

Alyas: Evidence-Grounded Synthetic Populations for Simulated Social and Market Research

Luigi Petrucco, Samuel Bortolin, Elena Leonelli, Eric Suardi,
Elena Zorzi, Daniele Miorandi, Davide Piazza

Affiant S.r.l., Trento, Italy

Abstract

Large language models are increasingly used in social-science and market research to stand in for human respondents, exploring questions ahead of fieldwork or piloting instruments. Existing systems address one facet at a time (a population’s demographics, an individual’s profile, or an answer distribution elicited from a single prompt), and few serve users without social-science training. We present ALYAS, a research-simulation platform in which an agentic methodologist assistant guides the user from a research goal to a study design, laid out as a branching graph of research artifacts, and simulates its field phase over a synthetic population grounded in institutional statistical sources. On held-out Italian and US microdata, the generated populations reconstruct demographic structure the engine was not given, and, on the Italian microdata, profile-conditioned respondents answer close to the real subgroups they simulate.

1 Introduction

Since Argyle et al. (2023) showed that a language model conditioned on real respondent profiles can reproduce marginal and correlational structure of human survey samples, LLM-simulated respondents have become an active research area and industry practice (Park et al., 2024; Maier et al., 2025). A parallel reliability literature documents where such simulations diverge from human data: misaligned subgroup opinion distributions (Santurkar et al., 2023), prompt and option-order sensitivity (Bisbee et al., 2024), and systematic distortions (Zhou et al., 2025a; Paglieri et al., 2026). Hullman et al. (2026) argue that simulated-respondent claims require both a defensible target population and evidence that the simulation tracks it; most current systems meet these conditions only partially.

Prior systems ground the simulated sample to varying degrees: Argyle et al. (2023) condition on

profiles resampled from an existing survey; SocioVerse (Zhang et al., 2025) aligns a social-media-derived user pool to census and poll marginals; demographic-tree generation (Chen et al., 2026) models a topic-conditioned joint over demographics with LLM-estimated conditionals. In all three, the population either depends on survey data collected for another purpose or rests on inferred rather than observed structure, and none keeps a per-attribute record linking each marginal or correlation to the evidence that justifies it.

ALYAS is a web-based research-simulation platform pairing an agentic *methodologist assistant* with a *population engine*, and it addresses both sides: the assistant guides the user from a rough brief through a branching research design (quantitative, qualitative, or mixed) to a simulated field phase and a report, over a population grounded in institutional statistical sources (§3). The construction is inspectable: the user can see where each marginal and correlation came from (observed, estimated, or defaulted) and how closely the realized sample matches it.

Our contributions are as follows:

1. An end-to-end research-simulation workflow in which an agentic methodologist assistant develops the brief with the user, plans the study as a directed acyclic graph of research artifacts, and builds each node in turn, converging parallel lines of inquiry into a report whose claims do not outrun the design (§2).
2. A design-composition layer that assembles standard designs and instruments from the survey-methodology literature (convergent and explanatory-sequential designs, diverse item types, discussion guides) together with the user rather than fixed in advance (§2).
3. A population engine that sources marginals and pairwise joints from an open set of

institutional statistical sources, with web search and model estimation as documented fallbacks; reconstructs a global joint under provenance-weighted constraints; samples deterministically; and verifies every asserted relationship on the realized sample, reporting each attribute’s evidential basis (§3).

ALYAS is intended for researchers who simulate human respondents with language models and for social scientists who run surveys, to explore a question ahead of fieldwork or to pilot an instrument before it is fielded.

2 System Overview

ALYAS assists users without survey-methodology training in running a social or market-research study with simulated respondents. Its interface is a chat with an assistant, the methodologist, which guides the user through the design of the study, executes it, and stays available to explain its choices and take feedback that steers or extends the research. Behind the chat is a single tool-using agent, the *orchestrator*, prompted as a senior research methodologist (Figure 1); its tools materialize and manipulate the research elements (target population, survey instrument, response dataset), which appear in the workspace beside the chat as the conversation progresses (Figure 2). To begin, the user can provide a document of study requirements, which the methodologist probes with clarification questions; alternatively, its questions bring the research question into focus and operationalize it as a *brief*: a short statement of the study objective and the decision it informs. Three standing rules govern the agent: tie every proposal to that decision, invent no data, and keep reported claims within what the design supports, even under user pressure.

The study as a graph. From the brief, the methodologist designs the steps needed to answer the study’s objectives, drawing on a catalog of standard methods, and maps this process onto a directed acyclic graph executed as a workflow: nodes are briefs, populations, instruments, response datasets, and reports; edges record what each artifact derives from. These dependencies stay in force (data collection, for instance, requires both a population and a questionnaire) as the methodologist drafts and edits the study under the user’s guidance: a node is built once its dependencies exist, and editing one whose dependents

are already materialized marks those as stale, to be updated or rediscussed with the user. Nodes persist across such revisions, each edit stored with the exchange that motivated it, so the path from brief to conclusions can be retraced. Node types may recur, so a question can branch and reconverge: a goal with distinct sub-questions branches into one *sub-brief* for each sub-question, a purposive subset of one sample can be carried forward as a new population (anchoring a qualitative follow-up to a quantitative sample), and converging lines end in a single report that synthesizes their datasets.

Designs and instruments. The methodologist proposes a design (qualitative, quantitative, convergent, or explanatory-sequential) and instantiates it with methods from a catalog, a target sample size, and a plan for combining its lines of evidence, all user-revisable. Instruments are structured questionnaires over eleven item types (single-choice and Likert through ranking, conjoint (attribute trade-off choices), and MaxDiff (best–worst scaling)) or discussion guides for in-depth interviews; a simulated field phase (§4) administers each instrument and feeds a design-aware report.

The population. The target population is produced by a separate population engine: a harness in which a second agent turns a natural-language description of the population and its attributes of interest into a calibrated population profile grounded in reference statistics, web search, and documented estimates (§3). In the graph the population is a node: the methodologist proposes a frame (who is sampled, the scope, the variables), the user confirms it, and the engine calibrates it, reporting per variable the source of its statistics and the fit of the realized sample; later changes go through an explicit, recorded recalibration.

Implementation. ALYAS is built on LLMs but does not depend on a particular one: it can be configured to run with different models, to leverage the state of the art. Orchestration is demanding, as the agent has to reason upon and interact with a complex graph of artifacts across a long conversation. To meet this demand, at the time of evaluation ALYAS ran Sonnet 5, delegating web search to Perplexity Sonar and numeric estimation to Gemini 3.5 Flash.

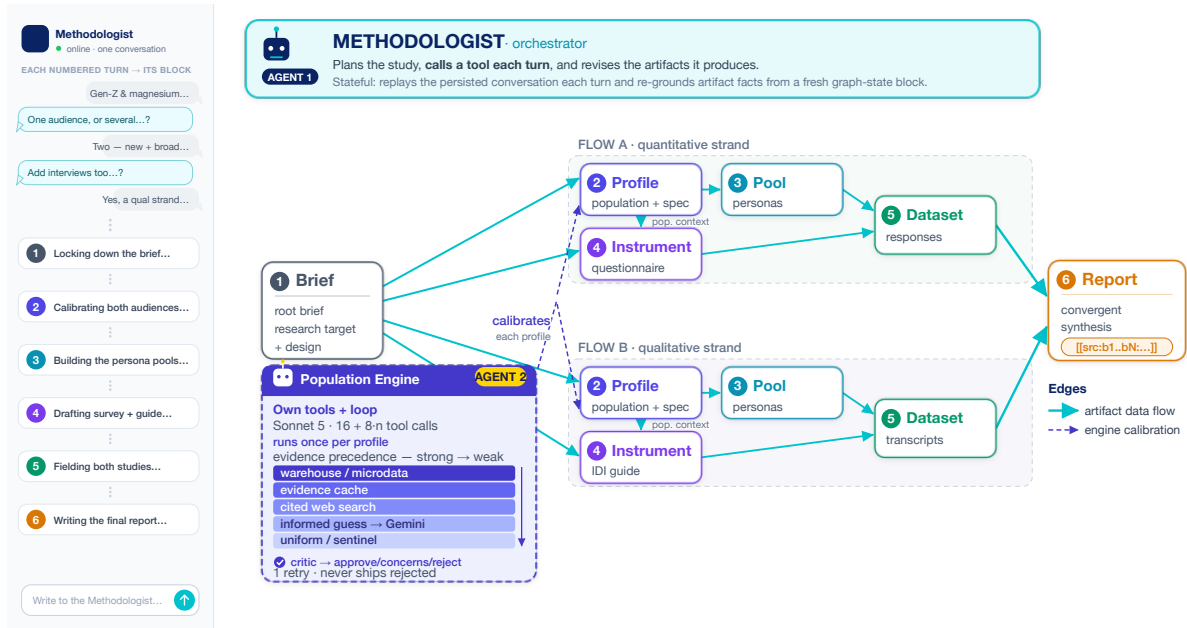


Figure 1: A simplified overview of the ALYAS methodologist orchestrating a branching graph of research artifacts.

3 The Population Engine

The population engine turns the confirmed frame into a synthetic population: attribute rows whose composition and dependence structure follow externally sourced statistics.

Sourcing. For each attribute and attribute pair, an agent seeks evidence in a documented fallback order. It first consults a reference warehouse of institutional statistical sources (such as Eurostat and ISTAT aggregates, US Census microdata, and major social surveys); failing that, it spawns a dedicated search agent that queries the web in natural language and returns cited figures; next comes a model-generated estimate, and finally an explicit low-trust default. Every sourced quantity records its provenance, carried into the reconstruction.

Reconstruction and sampling. The target joint over the coupled attributes is the penalized maximum-entropy distribution, in the lineage of iterative proportional fitting and minimum-discrimination-information estimation (Deming and Stephan, 1940; Csiszár, 1975): starting from independence, each constraint is matched in proportion to the trust its provenance supports, entering as an exponent on each constraint’s fitting update (0.9 warehouse, 0.7 web, 0.5–0.55 model estimates, 0.3 defaults), so disagreeing sources are balanced with the higher-trust one dominating, iteration stops at a fixed tolerance, and independence remains the fallback where no joint evi-

dence exists. Since the full cell table over all coupled attributes is too large to enumerate, the population is drawn by a factored Gibbs-style sampler targeting the same distribution, with a calibration pass correcting the small shift dense cross-couplings induce, preserving every declared association. Sampling is seeded for reproducibility.

Verification. After sampling, every asserted marginal and joint is re-computed from the realized rows and compared to its target within sample-size-aware tolerances, yielding a coverage map labeling each relationship as evidence-backed, estimated, or left at the maximum-entropy default. An independent critic agent reviews coverage, provenance health, plausibility, recency, and joint coherence, and rejects when a concern is blocking and actionable, reopening the search loop once.

4 Simulated Fieldwork

The field phase materializes the sampled population as respondents and collects their answers; it is LLM-driven and inherits the model’s limitations.

Persona materialization. Each sampled attribute row is expanded by a language model into a narrative persona (occupation, daily routine, communication style), under the constraint that every sampled attribute is preserved exactly: categorical codes become lived detail (an age bracket becomes a specific age), never altered, and no attribute is in-

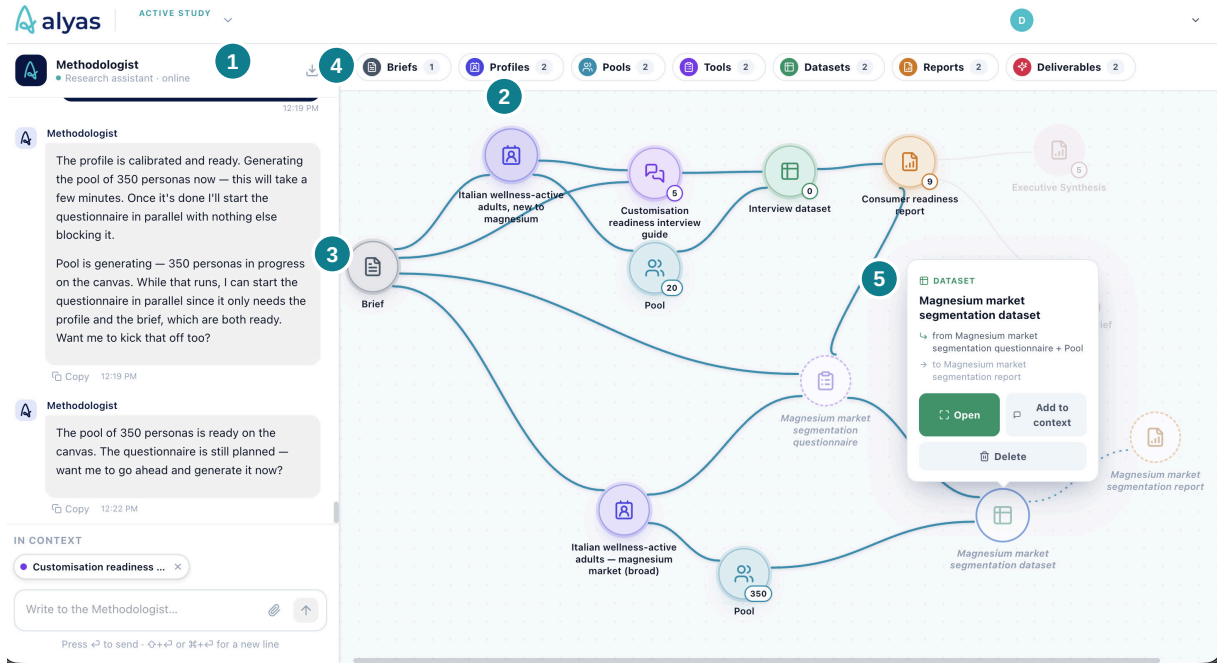


Figure 2: The ALYAS workspace: (1) the methodologist chat; (2) the study as a directed acyclic graph of research artifacts joined by lineage edges, with parallel branches and planned (dashed) nodes; (3) the brief node the graph derives from; (4) artifact-kind filters; (5) node context menu. The study shown branches a market brief into a qualitative and a quantitative line, each ending in its own report.

vented. The structured attributes remain attached to the persona for analysis.

Questionnaire execution. Each respondent answers the whole questionnaire in a single language-model call, independent across personas; type-specific executors and schema-validated parsers handle each item. The respondent prompt addresses two documented failure modes of LLM survey answers (convergence to the scale midpoint and a generic survey register leaking into open text) (Bisbee et al., 2024; Zhou et al., 2025a) by instructing personas to anchor on their strongest relevant trait and answer plausibly rather than ideally. Analyses are computed per item type (frequencies, central tendency and dispersion, average ranks, coded themes), and every run records a snapshot (seed, model, prompt versions) sufficient to reproduce it.

In-depth interviews. For qualitative designs, an interviewer agent conducts multi-turn interviews from the discussion guide, adapting follow-ups turn by turn; output is transcripts, coded themes, and saturation reporting (whether further interviews still surface new themes). In the explanatory-sequential design, a second interview phase runs on a subset of the same pool.

Exploration and reporting. After simulation, the user can explore each respondent’s results or group them by population attributes, open a chat to question individual personas, or convene multi-persona discussion panels. Results are aggregated into a design-aware report whose limitations section is populated from the design’s limits and the population run’s coverage map; a report over converged lines takes the strictest of their limits, and its claims carry inline citations resolving to the underlying evidence. From a report the user can generate deliverables: renderings of the findings for a given audience, carrying its citations.

5 Evaluation

Two experiments evaluate the population and respondent-simulation components: Experiment A the fidelity of the generated populations, Experiment B that of the simulated respondents.

Experiment A: population fidelity. We evaluate the population engine’s ability to reconstruct attributes it was not given, on two corpora with published microdata: the Italian *Aspetti della vita quotidiana* survey (AVQ 2023) and the US OpinionQA collection of Pew American Trends Panel waves (Santurkar et al., 2023). For each, the engine ingests a demographic core from the ware-

System	User assistance		Demographic fidelity				Elicitation		
	<i>Guided design</i>	<i>No-code pop.</i>	<i>End-to-end</i>	<i>Stat. grounding</i>	<i>Provenance</i>	<i>Fidelity check</i>	<i>Surveys & qualitative</i>	<i>Human validation</i>	<i>Claim governance</i>
ALYAS	✓	✓	✓	✓	✓	✓	✓	~	✓
S-Researcher (Wang et al., 2026)	✓	~	✓	—	—	—	~	~	~
UXAgent (Lu et al., 2025)	~	✓	~	—	—	—	✓	—	—
SOTOPIA-S4 (Zhou et al., 2025b)	~	~	~	—	—	—	~	—	—
SocioVerse (Zhang et al., 2025)	—	—	—	~	—	~	✓	✓	—
AgentSociety (Piao et al., 2025)	—	—	—	~	—	—	✓	~	—
AgentTorch (Chopra et al., 2025)	—	—	—	✓	—	~	—	~	—
Dynamo-K (Kang, 2026)	—	—	—	✓	—	—	~	✓	—
HAG (Chen et al., 2026)	—	~	—	~	—	~	~	~	—
SPIRIT (Li and Conrad, 2026)	—	—	—	~	~	~	✓	✓	—
Synthia (Rahimzadeh et al., 2026)	—	—	—	~	—	~	✓	✓	—
Pop.-aligned personas (Hu et al., 2025)	—	—	—	~	—	~	~	~	—
Persona Generators (Paglieri et al., 2026)	—	~	—	—	—	~	~	~	—
SubPOP FT (Suh et al., 2025)	—	—	—	~	—	~	~	✓	—
SSR elicitation (Maier et al., 2025)	—	—	—	~	—	—	✓	✓	—

Table 1: The authors’ reading of persona-simulation systems, 2025–June 2026 (papers only): ✓ present, ~ partial, — absent or no support found. *User assistance*: design guidance from a free-text brief; no-code population specification; end-to-end workflow. *Demographic fidelity*: composition and dependence between attributes from real reference statistics; per-variable provenance exposed to the user; quantitative check of the realized population against the reference. *Elicitation*: typed survey instruments and/or a qualitative mode; validation against human responses (ALYAS: partial, against microdata distributions, §5); enforced limits on reported claims.

house and must reconstruct, through the sourcing fallback of §3, a held-out block (five household and status attributes for ISTAT, Figure 3; four researched attitudes for OpinionQA), scored against microdata not used in sourcing. We compare the engine with and without its web-search fallback (*web-grounded* vs. *native*), against a marginals-only *independence* baseline (recovered marginals crossed as if independent) and a uniform null, over three agent runs. Comparisons use total-variation distance, $\text{TVD}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|$: the share of probability mass placed differently, zero for exact agreement (lower is better); Cramér’s V measures the strength of pairwise association.

Reconstructed marginals land close to truth: all five ISTAT block attributes at $\text{TVD} \leq 0.07$ and the four OpinionQA attitudes at mean $\text{TVD} \approx 0.07$, against uniform-null ceilings of 0.20–0.44. The drawn population also reproduces the strength of pairwise dependence, not just its presence: the ingested core matches truth almost exactly, and the reconstructed block associations transfer with 47 of 66 pairs within 0.05 of the true Cramér’s V (Figure 3). Modeling the joint beats independence on both corpora: on ISTAT, the mean joint TVD over block-involving pairs is 0.038 with web

grounding and 0.065 native, against 0.074 for the independence baseline; on OpinionQA it is 0.092 (web) and 0.097 (native) against 0.114 (independence); all far below the uniform null shown in Figure 3 (bottom right). Live web retrieval helps unevenly; the dependable gains come from warehouse grounding and joint modeling, not from search.

Experiment B: subgroup response fidelity. Experiment B asks whether profile-conditioned respondents answer like the specific people the profile describes, rather than like the general population. From the AVQ 2023 microdata (41,750 respondents) we draw five *twin groups*: twenty real respondents each, identical on the eight grounding attributes and within five years of age, chosen among the 94 qualifying profiles to span mainstream and outlying cases; each group’s observed answers are the ground truth for its simulated twins. Each group answers sixteen AVQ items selected on the microdata for demographic signal (weighted Cramér’s V against sex, age, and education) under four conditions using the profile progressively more: none (*generic*), a single call describing the shared profile (*group-level*), twenty

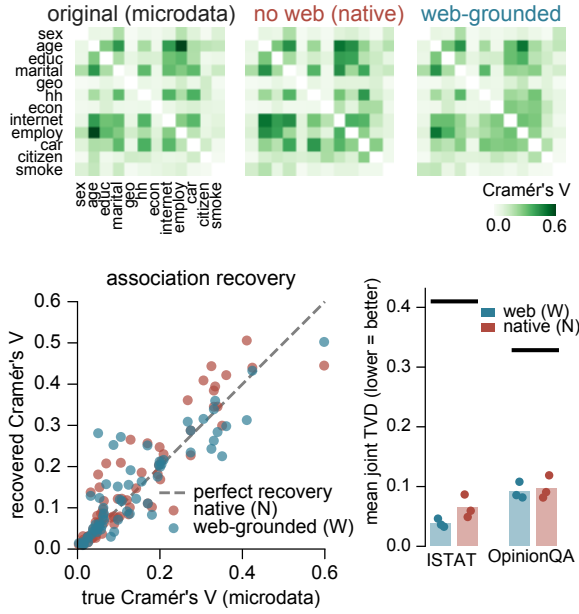


Figure 3: Association fidelity on ISTAT. *Top*: pairwise Cramér’s V among the seven core and five reconstructed attributes: microdata (left), native (centre), and web-grounded (right) reconstructions on a shared scale. *Bottom left*: recovered versus true V per pair (mean over three runs); the identity line is exact recovery. *Bottom right*: mean pairwise joint TVD (bars) with the individual runs (dots), over block-involving pairs (ISTAT) and demographic/attitude pairs (OpinionQA, family-income pairs excluded); black bars mark the flat null, the TVD of a uniform joint.

profile personas, and twenty personas with added *backstories* (§4); three replicates per condition (Sonnet 5). Each estimate is scored by its TVD to the group’s observed distribution and benchmarked against its multinomial $n=20$ resampling envelope: an estimate inside it is as close as a second real group of twenty would be. Analysis covers the 58 of 80 (group, item) cells where the group deviates from the national marginal beyond that envelope.

The generic prompt (mean TVD 0.370) matches the national-marginal baseline (0.365): unconditioned, the model reproduces the population pattern, not the group’s (Figure 4, top). Profile conditioning cuts the error roughly threefold, backstories add a small further gain (Figure 4, bottom), and 45–52% of conditioned estimates fall within the envelope, against 2% for the generic prompt (over all 80 cells: TVD 0.307 generic vs. 0.133–0.136 conditioned, baseline 0.299).

6 Related Work and Comparison

LLMs as survey respondents. Argyle et al.

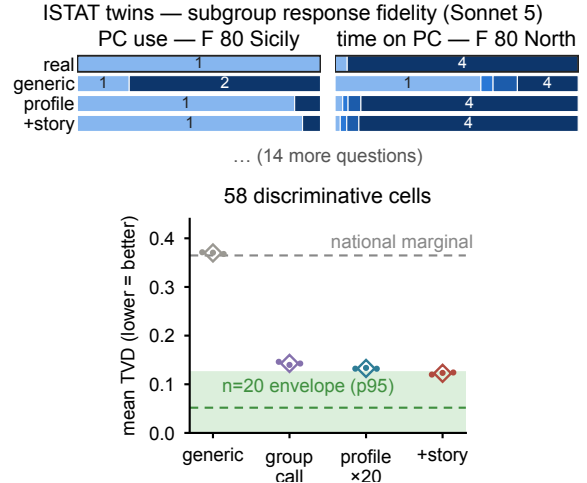


Figure 4: Subgroup response fidelity on ISTAT twins (Sonnet 5). *Top*: two exemplar (group, item) cells: the real twin distribution (ink edge) against the generic, profile-persona, and profile+backstory estimates. *Bottom*: mean TVD per condition over the 58 discriminative cells (small dots: replicates), against the national-marginal baseline (grey dashed) and the $n=20$ resampling envelope (band to p95, dashed median); estimates inside the band match a second real twin sample.

(2023) introduced “silicon sampling”: conditioning a model on backstories resampled from real ANES respondents, recovering marginal and correlational structure of the human data; ALYAS does not require respondent-level survey microdata covering the target population. Park et al. (2024) ground generative agents in a two-hour interview per participant, predicting individual responses at 0.83 normalized accuracy relative to test–retest reliability (per-person grounding ALYAS trades for statistical grounding); Maier et al. (2025) find behavioral validity achievable but variable.

Grounded populations. SocioVerse (Zhang et al., 2025) aligns a $\sim 10\text{M}$ -profile pool to census and poll marginals by raking, improving election simulation; ALYAS instead assembles a population on demand for any frame, with per-quantity provenance. Demographic-tree generation (Chen et al., 2026) is closest on grounding (a topic-conditioned demographic joint with real-WVS leaves, beating the product of marginals), though its conditionals are LLM-estimated, where ALYAS reads observed contingencies. Table 1 and the broader literature cover population steering via psychometrics, trait diversity, census calibration, or independent marginals (Hu et al., 2025; Paglieri

et al., 2026; NVIDIA, 2025; Chopra et al., 2025; Sun et al., 2024) and survey-ready populations from microdata (Piao et al., 2025; Kang, 2026; Li and Conrad, 2026; Rahimzadeh et al., 2026; Suh et al., 2025).

Limitations

ALYAS simulates respondents with a language model and inherits its limitations; simulated results complement, not replace, human studies (Maier et al., 2025). Population fidelity is bounded by the evidence the sourcing agent can find: reconstructed attributes carry residual error, and absent joint evidence the reconstruction falls back to maximum entropy. Our response evidence so far covers only subgroup answers on public ISTAT microdata (§5). The methodologist assistant, design composition, and claim governance are demonstrated in the system but not evaluated here.

Ethics Statement

ALYAS produces synthetic survey respondents, not human data; the two must not be conflated, and its output carries the evidential status of a simulation, not of fieldwork. The salient harms are misplaced confidence and the manufacture of an apparent public consensus, which we counter by holding reported claims to the limits of the design and by exposing each attribute’s provenance. Because a language model can encode and amplify social bias, simulated subgroups, especially small or marginalized ones, may misrepresent the people they stand for (Sun et al., 2024), so results warrant validation against human data before they inform decisions. The reference warehouse holds institutional aggregates and public-use microdata, so no personally identifying information is involved.

Acknowledgments

We thank Prof. Ivano Bison for generously sharing questionnaire data that informed some analyses of the system. The authors used AI assistants both in building the system (coding agents contributed to the implementation of ALYAS and of its evaluation pipeline) and in preparing this paper, for brainstorming and editing text. All AI-assisted content was reviewed, edited, and verified by the authors, who take full responsibility for the methods, results, and text.

References

- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.
- Rongxin Chen, Tianyu Wu, Bingbing Xu, Jiatang Luo, Xiucheng Xu, and Huawei Shen. 2026. HAG: Hierarchical demographic tree-based agent generation for topic-adaptive simulation. *Preprint*, arXiv:2601.05656.
- Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. 2025. On the limits of agency in agent-based models. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.
- Imre Csiszár. 1975. *I-divergence geometry of probability distributions and minimization problems*. *The Annals of Probability*, 3(1):146–158.
- W. Edwards Deming and Frederick F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.
- Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025. Population-aligned persona generation for LLM-based social simulation. *Preprint*, arXiv:2509.10127.
- Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2026. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *Preprint*, arXiv:2602.15785.
- Sungwoo Kang. 2026. Diagnosing Korean-language LLM political bias via census-grounded agent simulation. *Preprint*, arXiv:2605.18395.
- Mao Li and Frederick G. Conrad. 2026. Persona-based simulation of human opinion at population scale. *Preprint*, arXiv:2603.27056.
- Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Zheshen Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025. UXAgent: A system for simulating usability testing of web design with LLM agents. *Preprint*, arXiv:2504.09407.
- Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. 2025.

- LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. *Preprint*, arXiv:2510.08338.
- NVIDIA. 2025. *Nemotron-Personas-USA: Synthetic personas grounded in U.S. demographics*. Hugging Face dataset.
- Davide Paglieri, Logan Cross, William A. Cunningham, Joel Z. Leibo, and Alexander Sasha Vezhn-evets. 2026. *Persona generators: Generating diverse synthetic personas for arbitrary contexts*. *Preprint*, arXiv:2602.03545.
- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. *Generative agent simulations of 1,000 people*. *Preprint*, arXiv:2411.10109.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. *AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society*. *Preprint*, arXiv:2502.08691.
- Vahid Rahimzadeh, Erfan Moosavi Monazzah, Mohammad Taher Pilehvar, and Yadollah Yaghoobzadeh. 2026. *Synthia: Scalable grounded persona generation from social media data*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. *Whose opinions do language models reflect?* In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. *Language model fine-tuning on scaled survey data for predicting distributions of public opinions*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. *Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information*. *Preprint*, arXiv:2402.18144.
- Lei Wang, Yuanzi Li, Jinchao Wu, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2026. *LLM agents as social scientists: A human-AI collaborative platform for social science automation*. *Preprint*, arXiv:2604.01520.
- Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, and 2 others. 2025. *SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users*. *Preprint*, arXiv:2504.10157.
- Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025a. *The PIMMUR principles: Ensuring validity in collective behavior of LLM societies*. *Preprint*, arXiv:2509.18052.
- Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, Hao Zhu, and Maarten Sap. 2025b. *SOTOPIA-S4: A user-friendly system for flexible, customizable, and large-scale social simulation*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 350–360, Albuquerque, New Mexico. Association for Computational Linguistics.