Mensch und Computer 2026 · Workshop · Duisburg · 30.08.2026

The Median Has No Edge Cases

An Evidence Map for Synthetic Users in UX Research

Prof. Dr. Andreas Hinderks · Hochschule Hannover

jan-andreas.hinderks@hs-hannover.de

Synthetic users promise user insights without recruiting a single user.

(Rosala & Moran, 2024)

1 · The question

A question with no single answer

Do synthetic users work?

The literature looks contradictory

- **Close alignment reported.** Simulated samples reproduced aggregate voting patterns and the results of classic behavioral experiments.
- **Systematic failure reported.** Re-analyses found collapsed variance, diverging correlations and results that shift with the model version.

Alignment: (Argyle et al., 2023; Aher et al., 2023)

Failure: (Bisbee et al., 2024; Gao et al., 2025)

User research does three different jobs

User research does three different jobs

- **Estimate averages** — a mean score, a majority preference, the direction of an effect.

User research does three different jobs

- **Estimate averages** — a mean score, a majority preference, the direction of an effect.
- **Map variation** — subgroup differences, divergent needs, the structure of the responses.

User research does three different jobs

- **Estimate averages** — a mean score, a majority preference, the direction of an effect.
- **Map variation** — subgroup differences, divergent needs, the structure of the responses.
- **Reach lived experience** — what a practice means to a person in their context.

A method can serve one function and fail at another. “Does it work?” bundles them into a single verdict.

2 · Evidence

What comparative studies show

Read by function, is the evidence still contradictory?

Where the evidence comes from

- **182 studies** of LLM-generated participants, covered by a systematic review.
- **Targeted additions** from social science, NLP and HCI.
- **No new systematic search.** The contribution is the reading by research function, not new coverage.

(Kuric et al., 2026)

Preprint, not peer-reviewed; the authors are affiliated with a user-research vendor.

Averages replicate — under conditions.

The optimist case is real

- **Argyle et al. (2023)** — GPT-3 with rich survey-based personas reproduced aggregate US voting patterns. They called it algorithmic fidelity.
- **Aher et al. (2023)** — classic behavioral experiments, including the Ultimatum Game, replicated at the aggregate level.
- **Horton et al. (2023)** — behavioral economics reruns, explicitly framed as a cheap way to pilot designs.
- **Hewitt et al. (2024)** — direction and rough size of treatment effects predicted across a large archive of survey experiments.

Hewitt et al.: preprint, not peer-reviewed

Horton et al.: NBER working paper

Three conditions recur

- **Aggregate outcomes** — means, majority directions, effect signs.
- **Tasks well covered in the training data** — US politics, classic experiments, common moral vignettes.
- **Rich human-derived context in the prompt.**

Where these conditions hold, simulated averages can be informative.
Where they do not, nothing follows from these results.

Same mean

(Bisbee et al., 2024)

different distribution

Bisbee and colleagues re-ran the silicon-sampling design. Simulated means matched human survey means, while variance collapsed, inter-variable correlations diverged, and results shifted across model versions. Optimists and pessimists can read the same simulation and both be right, depending on which statistic they report.

It is not only the variance

- **Gao et al. (2025)** — near-uniform response distributions where human responses are broad; failure modes shift unpredictably with language, role and safety tuning.
- **Domínguez-Olmedo et al. (2024)** — ordering and labeling artifacts larger than the signal of interest. A direct warning for anyone giving a standardized questionnaire to a model.
- **Almeida et al. (2024)** — direction of moral and legal judgments largely tracked, effect sizes exaggerated, and on one manipulation every tested model reversed the human result.

Two harms when a model speaks for a group

Misportrayal

– The model reflects what outsiders say **about** a group.

– Not the group’s own voice.

Flattening

– Diversity **within** the group disappears.

– One synthetic voice stands in for many real ones.

Both patterns: Wang, Morgenstern & Dickerson (2025). Text about a group is far more common in training data than text by a group.

Culture is where it breaks in this room

- **German vote choice** — a synthetic sample matched to a German election study was biased toward Green and Left parties. The model captured typical voters and missed the rest.
- **Open-ended German opinions** — uneven subgroup fit, worst for the groups furthest from the center.
- **Closest alignment with WEIRD populations** — Western, educated, industrialized, rich, democratic.

(von der Heyde et al., 2026; Ma et al., 2025; Atari et al., 2023)

Atari et al.: preprint, not peer-reviewed

The most direct HCI evidence is qualitative

- **Nineteen qualitative researchers** recreated one of their own interview studies with an LLM.
- **First surprise, then limits** — plausible narratives at first; over further turns, no consent, no palpability, no contextual depth.
- **A normative echo** — substitution conflicts with the goals of participation itself: representation, inclusion, understanding.

(Kapania et al., 2025; Agnew et al., 2024)

The negative evidence has limits too

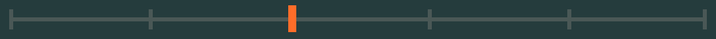
- **Model turnover** — several documented failures were measured on models that are no longer state of the art.
- **Contamination** inflates positive results wherever classic experiments or public survey items are replicated.
- **US, English, and widely varying prompting setups** — comparability across studies is limited.
- **The largest review is a preprint** written by researchers affiliated with a user-research vendor.

This is why the argument is anchored in mechanism, not in any single benchmark result.

3 · Mechanism

Why the split is built into the models

Immaturity, or architecture?



What the training pipeline predicts

(Ouyang et al., 2022)

What the training pipeline predicts

- **Next-token prediction** over a large corpus learns a weighted average of the training distribution. The most probable outputs sit near its center.

(Ouyang et al., 2022)

What the training pipeline predicts

- **Next-token prediction** over a large corpus learns a weighted average of the training distribution. The most probable outputs sit near its center.
- **Instruction tuning with human feedback** narrows that distribution further, because responses are optimized toward what raters prefer.

(Ouyang et al., 2022)

The median has a face: middle-aged, able-bodied, native-born, Caucasian, atheistic, male.

(Tan & Lee, 2025)

One objection, met head-on

(Horton et al., 2023)

One objection, met head-on

- **The objection.** At nonzero temperature a model returns a spread of answers, not a single point. Treat it as a configurable generator that can be assigned characteristics and sampled repeatedly.

(Horton et al., 2023)

One objection, met head-on

- **The objection.** At nonzero temperature a model returns a spread of answers, not a single point. Treat it as a configurable generator that can be assigned characteristics and sampled repeatedly.
- **The answer.** What matters is not whether the model varies, but whether its variation matches human variation. A model can sample widely and still concentrate most of its mass in the wrong place.

(Horton et al., 2023)

Persona prompting moves the median. It does not restore the distribution.

(Cheng et al., 2023; Wang et al., 2025)

Out of scope — digital twins

- **Fine-tuning on rich individual-level data** injects exactly the human variance that generic prompting lacks.
- **Early practitioner evidence** suggests grounded models outperform persona-prompted ones.
- **But the economics change.** Each simulated person needs substantial human data — which removes the cost argument that motivates synthetic users in the first place.

(Budiu, 2025)

Everything that follows applies to prompt-based simulation of generic or persona-described users.

Does this transfer to UX research?

The mechanism transfers

- Variance collapse, default-persona drift and response artifacts are properties of the model, not of the survey domain.
- They operate before any domain content enters the prompt.

The tasks are untested

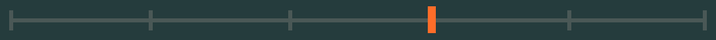
- Do synthetic walkthroughs find the usability problems that human tests find?
- Do simulated questionnaire responses preserve psychometric structure, not only scale means?
- How does synthetic think-aloud output compare to human verbalization?

The burden of proof lies with the claim that UX is an exception. (Lewis & Sauro, 2026)

4 · The map

An evidence map for synthetic users

Where exactly is the line?



Two questions place any research activity

Two questions place any research activity

- **What kind of knowledge does the activity need?** Averages, variation and subgroups, or lived experience.

Two questions place any research activity

- **What kind of knowledge does the activity need?** Averages, variation and subgroups, or lived experience.
- **How much depends on its result?** Procedural, directional, or substantive.

How much depends on the result

How much depends on the result

- **Procedural** — does the study run at all? Instruments, survey logic, moderator training.

How much depends on the result

- **Procedural** — does the study run at all? Instruments, survey logic, moderator training.
- **Directional** — what to study next: prioritization, sampling, which hypothesis to test.

How much depends on the result

- **Procedural** — does the study run at all? Instruments, survey logic, moderator training.
- **Directional** — what to study next: prioritization, sampling, which hypothesis to test.
- **Substantive** — claims about users and products; decisions that change the design.

Nine activities, placed by two questions

How much depends on the result →

Kind of knowledge the activity needs →

	Averages aggregate scores, majority directions	**Variation & subgroups** differences, response structure	**Lived experience** meaning, context, representation
Substantive claims about users, products; design decisions	Summative evaluation from synthetic panels (UEQ, SUS); market estimates	Personas and segmentations presented as empirical findings about real users	Insight generation; interviews as data; research on marginalized populations
Directional what to study next; prioritization, sampling	Hypothesis triage; effect-direction screening; proto-personas as hypotheses	Segment prioritization or sampling plans derived from simulated heterogeneity	Scoping research directions for underrepresented groups from simulated voices
Procedural does the study run? instruments, training	Pilot instruments and survey logic; dry-run guides and analysis pipelines	Heterogeneous test inputs for branching and edge-case checks	Moderator training; role-played difficult interviews

Fig. 1: The placement, before the verdict. Hinderks (2026), redrawn.

One cell holds. Three hold with constraints. Five do not.

How much depends on the result →

Kind of knowledge the activity needs →

	Averages aggregate scores, majority directions	**Variation & subgroups** differences, response structure	**Lived experience** meaning, context, representation
Substantive claims about users, products; design decisions	Summative evaluation from synthetic panels (UEQ, SUS); market estimates	Personas and segmentations presented as empirical findings about real users	Insight generation; interviews as data; research on marginalized populations
Directional what to study next; prioritization, sampling	Hypothesis triage; effect-direction screening; proto-personas as hypotheses	Segment prioritization or sampling plans derived from simulated heterogeneity	Scoping research directions for underrepresented groups from simulated voices
Procedural does the study run? instruments, training	Pilot instruments and survey logic; dry-run guides and analysis pipelines	Heterogeneous test inputs for branching and edge-case checks	Moderator training; role-played difficult interviews

Fig. 1: Verdicts as supported by the evidence. Hinderks (2026), redrawn.

One cell holds. Three hold with constraints. Five do not.

How much depends on the result →

Kind of knowledge the activity needs →

	Averages aggregate scores, majority directions	**Variation & subgroups** differences, response structure	**Lived experience** meaning, context, representation
Substantive claims about users, products; design decisions	Summative evaluation from synthetic panels (UEQ, SUS); market estimates	Personas and segmentations presented as empirical findings about real users	Insight generation; interviews as data; research on marginalized populations
Directional what to study next; prioritization, sampling	Hypothesis triage; effect-direction screening; proto-personas as hypotheses	Segment prioritization or sampling plans derived from simulated heterogeneity	Scoping research directions for underrepresented groups from simulated voices
Procedural does the study run? instruments, training	**Defensible** Pilot instruments and survey logic; dry-run guides and analysis pipelines	Heterogeneous test inputs for branching and edge-case checks	Moderator training; role-played difficult interviews

Fig. 1: Verdicts as supported by the evidence. Hinderks (2026), redrawn.

One cell holds. Three hold with constraints. Five do not.

How much depends on the result →

Kind of knowledge the activity needs →

	Averages aggregate scores, majority directions	**Variation & subgroups** differences, response structure	**Lived experience** meaning, context, representation
Substantive claims about users, products; design decisions	Summative evaluation from synthetic panels (UEQ, SUS); market estimates	Personas and segmentations presented as empirical findings about real users	Insight generation; interviews as data; research on marginalized populations
Directional what to study next; prioritization, sampling	**Conditional** Hypothesis triage; effect-direction screening; proto-personas as hypotheses Constraint: human validation required	Segment prioritization or sampling plans derived from simulated heterogeneity	Scoping research directions for underrepresented groups from simulated voices
Procedural does the study run? instruments, training	**Defensible** Pilot instruments and survey logic; dry-run guides and analysis pipelines	**Conditional** Heterogeneous test inputs for branching and edge-case checks Constraint: tests mechanics, not real variance	**Conditional** Moderator training; role-played difficult interviews Constraint: builds skill, produces no data

Fig. 1: Verdicts as supported by the evidence. Hinderks (2026), redrawn.

One cell holds. Three hold with constraints. Five do not.

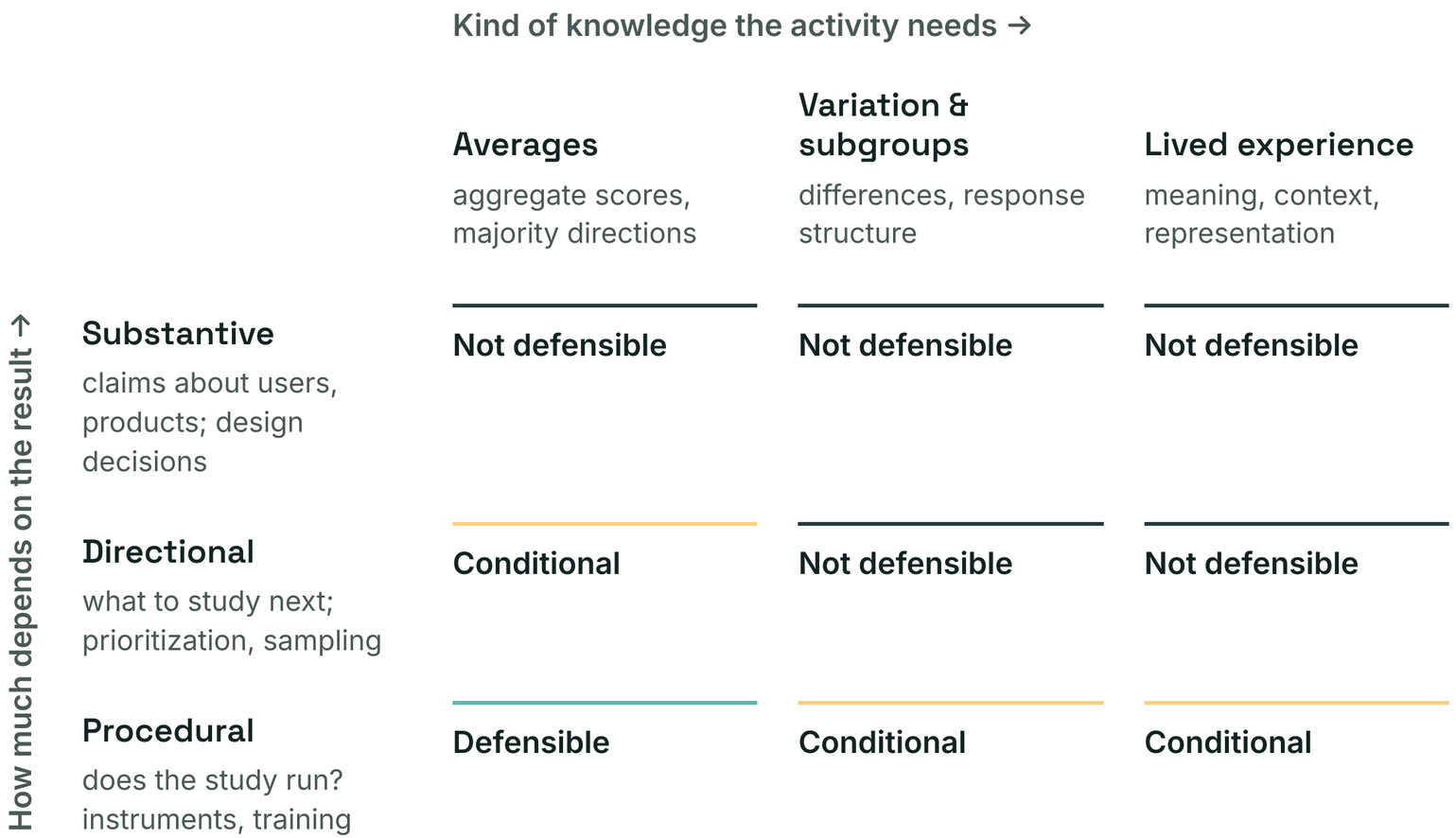
How much depends on the result →

Kind of knowledge the activity needs →

	Averages aggregate scores, majority directions	Variation & subgroups differences, response structure	Lived experience meaning, context, representation
Substantive claims about users, products; design decisions	**Not defensible** Summative evaluation from synthetic panels (UEQ, SUS); market estimates	**Not defensible** Personas and segmentations presented as empirical findings about real users	**Not defensible** Insight generation; interviews as data; research on marginalized populations
Directional what to study next; prioritization, sampling	**Conditional** Hypothesis triage; effect-direction screening; proto-personas as hypotheses Constraint: human validation required	**Not defensible** Segment prioritization or sampling plans derived from simulated heterogeneity	**Not defensible** Scoping research directions for underrepresented groups from simulated voices
Procedural does the study run? instruments, training	**Defensible** Pilot instruments and survey logic; dry-run guides and analysis pipelines	**Conditional** Heterogeneous test inputs for branching and edge-case checks Constraint: tests mechanics, not real variance	**Conditional** Moderator training; role-played difficult interviews Constraint: builds skill, produces no data

Fig. 1: Verdicts as supported by the evidence. Hinderks (2026), redrawn.

How to read the map



Place an activity by the **strongest** kind of knowledge it needs and by the **highest** weight resting on its result.

In case of doubt, classify toward the upper right.

Fine-tuned, individual-level digital twins are out of scope.

Fig. 1: Reading rule for the evidence map. Hinderks (2026), redrawn.

One defensible cell, three conditional ones

Defensible: procedural work on averages

- Pilot instruments and survey logic.
- Dry-run interview guides.
- Rehearse analysis pipelines.
- The only question is whether the study runs. A plausible average participant is enough for that.

Conditional: three cells, three constraints

- **Hypothesis triage** — every simulated result validated with humans before resources are committed.
- **Heterogeneous test inputs** — tests survey mechanics, does not estimate real variance.
- **Moderator training** — builds researcher skill, produces no data.

(Hämäläinen et al., 2023; Horton et al., 2023; Hewitt et al., 2024; Kuric et al., 2026)

What is not defensible

- **Summative evaluation from synthetic panels** — response artifacts dominate the signal and failure is unpredictable.
- **Personas and segmentations presented as findings** — variance collapse and stereotyping are inherited wholesale.
- **Sampling plans derived from simulated heterogeneity** — built on variance the model does not have.
- **Insight generation, interviews as data, research on marginalized populations** — no consent, no palpability, misportrayal and flattening.

(Bisbee et al., 2024; Domínguez-Olmedo et al., 2024; Gao et al., 2025; Kapania et al., 2025; Wang et al., 2025)

Two rules for every use that survives

- **Disclose synthetic provenance** wherever results are reported. Presenting synthetic findings as human findings damages the credibility of user research itself.
- **Pin and document the configuration** — model, version, prompts, parameters. Results do not survive version changes.

(Rosala & Moran, 2024; Kapania et al., 2025; Batzner et al., 2025; Gao et al., 2025)

5 · What follows

A research agenda for HCI

What would settle this?

Three studies this community can run

- **Controlled comparison on established UX questionnaires** — evaluated on item variances, factor structure and reliability, not only on scale means.
- **Usability-problem discovery** — synthetic walkthroughs against human tests on the same product: hit rates, severity calibration, false-positive profiles.
- **A German-language replication** — extending the election-study line to UX constructs, in this community's own context.

Plus reporting standards: model and version, the full prompting setup, the research function the synthetic data served, and the human validation performed.

(Batzner et al., 2025)

Synthetic users can rehearse research. They cannot perform it.

The median has no edge cases — and edge cases are where load-bearing user research earns its keep.

References

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3613904.3642703

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202, pp. 337–371).

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence, 333*, 104145. https://doi.org/10.1016/j.artint.2024.104145

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* PsyArXiv. Preprint, not peer-reviewed.

Batzner, J., Stocker, V., Tang, B., Natarajan, A., Chen, Q., Schmid, S., & Kasneci, G. (2025). Whose personae? Synthetic persona experiments in LLM research and pathways to transparency. In *Proceedings of the 2025 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1609/aies.v8i1.36553

References

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401–416.

Budiu, R. (2025). *Evaluating AI-simulated behavior: Insights from three studies on digital twins and synthetic users*. Nielsen Norman Group. https://www.nngroup.com/articles/ai-simulations-studies/

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Domínguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*.

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences, 122*(24), e2501660122. https://doi.org/10.1073/pnas.2501660122

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3544548.3580688

References

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). *Predicting results of social science experiments using large language models*. Preprint, not peer-reviewed.

Horton, J. J., Filippas, A., & Manning, B. S. (2023). *Large language models as simulated economic agents: What can we learn from Homo silicus?* (NBER Working Paper No. 31122). National Bureau of Economic Research.

Kapania, S., Agnew, W., Eslami, M., Heidari, H., & Fox, S. E. (2025). "Simulacrum of stories": Examining large language models as qualitative research participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3706598.3713220

Kuric, E., Demcak, P., & Krajcovic, M. (2026). *Synthetic participants generated by large language models: A systematic literature review*. Research Square. Preprint, not peer-reviewed. https://doi.org/10.21203/rs.3.rs-9057643/v1

Lewis, J., & Sauro, J. (2026). *A review of experiments with synthetic users*. MeasuringU. https://measuringu.com/review-of-experiments-with-synthetic-users/

Ma, B., Yoztyurk, B., Haensch, A.-C., Wang, X., Herklotz, M., Kreuter, F., Plank, B., & Aßenmacher, M. (2025). Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1785–1809). ACL. https://doi.org/10.18653/v1/2025.acl-long.90

References

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Rosala, M., & Moran, K. (2024). *Synthetic users: If, when, and how to use AI-generated "research"*. Nielsen Norman Group. https://www.nngroup.com/articles/synthetic-users/

Tan, B. C. Z., & Lee, R. K.-W. (2025). Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the ACL: Human Language Technologies* (Vol. 1, pp. 1075–1108). ACL.

von der Heyde, L., Haensch, A.-C., & Wenz, A. (2026). Vox populi, vox AI? Using large language models to estimate German vote choice. *Social Science Computer Review, 44*(3). https://doi.org/10.1177/08944393251337014

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-025-00986-z

Thank you.

Prof. Dr. Andreas Hinderks

Hochschule Hannover, Fakultät IV

jan-andreas.hinderks@hs-hannover.de · hinderks.org