1.200	**1.544**	0.134	0.325	1.630	2.423	0.171	0.420	1.226	0.926	0.125	0.300	1.633	3.392	0.176
GPT-5-mini	0.494	0.957	2.614	0.109	0.339	1.374	2.979	0.140	0.441	1.007	2.441	0.130	0.340	1.356	3.617	0.147
Qwen2.5-72B-Instruct	0.474	1.023	1.726	0.128	0.319	1.546	3.732	0.174	0.441	1.067	1.350	0.118	0.312	1.526	3.895	0.167
Qwen3-32B	0.479	1.044	1.669	0.128	0.325	1.550	3.142	0.169	0.428	1.109	0.820	0.138	0.311	1.557	3.982	0.165
Qwen3-30B-A3B	0.452	1.164	**1.073**	0.129	0.308	1.608	2.549	0.158	0.438	1.154	0.626	0.122	0.310	1.566	2.720	0.152
Llama-3.3-70B-Instruct	**0.517**	**0.920**	2.351	0.110	0.341	1.465	3.490	0.137	0.451	0.995	1.030	0.123	0.320	1.457	4.198	0.155
Gemma-3-27B-it	0.460	1.092	2.105	0.126	0.309	1.591	2.780	0.173	0.434	1.117	0.736	0.109	0.289	1.593	3.240	0.175
Mistral-Small-3.1-24B-Instruct	**0.508**	1.006	1.963	0.120	0.322	1.611	3.404	0.182	0.451	1.091	1.081	0.117	0.310	1.558	3.957	0.177
Avg. (Persona)	*0.484*	*1.023*	*2.009*	*0.121*	*0.334*	*1.526*	*3.360*	*0.159*	*0.440*	*1.071*	*1.256*	*0.122*	*0.322*	*1.494*	*3.824*	*0.161*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.459	0.972	1.189	0.120	0.357	1.391	3.591	0.148
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.464	1.010	0.884	0.107	0.355	1.421	3.321	0.147
Mimo-v2.5	–	–	–	–	–	–	–	–	0.445	0.987	1.321	0.131	0.347	1.397	4.085	0.157
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.463	1.038	**0.500**	**0.096**	0.356	1.455	3.404	0.154
LongCat-2.0	–	–	–	–	–	–	–	–	0.465	0.961	1.260	**0.091**	0.333	1.422	3.875	0.177
GPT-4o-mini	–	–	–	–	–	–	–	–	0.438	1.208	0.930	0.112	0.297	1.734	4.317	0.214
GPT-5-mini	–	–	–	–	–	–	–	–	0.446	1.014	0.774	0.114	0.343	1.355	3.159	0.156
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	**0.469**	1.039	1.005	0.111	0.332	1.497	4.074	0.170
Qwen3-32B	–	–	–	–	–	–	–	–	0.457	1.070	0.873	0.122	0.326	1.550	3.840	0.178
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.428	1.184	**0.417**	0.105	0.304	1.641	2.595	0.169
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.465	**0.954**	0.916	0.105	0.339	1.402	3.886	0.151
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.450	1.099	0.602	0.105	0.297	1.619	3.687	0.184
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.432	1.140	1.049	0.120	0.311	1.588	4.174	0.184
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.452*	*1.052*	*0.902*	*0.111*	*0.330*	*1.498*	*3.693*	*0.168*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.370	1.338	3.727	0.155	–	–	–	–	0.363	1.288	4.219	0.145
DeepSeek-V4-Flash	–	–	–	–	0.343	1.534	2.600	0.145	–	–	–	–	0.338	1.445	2.838	0.135
Mimo-v2.5	–	–	–	–	0.357	1.446	4.016	0.156	–	–	–	–	0.358	1.356	3.753	0.145
Mimo-v2.5-Pro	–	–	–	–	0.353	1.509	4.000	0.156	–	–	–	–	0.361	1.421	3.590	0.146
LongCat-2.0	–	–	–	–	0.352	1.429	4.024	0.150	–	–	–	–	0.351	1.426	4.233	0.156
GPT-4o-mini	–	–	–	–	0.339	1.468	2.803	0.142	–	–	–	–	0.326	1.444	3.454	0.148
GPT-5-mini	–	–	–	–	0.348	1.332	2.836	0.137	–	–	–	–	0.337	1.350	3.325	0.142
Qwen2.5-72B-Instruct	–	–	–	–	0.343	1.538	4.417	0.179	–	–	–	–	0.341	1.462	4.927	0.175
Qwen3-32B	–	–	–	–	0.339	1.563	3.536	0.162	–	–	–	–	0.322	1.526	3.887	0.151
Qwen3-30B-A3B	–	–	–	–	0.296	1.804	**1.859**	0.165	–	–	–	–	0.306	1.721	**2.186**	0.161
Llama-3.3-70B-Instruct	–	–	–	–	0.344	1.452	3.098	0.150	–	–	–	–	0.339	1.417	4.048	0.156
Gemma-3-27B-it	–	–	–	–	0.285	1.800	**1.698**	0.165	–	–	–	–	0.284	1.761	2.354	0.168
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.333	1.609	5.116	0.184	–	–	–	–	0.335	1.566	5.273	0.174
Avg. (PRA)	–	–	–	–	*0.339*	*1.525*	*3.364*	*0.157*	–	–	–	–	*0.335*	*1.476*	*3.699*	*0.154*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.371	**1.216**	3.983	0.141
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.346	1.451	2.840	0.139
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.367	1.323	4.177	0.152
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.367	1.364	3.325	0.140
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.358	1.363	3.314	0.146
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.312	1.617	3.178	0.190
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.344	1.347	2.683	**0.132**
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.344	1.506	5.050	0.179
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.326	1.559	3.119	0.151
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.295	1.762	**1.910**	0.171
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.341	1.396	3.393	0.145
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.309	1.686	2.638	0.178
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.324	1.580	4.645	0.187
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.339*	*1.474*	*3.404*	*0.158*
Training																
BERT	0.432	1.092	7.047	0.182	**0.382**	1.350	6.850	0.191	0.444	1.019	6.083	0.174	**0.411**	**1.179**	5.511	0.155
SFT (Qwen2.5-7B-Instruct)	0.497	**0.897**	2.966	0.114	**0.374**	**1.280**	3.080	**0.133**	**0.475**	**0.942**	2.094	0.121	**0.384**	1.251	3.852	**0.130**
SFT (Qwen3-8B)	0.501	0.952	2.500	0.117	0.373	**1.321**	3.144	**0.134**	0.468	0.979	1.958	0.119	0.365	1.375	3.402	0.141

Table 28: EVS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.394	0.827	12.898	0.290	0.467	0.881	11.265	0.284	0.424	0.893	12.196	0.315	0.458	0.845	11.480	0.274
DeepSeek-V4-Flash	0.336	0.781	14.246	0.321	0.463	0.864	11.346	0.277	0.317	0.923	14.665	0.381	0.464	0.855	11.358	0.276
Mimo-v2.5	0.361	0.936	13.660	0.323	0.430	0.930	12.106	0.301	0.400	0.832	12.766	0.295	0.457	0.866	11.520	0.283
Mimo-v2.5-Pro	0.366	0.860	13.543	0.346	0.408	0.975	12.624	0.317	0.423	0.786	12.240	0.283	0.442	0.915	11.866	0.300
LongCat-2.0	0.349	0.822	13.953	0.293	0.430	0.912	12.106	0.304	0.433	0.713	12.006	0.259	0.422	0.913	12.314	0.299
GPT-4o-mini	0.351	0.858	13.894	0.338	0.433	0.942	12.041	0.309	0.342	1.006	14.095	0.359	0.371	1.003	13.493	0.327
GPT-5-mini	0.361	0.898	13.660	0.313	0.396	0.918	12.899	0.305	0.316	0.966	14.694	0.371	0.387	0.959	13.127	0.321
Qwen2.5-72B-Instruct	0.374	0.990	13.367	0.330	0.436	0.896	11.960	0.284	0.385	0.847	13.116	0.303	0.423	0.890	12.293	0.287
Qwen3-32B	0.361	0.936	13.660	0.323	0.383	0.972	13.190	0.323	0.388	0.812	13.043	0.288	0.441	0.921	11.886	0.300
Qwen3-30B-A3B	0.387	0.779	13.074	0.335	0.402	0.895	12.753	0.301	0.371	0.789	13.438	0.326	0.383	0.981	13.209	0.323
Llama-3.3-70B-Instruct	0.346	0.878	14.011	0.344	0.416	0.961	12.430	0.316	0.397	0.889	12.839	0.308	0.443	0.933	11.846	0.305
Gemma-3-27B-it	0.377	0.720	13.308	0.301	0.436	0.902	11.977	0.296	0.339	0.819	14.154	0.332	0.422	0.915	12.314	0.299
Mistral-Small-3.1-24B-Instruct	0.354	0.796	13.836	0.324	0.403	0.926	12.737	0.309	0.371	0.784	13.423	0.320	0.415	0.934	12.476	0.302
Avg. (Zero-shot)	*0.363*	*0.852*	*13.624*	*0.322*	*0.423*	*0.921*	*12.264*	*0.302*	*0.377*	*0.851*	*13.283*	*0.318*	*0.425*	*0.918*	*12.245*	*0.300*
Persona																
DeepSeek-V4-Pro	0.463	0.664	1.282	0.134	0.462	0.818	2.232	0.147	0.445	0.700	2.119	0.128	0.485	0.809	1.992	0.152
DeepSeek-V4-Flash	0.461	0.667	1.300	0.115	0.467	**0.771**	1.529	0.123	**0.468**	0.663	1.341	0.118	0.465	0.804	1.906	0.136
Mimo-v2.5	0.458	0.679	0.833	**0.114**	0.473	0.800	**0.969**	0.136	0.442	0.727	0.596	0.118	0.471	0.855	1.925	0.162
Mimo-v2.5-Pro	0.412	0.712	1.128	**0.108**	0.468	0.784	0.970	**0.109**	0.446	0.689	1.276	0.109	0.472	0.804	1.212	0.132
LongCat-2.0	0.473	**0.641**	1.763	0.136	0.454	0.829	2.230	0.140	0.463	0.661	0.978	0.135	0.456	0.816	2.326	0.156
GPT-4o-mini	0.374	0.845	**0.578**	0.154	0.387	0.995	2.165	0.182	0.362	0.906	0.656	0.154	0.377	1.045	2.265	0.210
GPT-5-mini	0.443	0.684	1.371	0.115	0.408	0.852	2.928	0.162	0.450	0.680	1.096	0.115	0.440	0.845	2.619	0.165
Qwen2.5-72B-Instruct	0.415	0.687	1.117	0.124	0.436	0.845	1.496	**0.122**	0.445	0.692	1.048	0.130	0.428	0.895	2.001	0.167
Qwen3-32B	0.445	0.723	1.842	0.148	0.375	0.989	2.891	0.199	0.424	0.779	0.862	0.145	0.384	1.005	2.765	0.228
Qwen3-30B-A3B	0.415	0.730	0.852	0.143	0.407	0.850	1.032	0.152	0.435	0.713	0.814	0.140	0.398	0.937	1.996	0.184
Llama-3.3-70B-Instruct	**0.476**	**0.641**	1.779	0.122	0.447	0.853	2.576	0.151	0.459	0.680	1.437	0.115	0.453	0.863	2.563	0.161
Gemma-3-27B-it	0.394	0.774	**0.640**	0.117	0.422	0.868	1.449	0.138	0.407	0.775	0.573	0.106	0.428	0.922	1.362	0.174
Mistral-Small-3.1-24B-Instruct	**0.483**	**0.623**	1.592	0.118	0.454	0.841	2.266	0.153	0.456	0.691	1.577	0.146	0.448	0.869	2.762	0.171
Avg. (Persona)	*0.439*	*0.698*	*1.237*	*0.127*	*0.435*	*0.854*	*1.903*	*0.147*	*0.439*	*0.720*	*1.106*	*0.128*	*0.439*	*0.882*	*2.130*	*0.169*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.462	0.672	**0.169**	**0.083**	**0.533**	0.776	1.116	**0.117**
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.465	**0.655**	**0.172**	0.105	0.490	0.793	1.094	**0.117**
Mimo-v2.5	–	–	–	–	–	–	–	–	0.451	0.714	0.461	0.106	0.518	0.821	1.710	0.150
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.457	0.687	0.201	0.103	0.510	0.791	1.095	**0.116**
LongCat-2.0	–	–	–	–	–	–	–	–	0.460	0.673	0.659	0.120	0.489	0.841	2.350	0.150
GPT-4o-mini	–	–	–	–	–	–	–	–	0.403	0.848	0.346	0.124	0.428	0.966	1.932	0.191
GPT-5-mini	–	–	–	–	–	–	–	–	0.437	0.708	0.498	0.112	0.511	0.767	2.060	0.146
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	**0.477**	**0.652**	0.545	**0.096**	0.484	0.827	0.970	0.129
Qwen3-32B	–	–	–	–	–	–	–	–	0.445	0.711	0.449	0.104	0.428	0.937	1.540	0.176
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.438	0.742	0.270	0.134	0.423	0.943	**0.649**	0.165
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.467	0.659	0.735	0.119	0.484	0.819	1.711	0.149
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.421	0.758	0.274	0.124	0.462	0.867	0.853	0.161
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.445	0.728	0.231	0.112	0.495	0.861	1.751	0.150
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.448*	*0.708*	*0.385*	*0.111*	*0.481*	*0.847*	*1.449*	*0.148*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.463	0.852	3.068	0.157	–	–	–	–	0.488	0.797	3.524	0.168
DeepSeek-V4-Flash	–	–	–	–	0.445	0.892	2.425	0.142	–	–	–	–	0.437	0.911	2.535	0.149
Mimo-v2.5	–	–	–	–	0.448	0.874	2.165	0.157	–	–	–	–	0.474	0.862	2.518	0.162
Mimo-v2.5-Pro	–	–	–	–	0.474	0.821	0.974	0.126	–	–	–	–	0.441	0.898	1.513	0.122
LongCat-2.0	–	–	–	–	0.474	0.820	2.906	0.158	–	–	–	–	0.480	0.821	2.969	0.164
GPT-4o-mini	–	–	–	–	0.401	0.958	1.535	0.171	–	–	–	–	0.412	0.955	1.627	0.185
GPT-5-mini	–	–	–	–	0.436	0.869	2.763	0.165	–	–	–	–	0.470	0.830	3.122	0.178
Qwen2.5-72B-Instruct	–	–	–	–	0.456	0.829	2.205	0.137	–	–	–	–	0.442	0.881	3.158	0.147
Qwen3-32B	–	–	–	–	0.420	0.957	2.105	0.183	–	–	–	–	0.409	1.033	2.356	0.189
Qwen3-30B-A3B	–	–	–	–	0.356	1.068	**0.887**	0.136	–	–	–	–	0.354	1.107	0.811	0.163
Llama-3.3-70B-Instruct	–	–	–	–	0.460	0.850	2.552	0.166	–	–	–	–	0.450	0.850	2.550	0.155
Gemma-3-27B-it	–	–	–	–	0.391	0.973	1.131	0.144	–	–	–	–	0.398	0.998	1.465	0.164
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.434	0.940	2.784	0.170	–	–	–	–	0.452	0.936	3.081	0.189
Avg. (PRA)	–	–	–	–	*0.435*	*0.900*	*2.115*	*0.155*	–	–	–	–	*0.439*	*0.914*	*2.402*	*0.164*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	**0.519**	**0.715**	2.617	0.164
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.481	0.834	2.107	0.139
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	**0.519**	0.769	2.207	0.161
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.496	0.747	1.192	0.142
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.501	0.784	2.753	0.177
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.424	0.965	1.561	0.177
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.477	0.795	2.604	0.172
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.504	0.762	1.634	0.124
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.440	0.900	0.932	0.152
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.391	1.041	**0.804**	0.141
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.468	0.858	1.895	0.158
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.454	0.890	1.607	0.150
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.469	0.864	1.708	0.159
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.472*	*0.840*	*1.817*	*0.155*
Training																
BERT	0.438	0.730	5.707	0.184	**0.507**	0.789	3.934	0.166	0.431	0.699	4.428	0.180	0.518	0.764	4.068	0.160
SFT (Qwen2.5-7B-Instruct)	0.435	0.740	2.567	0.173	**0.487**	**0.779**	3.517	0.162	0.424	0.723	2.118	0.173	0.505	**0.746**	3.694	0.159
SFT (Qwen3-8B)	0.438	0.710	2.589	0.167	0.468	0.802	2.255	0.146	0.423	0.740	2.219	0.170	0.481	0.769	2.210	0.138

Table 29: GSS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.454	**0.864**	11.472	0.303	0.439	0.908	11.890	0.318	0.437	0.922	11.839	0.297	0.431	0.892	12.094	0.323
DeepSeek-V4-Flash	0.433	0.960	11.959	0.327	0.440	0.962	11.878	0.332	0.404	1.111	12.582	0.340	0.448	0.907	11.718	0.323
Mimo-v2.5	0.419	0.923	12.284	0.314	0.430	0.938	12.113	0.327	0.420	0.951	12.217	0.307	0.429	0.938	12.146	0.332
Mimo-v2.5-Pro	0.421	0.894	12.238	0.314	0.434	0.940	12.011	0.332	0.427	0.950	12.073	0.313	0.425	0.940	12.237	0.335
LongCat-2.0	0.434	0.883	11.936	0.310	0.434	0.921	12.011	0.329	0.443	0.906	11.701	0.295	0.424	0.928	12.259	0.333
GPT-4o-mini	0.387	0.992	13.026	0.344	0.401	1.004	12.764	0.349	0.358	1.060	13.643	0.354	0.400	0.929	12.809	0.338
GPT-5-mini	0.416	0.958	12.354	0.322	0.431	0.908	12.071	0.318	0.423	0.934	12.157	0.310	0.406	0.949	12.688	0.340
Qwen2.5-72B-Instruct	0.388	0.957	13.003	0.326	0.428	0.912	12.161	0.323	0.410	0.999	12.450	0.312	0.410	0.909	12.583	0.324
Qwen3-32B	0.401	1.008	12.701	0.333	0.411	0.965	12.541	0.345	0.393	1.073	12.840	0.328	0.411	0.976	12.553	0.341
Qwen3-30B-A3B	0.366	1.061	13.513	0.358	0.378	1.028	13.311	0.366	0.360	1.133	13.613	0.359	0.394	0.984	12.959	0.350
Llama-3.3-70B-Instruct	0.441	0.970	11.774	0.326	0.439	0.928	11.896	0.324	0.424	1.013	12.127	0.335	0.424	0.952	12.259	0.334
Gemma-3-27B-it	0.393	0.939	12.887	0.335	0.417	0.931	12.408	0.329	0.412	0.918	12.414	0.312	0.393	0.957	12.967	0.344
Mistral-Small-3.1-24B-Instruct	0.405	0.968	12.609	0.322	0.430	0.874	12.101	0.312	0.401	1.007	12.666	0.317	0.427	0.901	12.199	0.325
Avg. (Zero-shot)	*0.412*	*0.952*	*12.443*	*0.326*	*0.424*	*0.940*	*12.243*	*0.331*	*0.409*	*0.998*	*12.486*	*0.322*	*0.417*	*0.935*	*12.421*	*0.334*
Persona																
DeepSeek-V4-Pro	0.452	0.912	4.014	0.186	0.473	0.803	3.745	0.181	0.435	0.926	2.108	0.135	0.483	0.766	3.653	0.168
DeepSeek-V4-Flash	**0.462**	0.879	3.752	0.192	0.476	0.800	3.769	0.190	0.438	0.916	2.148	0.155	0.488	0.770	3.927	0.179
Mimo-v2.5	0.430	0.933	3.073	0.189	0.478	0.799	3.353	0.179	0.444	0.893	2.069	0.149	0.467	0.811	3.605	0.177
Mimo-v2.5-Pro	0.437	0.917	**2.863**	**0.182**	0.479	0.797	3.257	0.173	0.448	0.878	1.670	0.138	0.484	0.777	3.231	0.169
LongCat-2.0	**0.461**	0.890	2.944	0.193	0.479	0.768	3.138	**0.167**	0.455	0.876	1.778	0.160	0.475	0.783	3.281	0.160
GPT-4o-mini	0.398	0.920	2.949	**0.182**	0.436	0.859	**1.993**	0.173	0.402	0.965	1.318	0.162	0.439	0.834	2.356	0.169
GPT-5-mini	0.431	0.894	2.931	**0.175**	0.457	0.823	3.480	0.174	0.424	0.891	1.548	0.150	0.456	0.820	3.250	0.177
Qwen2.5-72B-Instruct	0.404	0.928	3.349	0.201	0.446	0.817	3.116	0.182	0.410	0.934	1.597	0.164	0.444	0.804	2.950	0.175
Qwen3-32B	0.416	0.951	**2.859**	0.195	0.455	0.842	3.311	0.192	0.407	0.966	1.621	0.165	0.443	0.840	3.405	0.194
Qwen3-30B-A3B	0.403	0.973	4.164	0.207	0.415	0.874	2.988	0.213	0.403	0.977	2.160	0.175	0.422	0.866	3.420	0.207
Llama-3.3-70B-Instruct	0.454	**0.870**	3.330	0.184	0.464	0.797	3.850	0.191	0.451	0.874	2.071	0.164	0.465	0.790	4.347	0.197
Gemma-3-27B-it	0.381	0.997	3.886	0.229	0.416	0.870	3.059	0.208	0.382	0.994	1.978	0.196	0.425	0.852	3.421	0.202
Mistral-Small-3.1-24B-Instruct	0.407	0.945	3.093	0.215	0.448	0.839	3.774	0.207	0.427	0.947	1.738	0.170	0.445	0.827	3.634	0.191
Avg. (Persona)	*0.426*	*0.924*	*3.324*	*0.195*	*0.456*	*0.822*	*3.295*	*0.187*	*0.425*	*0.926*	*1.831*	*0.160*	*0.457*	*0.811*	*3.422*	*0.182*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.461	0.888	1.998	**0.119**	0.486	0.769	3.486	0.163
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.450	0.905	1.843	0.150	0.499	0.762	3.238	0.168
Mimo-v2.5	–	–	–	–	–	–	–	–	0.462	0.878	1.562	0.144	0.493	0.772	2.883	0.165
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	**0.463**	0.857	**1.286**	**0.133**	0.497	0.761	2.917	0.162
LongCat-2.0	–	–	–	–	–	–	–	–	**0.468**	**0.855**	1.484	0.157	0.482	0.781	3.206	0.173
GPT-4o-mini	–	–	–	–	–	–	–	–	0.417	0.977	1.668	0.167	0.449	0.856	2.553	0.180
GPT-5-mini	–	–	–	–	–	–	–	–	0.426	0.925	**0.816**	0.135	0.460	0.815	2.803	0.167
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.432	0.905	1.537	0.162	0.470	0.771	2.576	0.169
Qwen3-32B	–	–	–	–	–	–	–	–	0.414	0.965	1.445	0.162	0.441	0.842	3.285	0.192
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.407	0.993	1.589	0.164	0.424	0.914	3.357	0.197
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.455	**0.844**	2.168	0.143	0.481	0.764	3.762	0.181
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.397	0.980	1.621	0.190	0.436	0.846	2.789	0.198
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.427	0.981	1.787	0.178	0.455	0.849	3.542	0.188
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.437*	*0.919*	*1.600*	*0.154*	*0.467*	*0.808*	*3.107*	*0.177*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.478	0.804	4.474	0.202	–	–	–	–	0.497	0.763	4.563	0.189
DeepSeek-V4-Flash	–	–	–	–	0.481	0.794	3.984	0.195	–	–	–	–	0.485	0.775	3.971	0.179
Mimo-v2.5	–	–	–	–	0.471	0.807	2.968	**0.164**	–	–	–	–	0.461	0.812	2.984	**0.158**
Mimo-v2.5-Pro	–	–	–	–	**0.483**	0.803	3.479	0.183	–	–	–	–	0.491	0.752	3.535	0.169
LongCat-2.0	–	–	–	–	0.481	0.793	4.662	0.198	–	–	–	–	0.497	0.778	4.341	0.187
GPT-4o-mini	–	–	–	–	0.462	0.846	3.806	0.194	–	–	–	–	0.469	0.823	3.688	0.187
GPT-5-mini	–	–	–	–	0.474	0.777	3.664	0.184	–	–	–	–	0.479	0.791	3.931	0.184
Qwen2.5-72B-Instruct	–	–	–	–	0.472	0.793	4.423	0.197	–	–	–	–	0.485	0.755	4.043	0.183
Qwen3-32B	–	–	–	–	0.464	0.845	3.797	0.202	–	–	–	–	0.467	0.827	4.034	0.203
Qwen3-30B-A3B	–	–	–	–	0.395	1.001	**1.905**	0.172	–	–	–	–	0.394	1.001	**2.064**	0.177
Llama-3.3-70B-Instruct	–	–	–	–	0.471	0.814	4.313	0.196	–	–	–	–	0.484	0.789	4.395	0.202
Gemma-3-27B-it	–	–	–	–	0.404	0.936	2.038	0.173	–	–	–	–	0.405	0.926	2.076	0.175
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.459	0.856	4.973	0.218	–	–	–	–	0.473	0.809	4.774	0.201
Avg. (PRA)	–	–	–	–	*0.461*	*0.836*	*3.730*	*0.191*	–	–	–	–	*0.468*	*0.815*	*3.723*	*0.184*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.500	0.740	3.674	0.170
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	**0.507**	0.753	3.717	0.176
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.490	0.771	2.710	**0.157**
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.506	0.738	3.666	0.170
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.497	0.772	3.525	0.170
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.476	0.808	3.042	0.165
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.478	0.776	2.838	**0.158**
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.478	0.769	3.355	0.175
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.453	0.840	3.227	0.187
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.394	0.988	**2.016**	0.180
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.486	0.780	3.771	0.179
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.443	0.864	2.499	0.186
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.463	0.827	3.793	0.192
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.475*	*0.802*	*3.218*	*0.174*
Training																
BERT	0.369	1.158	10.218	0.348	**0.496**	**0.756**	4.938	0.195	0.397	1.130	4.336	0.270	**0.515**	**0.733**	5.988	0.211
SFT (Qwen2.5-7B-Instruct)	0.417	0.955	4.327	0.228	**0.496**	**0.755**	4.046	0.186	0.426	0.942	2.563	0.195	0.505	**0.711**	4.159	0.181
SFT (Qwen3-8B)	0.405	0.986	5.272	0.242	0.476	0.794	4.185	0.193	0.427	0.977	3.825	0.208	0.482	0.770	4.272	0.193

Table 30: ISSP results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	0.282	1.085	15.301	0.276	0.308	1.172	14.708	0.267	0.397	0.840	12.652	0.213	0.296	1.319	14.948	0.299
DeepSeek-V4-Flash	0.250	1.186	16.036	0.301	0.257	1.244	15.886	0.293	0.389	0.929	12.830	0.236	0.303	1.204	14.783	0.272
Mimo-v2.5	0.309	1.096	14.688	0.277	0.336	0.908	14.066	0.223	0.336	1.030	14.046	0.260	0.279	1.295	15.345	0.302
Mimo-v2.5-Pro	0.277	1.117	15.423	0.282	0.341	0.910	13.959	0.224	0.273	1.120	15.500	0.283	0.312	1.135	14.584	0.277
LongCat-2.0	0.319	0.963	14.443	0.243	0.317	1.029	14.494	0.253	0.430	0.778	11.880	0.197	0.316	1.115	14.485	0.274
GPT-4o-mini	0.356	0.894	13.586	0.228	0.259	1.334	15.833	0.312	0.325	1.050	14.313	0.265	0.267	1.372	15.610	0.322
GPT-5-mini	0.330	1.005	14.198	0.254	0.337	1.101	14.039	0.249	0.313	0.987	14.580	0.249	0.276	1.300	15.412	0.293
Qwen2.5-72B-Instruct	0.271	1.021	15.546	0.258	0.359	**0.859**	13.530	0.211	0.361	0.856	13.482	0.216	0.306	1.221	14.717	0.281
Qwen3-32B	0.293	1.191	15.056	0.302	0.290	1.267	15.137	0.294	0.276	1.206	15.441	0.304	0.250	1.500	16.007	0.344
Qwen3-30B-A3B	0.293	1.144	15.056	0.289	0.252	1.377	15.993	0.322	0.334	1.061	14.105	0.268	0.263	1.453	15.709	0.333
Llama-3.3-70B-Instruct	0.362	0.872	13.464	0.221	0.312	1.167	14.628	0.268	0.384	0.874	12.948	0.221	0.287	1.290	15.147	0.302
Gemma-3-27B-it	0.303	1.048	14.811	0.258	0.345	0.915	13.851	0.225	0.402	0.831	12.533	0.210	0.309	1.027	14.651	0.252
Mistral-Small-3.1-24B-Instruct	0.383	**0.771**	12.974	0.195	0.326	1.103	14.307	0.250	**0.439**	**0.729**	11.672	0.185	0.290	1.221	15.081	0.275
Avg. (Zero-shot)	*0.310*	*1.030*	*14.660*	*0.260*	*0.311*	*1.107*	*14.649*	*0.261*	*0.358*	*0.945*	*13.537*	*0.239*	*0.289*	*1.266*	*15.114*	*0.294*
Persona																
DeepSeek-V4-Pro	0.351	0.888	3.978	**0.117**	0.330	0.964	3.790	0.136	0.399	0.818	2.689	**0.077**	0.342	0.987	4.549	0.148
DeepSeek-V4-Flash	0.362	0.872	5.778	0.127	0.350	0.892	3.368	0.131	0.379	0.854	1.297	0.094	0.332	0.980	3.945	0.143
Mimo-v2.5	0.298	1.069	3.681	0.166	0.314	1.020	3.440	0.129	0.365	0.924	1.089	0.127	0.313	1.079	4.112	0.158
Mimo-v2.5-Pro	0.383	0.910	3.541	0.123	0.344	0.930	1.978	0.118	0.384	0.871	1.046	0.084	0.349	0.983	3.831	0.140
LongCat-2.0	0.346	0.920	2.186	0.137	0.347	0.897	4.316	0.136	0.379	0.866	2.038	0.140	0.330	0.984	5.155	0.174
GPT-4o-mini	0.330	0.957	2.635	0.119	0.330	0.990	2.805	0.129	0.347	0.981	0.909	0.108	0.302	1.086	3.479	0.156
GPT-5-mini	0.367	0.915	3.953	0.153	0.333	0.921	3.693	0.130	0.388	0.841	1.297	0.126	0.306	1.076	4.416	0.165
Qwen2.5-72B-Instruct	0.335	0.883	3.248	0.134	0.337	0.937	3.256	0.132	0.349	0.914	1.762	0.136	0.319	1.027	3.790	0.155
Qwen3-32B	0.271	1.048	3.550	0.150	0.324	1.050	4.457	0.171	0.331	0.974	1.670	0.121	0.259	1.197	4.283	0.191
Qwen3-30B-A3B	0.287	1.122	**1.946**	0.152	0.288	1.120	2.307	0.153	0.285	1.164	1.537	0.139	0.266	1.254	3.264	0.186
Llama-3.3-70B-Instruct	**0.399**	**0.819**	5.123	0.127	0.350	0.945	3.497	0.131	0.388	0.821	2.191	0.104	0.325	1.039	4.562	0.166
Gemma-3-27B-it	**0.388**	0.888	3.057	0.125	0.315	1.003	2.417	0.134	0.334	0.965	0.791	0.121	0.302	1.069	3.261	0.151
Mistral-Small-3.1-24B-Instruct	0.372	0.878	**1.896**	**0.093**	0.317	0.993	3.274	0.145	0.356	0.876	1.344	0.102	0.290	1.152	4.852	0.182
Avg. (Persona)	*0.345*	*0.936*	*3.429*	*0.133*	*0.329*	*0.974*	*3.277*	*0.136*	*0.360*	*0.913*	*1.512*	*0.114*	*0.310*	*1.070*	*4.115*	*0.163*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	**0.432**	**0.771**	1.173	0.090	**0.394**	0.924	2.768	**0.121**
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.396	0.823	**0.345**	**0.068**	0.388	0.943	2.645	**0.125**
Mimo-v2.5	–	–	–	–	–	–	–	–	0.393	0.862	0.940	0.104	0.381	0.999	3.550	0.163
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.409	0.829	0.595	0.088	**0.404**	0.948	2.799	0.139
LongCat-2.0	–	–	–	–	–	–	–	–	0.409	0.852	0.492	0.100	0.371	0.968	3.747	0.145
GPT-4o-mini	–	–	–	–	–	–	–	–	0.379	1.001	0.531	0.146	0.320	1.131	3.391	0.180
GPT-5-mini	–	–	–	–	–	–	–	–	0.376	0.896	0.562	0.120	0.333	1.057	4.524	0.178
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.403	0.802	1.476	0.115	0.346	1.004	3.630	0.163
Qwen3-32B	–	–	–	–	–	–	–	–	0.392	0.888	0.494	0.105	0.332	1.062	3.519	0.161
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.341	1.050	**0.264**	0.123	0.303	1.204	2.272	0.149
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.425	0.794	1.546	0.081	0.374	0.981	3.624	0.146
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.365	0.933	0.989	0.102	0.356	1.039	2.683	0.141
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.372	0.933	0.943	0.128	0.341	1.132	3.735	0.170
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.392*	*0.879*	*0.796*	*0.105*	*0.357*	*1.030*	*3.299*	*0.152*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.344	0.900	4.524	0.142	–	–	–	–	0.349	1.006	5.731	0.169
DeepSeek-V4-Flash	–	–	–	–	0.323	1.045	1.994	**0.114**	–	–	–	–	0.319	1.111	2.837	0.145
Mimo-v2.5	–	–	–	–	0.331	1.014	2.126	0.115	–	–	–	–	0.307	1.111	3.606	0.161
Mimo-v2.5-Pro	–	–	–	–	0.347	0.886	3.009	**0.112**	–	–	–	–	0.338	1.013	4.461	0.161
LongCat-2.0	–	–	–	–	**0.383**	**0.860**	5.874	0.143	–	–	–	–	0.368	0.971	6.821	0.178
GPT-4o-mini	–	–	–	–	0.306	1.062	2.701	0.124	–	–	–	–	0.300	1.101	2.863	0.136
GPT-5-mini	–	–	–	–	0.356	0.906	3.550	0.124	–	–	–	–	0.338	1.059	4.827	0.175
Qwen2.5-72B-Instruct	–	–	–	–	0.334	0.943	5.136	0.147	–	–	–	–	0.342	1.029	5.714	0.173
Qwen3-32B	–	–	–	–	0.315	1.076	3.185	0.158	–	–	–	–	0.283	1.162	4.171	0.193
Qwen3-30B-A3B	–	–	–	–	0.230	1.364	**1.075**	0.168	–	–	–	–	0.239	1.443	**1.414**	0.164
Llama-3.3-70B-Instruct	–	–	–	–	0.348	0.944	4.189	0.132	–	–	–	–	0.335	1.034	4.712	0.172
Gemma-3-27B-it	–	–	–	–	0.276	1.192	**1.696**	0.128	–	–	–	–	0.283	1.240	2.375	0.157
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.334	0.987	3.985	0.139	–	–	–	–	0.330	1.076	4.319	0.173
Avg. (PRA)	–	–	–	–	*0.325*	*1.014*	*3.311*	*0.134*	–	–	–	–	*0.318*	*1.104*	*4.142*	*0.166*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.386	**0.920**	3.703	0.150
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.366	1.013	3.981	0.148
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.351	1.039	3.596	0.178
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.381	0.954	3.969	0.147
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.376	0.958	4.702	0.156
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.284	1.131	2.736	0.157
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.329	1.065	3.331	0.152
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.369	1.010	4.033	0.164
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.292	1.190	4.374	0.195
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.257	1.402	**1.799**	0.173
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	**0.394**	0.953	3.946	0.147
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.300	1.174	2.170	0.163
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.349	1.102	3.530	0.177
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.341*	*1.070*	*3.528*	*0.162*
Training																
BERT	0.383	0.920	8.830	0.203	0.371	0.892	9.471	0.200	0.415	0.920	7.464	0.181	0.378	**0.891**	8.320	0.175
SFT (Qwen2.5-7B-Instruct)	0.335	0.989	3.356	0.123	**0.378**	0.913	2.425	0.117	0.361	0.901	1.921	0.109	0.318	1.023	4.380	0.154
SFT (Qwen3-8B)	0.330	1.048	2.378	0.129	0.300	1.045	3.222	0.143	0.365	0.945	0.796	0.092	0.307	1.091	4.336	0.167

Table 31: WVS results across four evaluation settings. **Dark green** indicates best performance, **sage green** indicates second best.

Model	Test-UU				Test-US				Test-SU				Test-SS			
	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD	Acc.	MAE	KL	EMD
Zero-shot																
DeepSeek-V4-Pro	**0.497**	1.146	10.414	0.251	0.386	1.348	12.964	0.335	0.445	1.305	11.591	0.301	0.399	1.283	12.684	0.336
DeepSeek-V4-Flash	0.436	1.367	11.813	0.314	0.384	1.373	12.997	0.351	0.443	1.259	11.627	0.316	0.412	1.257	12.374	0.329
Mimo-v2.5	**0.494**	1.167	10.486	0.262	0.400	1.315	12.637	0.336	**0.485**	1.146	10.654	0.273	0.396	1.299	12.742	0.340
Mimo-v2.5-Pro	0.491	1.176	10.558	0.264	0.392	1.356	12.822	0.340	0.476	1.180	10.883	0.285	0.408	1.324	12.477	0.340
LongCat-2.0	0.439	1.378	11.755	0.308	0.391	1.293	12.855	0.327	0.439	1.261	11.725	0.320	0.400	1.275	12.657	0.331
GPT-4o-mini	0.438	1.242	11.784	0.295	0.352	1.456	13.749	0.369	0.419	1.201	12.178	0.309	0.365	1.370	13.455	0.358
GPT-5-mini	0.476	1.325	10.904	0.279	0.380	1.350	13.091	0.337	0.431	1.484	11.907	0.319	0.386	1.323	12.971	0.343
Qwen2.5-72B-Instruct	0.465	1.218	11.149	0.279	0.390	1.275	12.862	0.326	0.470	1.218	11.012	0.299	0.407	1.222	12.486	0.318
Qwen3-32B	0.443	1.281	11.654	0.297	0.372	1.419	13.280	0.354	0.467	1.193	11.076	0.289	0.388	1.397	12.935	0.357
Qwen3-30B-A3B	0.408	1.330	12.461	0.339	0.324	1.509	14.384	0.401	0.405	1.332	12.511	0.338	0.341	1.423	14.002	0.388
Llama-3.3-70B-Instruct	0.443	1.224	11.668	0.296	0.358	1.357	13.596	0.365	0.464	1.166	11.146	0.292	0.368	1.397	13.397	0.379
Gemma-3-27B-it	0.465	1.247	11.149	0.269	0.384	1.330	13.008	0.332	0.455	1.259	11.367	0.291	0.389	1.311	12.899	0.338
Mistral-Small-3.1-24B-Instruct	0.450	1.224	11.495	0.273	0.405	1.231	12.528	0.312	0.459	1.100	11.265	0.270	0.384	1.317	13.015	0.334
Avg. (Zero-shot)	*0.457*	*1.256*	*11.330*	*0.287*	*0.378*	*1.355*	*13.136*	*0.345*	*0.451*	*1.239*	*11.457*	*0.300*	*0.388*	*1.323*	*12.930*	*0.345*
Persona																
DeepSeek-V4-Pro	0.463	1.185	4.649	0.176	0.424	1.179	4.352	0.183	0.476	1.170	3.214	0.184	0.419	1.176	4.295	0.201
DeepSeek-V4-Flash	0.470	1.180	4.447	0.175	0.440	1.173	3.840	0.179	0.475	1.142	3.171	0.171	0.438	1.176	4.359	0.187
Mimo-v2.5	0.487	1.131	4.151	0.160	0.435	1.159	3.354	0.183	0.475	1.120	2.807	0.172	0.426	1.175	4.095	0.198
Mimo-v2.5-Pro	0.475	1.178	4.154	0.168	0.434	1.188	3.634	**0.172**	0.468	1.163	3.217	0.173	0.436	1.179	3.990	0.192
LongCat-2.0	0.459	1.176	4.032	0.167	0.437	1.114	3.366	0.177	0.467	1.144	3.032	0.167	0.424	1.144	4.052	0.190
GPT-4o-mini	0.448	1.225	3.938	0.179	0.386	1.272	2.864	0.202	0.455	1.154	2.275	0.188	0.378	1.272	3.042	0.209
GPT-5-mini	0.453	**1.115**	**3.081**	**0.150**	0.422	**1.102**	3.451	0.175	0.463	**1.093**	**1.965**	**0.143**	0.405	1.146	3.692	0.191
Qwen2.5-72B-Instruct	0.465	1.147	3.831	0.167	0.419	1.187	3.226	0.189	0.465	1.118	2.880	0.175	0.407	1.208	3.814	0.207
Qwen3-32B	0.463	1.183	5.020	0.169	0.410	1.233	3.812	0.201	0.460	1.149	3.700	0.184	0.408	1.213	4.244	0.211
Qwen3-30B-A3B	0.435	1.229	**3.083**	0.170	0.391	1.237	2.988	0.199	0.436	1.213	1.976	0.173	0.388	1.248	3.483	0.211
Llama-3.3-70B-Instruct	0.475	**1.125**	4.592	0.161	0.427	1.137	4.182	0.186	0.475	1.098	3.416	0.171	0.428	1.154	4.885	0.199
Gemma-3-27B-it	0.427	1.227	5.081	0.210	0.382	1.231	3.375	0.196	0.426	1.219	3.579	0.221	0.384	1.232	3.775	0.207
Mistral-Small-3.1-24B-Instruct	0.460	1.153	4.615	0.188	0.407	1.242	4.169	0.202	0.474	1.139	4.319	0.199	0.403	1.251	4.455	0.217
Avg. (Persona)	*0.460*	*1.173*	*4.206*	*0.172*	*0.416*	*1.189*	*3.586*	*0.188*	*0.463*	*1.148*	*3.042*	*0.178*	*0.411*	*1.198*	*4.014*	*0.202*
HQA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	0.475	1.192	2.660	0.172	0.434	1.172	3.830	0.198
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	0.472	1.159	3.362	0.179	0.446	1.158	3.525	0.185
Mimo-v2.5	–	–	–	–	–	–	–	–	**0.482**	1.125	3.360	0.163	0.451	1.134	3.802	0.195
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	0.475	1.182	3.391	0.178	0.448	1.153	3.842	0.189
LongCat-2.0	–	–	–	–	–	–	–	–	0.463	1.162	2.829	0.182	0.451	1.111	3.634	0.186
GPT-4o-mini	–	–	–	–	–	–	–	–	0.449	1.250	2.949	0.206	0.400	1.273	3.562	0.229
GPT-5-mini	–	–	–	–	–	–	–	–	0.473	**1.088**	2.440	**0.152**	0.412	1.146	3.578	0.189
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	0.470	1.134	3.349	0.183	0.417	1.194	3.974	0.209
Qwen3-32B	–	–	–	–	–	–	–	–	0.467	1.170	3.504	0.199	0.420	1.193	4.062	0.209
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	0.445	1.232	**1.678**	0.168	0.405	1.230	3.257	0.211
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	0.474	1.114	2.911	0.180	0.437	1.121	4.026	0.192
Gemma-3-27B-it	–	–	–	–	–	–	–	–	0.450	1.218	2.938	0.192	0.396	1.227	3.255	0.200
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	0.456	1.207	4.167	0.227	0.420	1.262	4.033	0.223
Avg. (HQA)	–	–	–	–	–	–	–	–	*0.466*	*1.172*	*3.041*	*0.183*	*0.426*	*1.183*	*3.722*	*0.201*
PRA																
DeepSeek-V4-Pro	–	–	–	–	0.447	1.184	3.953	0.190	–	–	–	–	0.453	1.195	4.579	0.205
DeepSeek-V4-Flash	–	–	–	–	0.448	1.194	3.886	0.185	–	–	–	–	0.448	1.200	4.066	0.196
Mimo-v2.5	–	–	–	–	0.441	1.210	3.739	0.185	–	–	–	–	0.447	1.188	3.885	0.196
Mimo-v2.5-Pro	–	–	–	–	0.444	1.206	3.330	**0.171**	–	–	–	–	0.447	1.211	3.500	0.180
LongCat-2.0	–	–	–	–	0.457	1.161	4.089	0.186	–	–	–	–	0.465	1.174	4.656	0.197
GPT-4o-mini	–	–	–	–	0.430	1.228	2.848	0.175	–	–	–	–	0.424	1.239	3.030	0.178
GPT-5-mini	–	–	–	–	0.437	1.163	2.913	0.174	–	–	–	–	0.431	1.208	3.578	0.186
Qwen2.5-72B-Instruct	–	–	–	–	0.434	1.213	3.830	0.192	–	–	–	–	0.431	1.238	4.398	0.210
Qwen3-32B	–	–	–	–	0.422	1.253	3.655	0.207	–	–	–	–	0.430	1.234	3.899	0.217
Qwen3-30B-A3B	–	–	–	–	0.363	1.356	**2.272**	0.198	–	–	–	–	0.374	1.374	**2.236**	0.201
Llama-3.3-70B-Instruct	–	–	–	–	0.431	1.199	4.533	0.208	–	–	–	–	0.434	1.177	4.808	0.215
Gemma-3-27B-it	–	–	–	–	0.390	1.328	**2.498**	0.176	–	–	–	–	0.409	1.287	2.682	**0.165**
Mistral-Small-3.1-24B-Instruct	–	–	–	–	0.436	1.239	4.369	0.197	–	–	–	–	0.447	1.222	4.818	0.206
Avg. (PRA)	–	–	–	–	*0.429*	*1.226*	*3.532*	*0.188*	–	–	–	–	*0.434*	*1.227*	*3.856*	*0.196*
DACA																
DeepSeek-V4-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.461	1.153	4.976	0.218
DeepSeek-V4-Flash	–	–	–	–	–	–	–	–	–	–	–	–	0.448	1.202	4.098	0.196
Mimo-v2.5	–	–	–	–	–	–	–	–	–	–	–	–	0.448	1.177	3.654	0.201
Mimo-v2.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	0.459	1.166	3.757	0.186
LongCat-2.0	–	–	–	–	–	–	–	–	–	–	–	–	0.468	1.142	4.380	0.195
GPT-4o-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.423	1.285	3.773	0.202
GPT-5-mini	–	–	–	–	–	–	–	–	–	–	–	–	0.412	1.250	2.881	0.183
Qwen2.5-72B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.443	1.203	4.764	0.220
Qwen3-32B	–	–	–	–	–	–	–	–	–	–	–	–	0.433	1.229	3.754	0.207
Qwen3-30B-A3B	–	–	–	–	–	–	–	–	–	–	–	–	0.387	1.351	**2.294**	0.191
Llama-3.3-70B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.441	1.129	4.111	0.199
Gemma-3-27B-it	–	–	–	–	–	–	–	–	–	–	–	–	0.423	1.262	3.777	0.197
Mistral-Small-3.1-24B-Instruct	–	–	–	–	–	–	–	–	–	–	–	–	0.439	1.256	4.665	0.219
Avg. (DACA)	–	–	–	–	–	–	–	–	–	–	–	–	*0.437*	*1.216*	*3.914*	*0.201*
Training																
BERT	0.398	1.288	11.271	0.320	**0.477**	**1.063**	5.364	0.200	0.474	1.151	2.155	0.204	**0.530**	**0.990**	3.705	**0.173**
SFT (Qwen2.5-7B-Instruct)	0.466	1.158	5.176	0.188	**0.482**	1.115	3.853	0.179	0.477	1.120	3.796	0.183	**0.487**	**1.083**	4.595	0.192
SFT (Qwen3-8B)	0.482	1.161	4.782	**0.158**	0.473	1.107	4.430	0.184	0.481	1.135	4.378	0.182	0.480	1.102	4.848	0.192

J LIMITATIONS AND FUTURE WORK

Unverifiable training-data contamination. All 15 source surveys are publicly released, while the training corpora and procedures of both API-based and open-weight models are only partially documented or undisclosed. We therefore cannot determine whether, or to what extent, an evaluated model encountered survey questions, documentation, or related response data during pretraining or post-training. We control the evaluation protocol at test time by using identical splits, personas, questions, prompts, decoding settings, option shuffling, answer parsing, and metrics across models. However, these controls do not eliminate the possibility that models differ in their prior exposure to the underlying surveys, so differential training-data contamination remains an unresolved confound in cross-model comparison. The fact that even the strongest evaluated models reach only 40–55% accuracy suggests that any such exposure does not make the task trivial, but it does not rule out smaller effects on absolute performance or model rankings. Future work should incorporate evaluation data that tested models could not have encountered during training, such as newly collected or otherwise access-restricted survey data, to provide stronger contamination controls.