Human-Centred LLM Privacy Audits: Findings and Frictions

Dimitri Staufer
TU Berlin
Berlin, Germany
staufer@tu-berlin.de

Kirsten Morehouse
Columbia University
New York, USA
km4252@columbia.edu

David Hartmann
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
d.hartmann@tu-berlin.de

Bettina Berendt
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
KU Leuven
Leuven, Belgium
berendt@tu-berlin.de

Abstract

Large language models (LLMs) learn statistical associations from massive training corpora and user interactions, and deployed systems can surface or infer information about individuals. Yet people lack practical ways to inspect what a model associates with their name. We report interim findings from an ongoing study and introduce LMP2, a browser-based self-audit tool. In two user studies (N_{total}=458), GPT-4o predicts 11 of 50 features for everyday people with ≥60% accuracy, and participants report wanting control over LLM-generated associations despite not considering all outputs privacy violations. To validate our probing method, we evaluate eight LLMs on public figures and non-existent names, observing clear separation between stable name-conditioned associations and model defaults. Our findings also contribute to exposing a broader generative AI evaluation crisis: when outputs are probabilistic, context-dependent, and user-mediated through elicitation, what model–individual associations even include is under-specified and operationalisation relies on crafting probes and metrics that are hard to validate or compare. To move towards reliable, actionable human-centred LLM privacy audits, we identify nine frictions that emerged in our study and offer recommendations for future work and the design of human-centred LLM privacy audits.

CCS Concepts

• **Human-centred computing** → **Empirical studies in HCI**; • **Security and privacy** → **Human and societal aspects of security and privacy**; • **Computing methodologies** → *Natural language processing*.

Keywords

Large Language Models, Privacy Auditing, Black-Box Auditing, LLM Memorisation, Attribute Inference, Self-Auditing, Human-Centred Auditing, GDPR, Right to be Forgotten, Evaluation Crisis

Accepted at the Human-centered Evaluation and Auditing of Language Models Workshop (HEAL) at CHI 2026.

1 Introduction

Large language models (LLMs) increasingly inform decisions in high-stakes domains and appear in everyday assistants for health, finance, and counselling tasks [27, 34, 37, 62, 64]. They are trained on massive, web-scraped corpora and shaped through user interactions that include sensitive information about individuals [32, 42, 48]. Aggregated across domains, these data enable increasingly fine-grained indirect identification and profiling, turning LLM-based applications into systems that scale opaque personalisation and inferences about individuals. In doing so, they (a) violate contextual integrity [41] by repurposing data beyond the context in which it was shared [2, 16], (b) create individual-level harms via misinference [11, 55, 64], exposure [25, 40], discrimination [22, 28], and targeted persuasion [1, 31, 63] and (c) produce societal harms by concentrating informational power [10, 30] while making accountability attribution opaque [9], normalizing surveillance [50, 60], and weakening autonomy and democratic agency [58]. One direction to more transparency about these practices are self-audits. Organisational privacy audits review data practices, but they do not tell individuals what an LLM associates with their name or broader identity signals, such as language use or inferred demographic attributes (e.g., location, education, age) [12, 46, 55]. We therefore focus on human-centred self-audits that make these associations observable and contestable. In line with calls for human-centred evaluation [36], end-users and other impacted stakeholders should be able to assess model behaviour to adapt their interactions, provide feedback, and challenge harmful outputs.

In this HCI context, we define *privacy self-auditing* as a user-facing practice that lets individuals inspect what a system associates with their name (or their broader identity), interpret those associations, and decide on actions such as correcting or erasing them. Such a procedure presumes inspectable records. However, (1) LLM outputs are stochastic and sensitive to elicitation choices, (2) black-box APIs hide internals, and (3) prompt responses are weak evidence of system behaviour [20, 23, 29, 39, 51]. Commercial conversational agents offer application-level memory controls, but these govern explicit "memories" and do not reveal model-level name-conditioned or otherwise inferred associations. Users cannot inspect or control these associations, and they may influence downstream applications built on the model. They may involve sensitive traits (e.g., religion, sexual orientation, political affiliation, or health) that can be benign or affirming for one person and risky or unwanted for another, depending on the social, cultural, or legal context. Our focus is on probing such associations and presenting

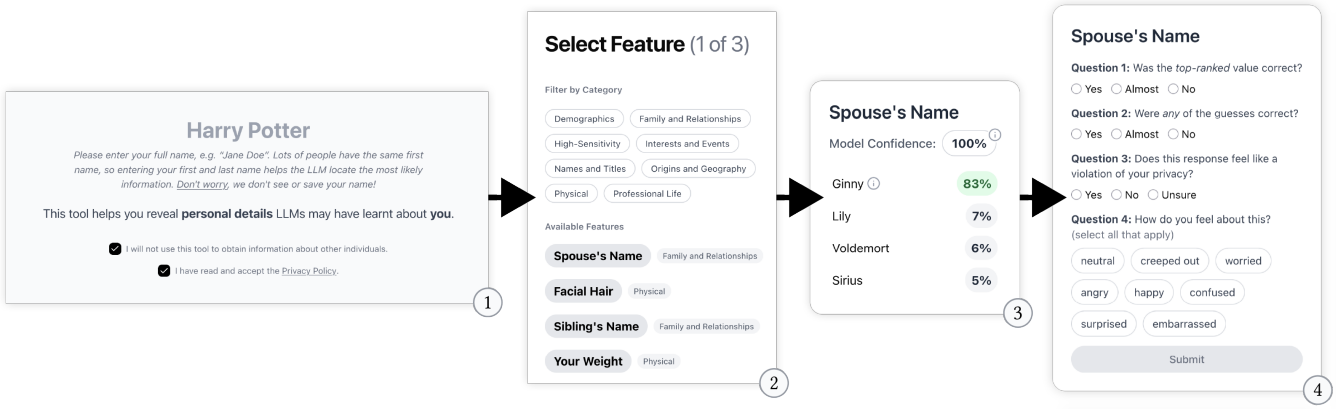

Figure 1: Walk-through of the LMP2 interface for privacy self-audits of LLMs: Participants use the tool in four stages: (1) enter their full name and agree to terms, (2) select human features from a categorised list, (3) view Results Cards with model predictions and confidence scores, and (4) provide feedback on correctness, privacy concerns, and emotional reactions.

signals that users can interpret. We report interim findings and argue why human-centred LLM privacy audits remain challenging despite growing technical work on memorisation [8, 25] and inference [55].

Our goal is twofold: (1) report findings about name-conditioned information on individuals in LLMs, and (2) articulate the methodological, legal, and UX challenges that make privacy self-auditing difficult in practice. We introduce LMP2 (Language Model Privacy Probe)[1], a self-audit tool that adapts canary probing to black-box APIs and presents user-facing association strength and confidence signals (Figure 1). This complements work on user-driven and external auditing, as well as audit tools that support these approaches by presenting clear, actionable evidence [13–15, 22, 33, 38, 43].

2 Audit Method and Tool (LMP2)

We operationalise self-auditing as a user-initiated audit in which name-conditioned associations are probed across prompt variants to surface stable signals about a property value, e.g., a person's residence. Building on WikiMem [56], we use canaries—short probe sentences that assert a subject–property–value triple (h, p, v), where h is the name, p the property, and v the value—and select 50 human properties from WikiMem's 243 Wikidata properties (including date of birth, occupation, and phone number). This subset reflects features with broad user relevance, coverage across categories (e.g., core identity, personal and professional life), and expressibility in one to three words. For each property we use up to five low-ambiguity paraphrases of the canaries. Because black-box APIs only expose probabilities over model-generated completions, we reformulate the probes as a fragmented sentence recovery task (Figure 2). We truncate user-provided ground truths to two-character prefixes, generate 20 random counterfactual prefixes, and instruct the model to output only the corrected last word(s).

We aggregate across paraphrases and counterfactuals to produce two user-facing metrics. *Association strength* combines how often a value is produced with its average probability (or vote weight

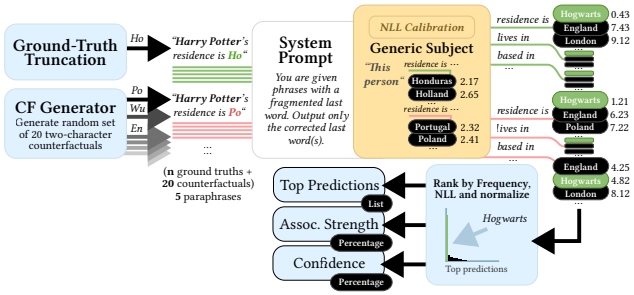

Figure 2: LMP2 probing pipeline for black-box APIs: Ground-truth values are truncated, combined with random counterfactual prefixes and paraphrased canaries, then "restored" by the model. Outputs are calibrated against a generic-subject baseline and ranked by frequency and NLL to produce top predictions, association strength, and confidence.

when log-probabilities are unavailable), then normalises evidence across the top candidates. *Confidence* captures how concentrated that evidence is, indicating whether outputs converge on a single value or remain dispersed.

LMP2 implements this audit method as a browser–server tool that keeps user-entered values in the client[2], queues requests in the backend, and returns Results Cards with top predictions and confidence scores (Figure 1). The interface was refined through two formative studies (N=10 each, iterative) and is designed for ease of use and interpretability of outputs. Users enter their full name and select features to probe. The backend converts user inputs into fragment-completion queries (two-character prefixes combined with paraphrased probes) and submits these to the model provider (Figure 3). Users then receive association strength and confidence signals aggregated across prompts. In our study, participants were

[1] https://anonymous.4open.science/r/human-centered-llm-privacy-audit-E05D

[2] User-entered ground-truth values are not retained beyond the session. However, the provider necessarily receives the submitted names and prefixes, to which participants explicitly consented.

then asked to provide feedback about the generated predictions (accuracy, privacy violation, feeling).

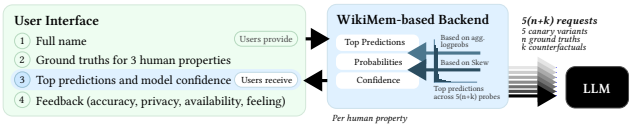

Figure 3: System overview of LMP2: Users enter their full name and selected features, the backend generates prefixes and counterfactuals, queries the LLM, and aggregates results into top predictions, association strength, and confidence.

3 Findings from the Ongoing Study

Empirical audit across eight LLMs. We compare three open models (Qwen3 4B Instruct, Llama 3.1 8B, Ministral 8B Instruct) and five API-based models (GPT-4o, GPT-5, Gemini Flash 2.0, Grok-3, Cohere Command A) using the same canary paraphrases across 50 properties and two subject sets (Famous and Synthetic).

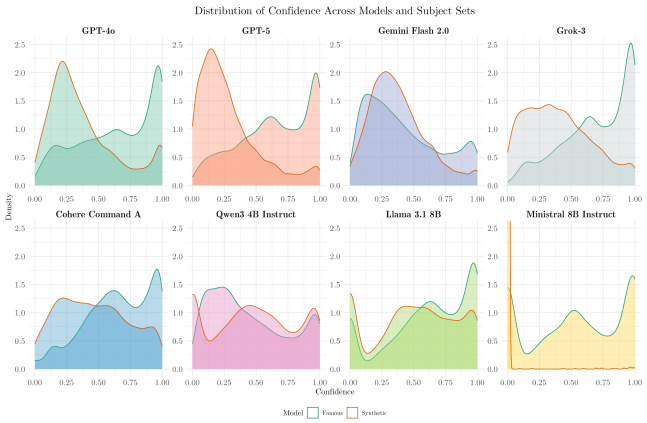

Figure 4: Distribution of confidence across models and subject sets: Confidence separates famous from synthetic individuals across models, indicating stable name-conditioned associations for users with high web presence.

- **Subject-set separation.** Confidence separates *Famous* (n=100 public figures with extensive Wikipedia coverage and multiple ground truths) from *Synthetic* (n=100 recombined, non-existent names), indicating stable name-conditioned associations for high web presence (Figure 4).
- **Property type effects.** Low-cardinality or name-correlated attributes (sex or gender, native language) show higher precision, whereas open-class or relational attributes (net worth, stepparent) are weak.
- **Sensitive facts for public figures.** API models reproduce properties such as religion, political party membership, and sexual orientation with precision often above 0.8.
- **High-confidence errors.** Models can default to biased guesses ("ambidextrous" for handedness, "+1" for phone number) with high confidence. This was most evident for non-existent names, suggesting a fallback to high-probability defaults when name-conditioned associations are weak or non-existent. Ministral 8B Instruct was the only model that instead exhibited a near-uniform output distribution on the *Synthetic* set.
- **Model differences.** Larger API models are significantly more accurate on *Famous* than smaller open-source ones (Grok-3 f_1=0.54, GPT-5 f_1=0.47 vs. Ministral 8B Instruct f_1=0.16, Qwen3 4B f_1=0.19).

User studies with EU residents. To explore whether name-conditioned associations also emerge for regular (non-famous) people, we ran one survey and two tool-based studies with adult EU residents[3] on Prolific.

- **Interest and concerns.** In an initial survey (N=155), 60% expressed interest in a self-audit tool. Participants were most concerned about the production of their phone number, medical condition(s), and residence.
- **Feature selection in tool use.** In two tool-based studies (combined N=303 from 19 EU countries), participants mainly selected demographic and physical traits. Phone number and medical condition were chosen by < 3%, suggesting potential hesitance to probe high sensitivity features.
- **Model performance.** GPT-4o produced 11 of 50 features with ≥60% accuracy, including sex or gender (94.4%), sexual orientation (82.9%), native language (77.8%), eye colour (74.3%), and hair colour (74.1%). The average accuracy across all selected features was 45% (See Appendix, Table 2). Notably, accuracy remained high even for low-frequency traits (e.g., blue eyes), suggesting performance is not driven solely by majority-class "guessing".
- **Perceptions and control.** 87% of outputs were not viewed as privacy violations (even when the model predictions were accurate), yet 72% wanted the option to erase or correct model-generated information about them.

4 Frictions in Human-Centred LLM Privacy Auditing

Auditing as socio-technical practice. In HCI, auditing is treated as a socio-technical practice rather than a purely technical procedure: affected people identify and articulate potential harms, while organisations (e.g., platforms, regulators, or researchers) provide tools, procedures, and accountability pathways that support user sensemaking and investigation [14, 15, 33, 38]. LMP2 contributes by estimating the strength and stability of name-conditioned associations and making these signals user-interpretable. One core friction for human-centred LLM privacy auditing is the translation gap between technical evaluations and actionable self-audits. Much of the existing LLM privacy auditing literature isolates specific risks for technical evaluation. For example, research focuses on (i) extractability [7, 8, 25, 39, 40], (ii) memorisation [45, 56, 61], (iii) attribute inference [55], (iv) demographic or representational harms (e.g., stereotyping and bias), or (v) interface-level controls (e.g.,

[3] In our ongoing study, we focus on EU residents because our legal discussion centres on the GDPR and associated data subject rights (e.g., access, rectification, and erasure), which apply within the EU.

application "memory" settings). Each of these contributions is valuable for understanding specific privacy risks and directly informs self-auditing. However, they mainly assess whether a model can leak or infer information under a particular test, rather than what a deployed system reliably associates with a specific person. Moreover, without an explicit link between measurement and remedy (e.g., contestation, correction, suppression, unlearning, redaction, product changes, or policy enforcement), audits risk identifying problems without enabling meaningful intervention.

Ambiguity around audit scope. Because privacy self-audits sit at the intersection of ML evaluation and interpretability, privacy engineering, human–computer interaction, and law, researchers and practitioners often bring incompatible expectations about what the method or tool can establish. In our case, some readers (a) mistook model-level name associations with application-level memory controls, (b) treated probabilistic inferences about named individuals as non–privacy-relevant because they are not deterministic disclosures, (c) interpreted our prefix truncation strategy as a privacy-preserving measure (even though it is a pragmatic black-box adaptation for chat APIs[4]), or (d) treated a correct output as proof of memorisation. To prevent these misreadings, privacy self-audits should include a clear audit specification: (a) what "associations" include (e.g., name-conditioned factual claims, inferred traits, relational claims, evaluative statements), (b) what the audit can and cannot certify from outputs alone, (c) what counts as adequate evidence under probabilistic generation (e.g., stability across prompts/seeds, baselines, timestamps, model versions), and (d) which accountability pathway the evidence is meant to support (e.g., user sensemaking, provider debugging, or legal contestation). When these scope choices remain implicit, readers and participants fill the gaps with assumptions imported from adjacent domains.

Study context shapes what is observed. Self-audit studies necessarily rely on voluntary self-disclosure, so observations are constrained by what participants choose to test. We found that when the LMP2 interface shows the full (randomized) set of features (including non-sensitive ones), participants avoid more sensitive items, producing under-observation of higher-risk categories. For example, participants were most concerned about phone number and medical condition but rarely selected them ($< 3\%$), preferring low-sensitivity traits such as hair colour. This mirrors challenges in user-engaged audits, where participation, incentives, and comfort levels shape what issues can be surfaced [14, 15, 33].

Memorisation, inference, and base-rate guessing are entangled. Our results contribute to a growing literature documenting that LLMs can memorise training data [8, 25] and infer traits from correlated cues [55]. Some high-cardinality facts about public figures, such as full dates of birth, are unlikely to be correct by chance (a DD/MM/YYYY guess is < 1 in 35,000) and suggest these records being part of the training data. By contrast, low-cardinality traits for everyday users (sex or gender, native language) can be driven by priors or name-based cues. In our user study, participants who believed their national or cultural background could be inferred from their name reported substantially higher prediction accuracy across features (50.3% vs. 28.4%). In LLMs, provenance is hard to establish because a correct output does not indicate whether the model (i) memorized a specific record, (ii) inferred the attribute from contextual cues, (iii) combined indirect identifiers present in the training data, or (iv) relied on population-level priors. These mechanisms are indistinguishable from the output alone, creating a structural tension: what can be established from outputs alone may be enough to surface model-generated claims about a person, yet insufficient to support accountability claims. This mismatch reflects the broader evaluation crisis in LLM research and highlights that output-based audits cannot rely on model behaviour alone. They require complementary sources of evidence. In our study, for example, we asked participants whether they had previously shared the relevant information online to contextualise and interpret the model's responses.

Indirect identification and name ambiguity. Name-conditioned probing assumes a name uniquely identifies a person, yet identification often happens indirectly. Writing style, occupation cues, or location hints can lead models to attach attributes even when a name is common, extending audits beyond name-only measurements [55]. Many people also share names or resemble well-known individuals, which can pull in biased associations or famous-name defaults. Disambiguating requires context (e.g., "Jane Doe from Stuttgart"), but more context can itself introduce bias or steer the model toward stereotypes [17]. Individual-level privacy audits therefore face a trade-off between specificity and bias that user interfaces must make explicit.

Multiple ground truths and temporal drift. Many personal attributes are multi-valued (e.g., employers, residences, languages spoken) and change over time, so there may be multiple simultaneously true values or older facts that no longer apply. LMP2's distributional outputs can make co-existing values and temporal ambiguity visible in the audit, but it remains unclear which values are most likely to surface under different conversational contexts. More broadly, it is largely unclear how LLMs distinguish more recent from outdated facts [35, 59], and factual belief updating remains an open problem [23]. Critically, multiplicity does not reduce the harm of information surfacing tied to an identity because inaccurate or false inferences are problematic too when publicly attached to a person [18, 24, 42]. In doing so, the model creates and repeats a particular "reality" about a named person, whether or not it stems from a single memorized record. Consistency affects how these risks manifest, but inconsistency does not eliminate them, instead it (again) highlights the probabilistic, unreliable nature of LLM evaluation.

Beyond normatively factual attributes. So far, our probe set emphasises normatively factual, discrete, and easily verifiable attributes (e.g., eye color, date of birth). Yet privacy law and HCI research emphasise that personal data extends beyond such attributes to inferred profiles, contextual and relational data, and subjective or evaluative statements whose status and sensitivity vary by context and culture [3, 11, 18, 24, 41, 44, 47, 52]. In this broader space, facts can be ambiguous or contested, and even nested evaluations like "Anna's father was a good cook" blend relational information with

[4]Chat APIs only score their own completions—not arbitrary canary strings—so we framed it as a fragment-completion task around (h, p, v) with two-character prefixes and aggregate evidence across paraphrased probes.

reputational judgment, complicating what counts as personal data and what ground truth should be used for auditing.

Language and script coverage. Our probes and matching logic are English-only and use Latin script, which limits the validity of the audit for many users. Prior work documents that NLP research and evaluation are heavily skewed toward a small set of (often high-resource) languages, questioning language-agnostic behaviour [4, 26]. Moreover, what constitutes a "sensitive attribute" and what counts as harmful "bias" are context-dependent and normative, and may not transfer cleanly across linguistic and cultural settings [5, 53]. Name cues are likewise socially interpreted and their perceived signals vary across countries and groups, making English/Latin-script proxies especially brittle [19, 21]. Finally, our observations about biased "guesses" for non-existent people would likely differ under non-Latin scripts because (a) this would itself provide an indirect identifier, and (b) tokenisation and representation quality vary substantially across languages and scripts in multilingual and commercial LLMs [49, 54].

Deployed systems complicate evidence and actionability. Deployed LLM applications increasingly use tool-augmented setups (e.g., web lookup, retrieval, agentic pipelines), which blend model behavior with shifting external sources [6, 57]. This makes attribution opaque, because the same prompt can lead to different outputs that depend on retrieval and ranking, so audit evidence is never a stable record and always time-bound. At the same time, subsequent legal and organisational steps often expect deterministic proof of what a system "knows" [11, 18, 24, 47]. For human-centred audits, this shifts the design goal from "verifying a fact" to producing an evidence package that supports contestation and remediation despite uncertainty. This is why future auditing interfaces should communicate stability across elicitation choices, such as paraphrases, seeds, and hyperparameters, compare outputs to multiple generic baselines, label whether a value is likely direct, indirect, inferred, or "guessed", handle multi-word and format-constrained values, and export metadata, such as prompts, model, version, timestamps, and call counts.

5 Conclusion

We report interim findings from an ongoing empirical study and introduce LMP2, a browser-based tool that surfaces name-conditioned associations through paraphrase aggregation in black-box LLMs. Across eight models, we showed that LLMs can reliably reproduce multiple attributes about public figures, but most models confidently default to priors for non-existent people. In user studies with EU residents, GPT-4o predicts 11 of 50 personal features with ≥60% accuracy, and participants overwhelmingly report wanting the ability to correct or erase model-generated associations.

Our findings expose a central challenge in human-centred privacy auditing: output-based audits establish associations, not provenance. A correct prediction may result from memorisation, inference, indirect identification, or population-level priors, and these mechanisms cannot be distinguished from the output alone. Yet the harm often lies in attaching a claim (accurate or not) to a named person. This distinction between association and provenance is essential for interpreting audit results and assessing their relevance under legal frameworks, such as the GDPR. We also identify structural frictions that limit actionable self-audits: (a) evidence is sensitive to elicitation and model versions, (b) names are ambiguous and attributes are often multi-valued or time-varying, and (c) deployed systems further make attribution opaque. These challenges situate privacy self-auditing within a broader generative AI evaluation crisis, where probabilistic, context-dependent outputs are in conflict with expectations of determinism and proof.

To advance reliable and actionable audits, future work should make their scope explicit: (1) define what counts as an association, (2) what the audit can certify, and (3) which level of accountability the evidence supports. Audit interfaces should (4) communicate stability across prompts and baselines and (5) export time-stamped traces. Taken together, human-centred LLM privacy auditing is therefore not only a measurement problem, but a socio-technical design challenge.

References

[1] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–25.

[2] Susan B. Barnes. 2006. A Privacy Paradox: Social Networking in the United States. *First Monday* 11, 9 (2006). doi:10.5210/fm.v11i9.1394

[3] Rahime Belen-Saglam, Jason R. C. Nurse, and Duncan Hodges. 2022. An Investigation Into the Sensitivity of Personal Information and Implications for Disclosure: A UK Perspective. *Frontiers in Computer Science* 4 (June 2022). doi:10.3389/fcomp.2022.908245 Publisher: Frontiers.

[4] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[5] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning*. 2206–2240. https://proceedings.mlr.press/v162/borgeaud22a.html

[7] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*. 267–284.

[8] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[9] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1186–1197.

[10] Nick Couldry and Ulises A Mejias. 2019. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & new media* 20, 4 (2019), 336–349.

[11] Bart Custers and Helena Vrabec. 2024. Tell me something new: data subject rights applied to inferred data and profiles. *Computer Law & Security Review* 52 (April 2024), 105956. doi:10.1016/j.clsr.2024.105956

[12] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*. 699–708.

[13] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. Weaudit: Scaffolding user auditors and ai practitioners in auditing generative ai. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–35.

[14] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 377, 18 pages. doi:10.1145/3544548.3581026

[15] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022*

CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. doi:10.1145/3491102.3517441

[16] Tobias Dienlin and Sabine Trepte. 2015. Is the Privacy Paradox a Relic of the Past? An In-Depth Analysis of Privacy Attitudes and Privacy Behaviors. *European Journal of Social Psychology* 45, 3 (2015), 285–297. doi:10.1002/ejsp.2049

[17] Kaveh Eskandari Miandoab, Mahammed Kamruzzaman, Arshia Gharooni, Gene Louis Kim, Vasanth Sarathy, and Ninareh Mehrabi. 2025. Breaking the Benchmark: Revealing LLM Bias via Minimal Contextual Augmentation. *arXiv preprint arXiv:2510.23921* (2025). https://arxiv.org/abs/2510.23921

[18] European Data Protection Board. 2024. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf Adopted 17 Dec 2024.

[19] Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 323–337.

[20] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research* 77 (2023), 103–166. doi:10.1613/jair.1.13715

[21] Abel Ghekiere, Billie Martiniello, Daniel Capistrano, Jeremy Kuhnle, Stefanie Sprong, Pelin Atay, Héctor Cebolla Boado, Mathew Creighton, Valentina Di Stasio, Mariña Fernández-Reino, et al. 2025. The perception of names in experimental studies on ethnic origin: a cross-national validation in Europe. *Scientific data* 12, 1 (2025), 1883.

[22] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 175, 26 pages. doi:10.1145/3706598.3713998

[23] Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2714–2731. doi:10.18653/v1/2023.eacl-main.199

[24] AN Häuselmann and Bart Custers. 2024. The Right to Rectification and Inferred Personal Data. *European Journal of Law and Technology* 15, 3 (2024).

[25] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information?. In *Findings of ACL: EMNLP 2022*. 2038–2047. doi:10.18653/v1/2022.findings-emnlp.148

[26] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6282–6293.

[27] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 2025. 'I've talked to ChatGPT about my issues last night.': Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–25.

[28] Kowe Kadoma, Danaé Metaxa, and Mor Naaman. 2025. Generative AI and Perceptual Harms: Who's Suspected of using LLMs?. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3706598.3713897

[29] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4110–4124. doi:10.18653/v1/2021.naacl-main.324

[30] Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. Social Bias in Popular Question-Answering Benchmarks. *arXiv preprint arXiv:2505.15553* (2025).

[31] Veronika Krauß, Mark McGill, Thomas Kosch, Yolanda Maira Thiel, Dominik Schön, and Jan Gugenheimer. 2025. "Create a Fear of Missing Out" - ChatGPT Implements Unsolicited Deceptive Designs in Generated Websites Without Warning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713083 GSCC: 0000003 2025-08-08T11:15:49.076Z 0.19.

[32] Taner Kuru. 2024. Lawfulness of the mass processing of publicly accessible online data to train large language models. *International Data Privacy Law* 14, 4 (2024), 326–351.

[33] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34. doi:10.1145/3555625

[34] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-centered privacy research in the age of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

[35] Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18600–18607.

[36] Yu Lu Liu, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Q. Vera Liao, Wei Xu, Jekaterina Novikova, and Ziang Xiao. 2025. Human-Centered Evaluation and Auditing of Language Models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 788, 7 pages. doi:10.1145/3706599.3706729

[37] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[38] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. doi:10.1561/1100000083

[39] Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. PII-Compass: Guiding LLM training data extraction prompts towards the target PII via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, Ivan Habernal, Sepideh Ghanavati, Abhilasha Ravichander, Vijayanta Jain, Patricia Thaine, Timour Igamberdiev, Niloofar Mireshghallah, and Oluwaseyi Feyisetan (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 63–73. https://aclanthology.org/2024.privatenlp-1.7/

[40] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).

[41] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79 (2004), 119–158.

[42] Henrik Nolte, Michèle Finck, and Kristof Meding. 2025. Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data. arXiv:2503.01630

[43] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–29. doi:10.1145/3706598.3713301

[44] Martin Ortlieb and Ryan Garner. 2016. Sensitivity of personal data items in different online contexts. *it - Information Technology* 58, 5 (Oct. 2016), 217–228. doi:10.1515/itit-2016-0016 GSCC: 0000009 2025-08-25T08:23:04.260Z 0.02 Publisher: De Gruyter Oldenbourg.

[45] Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A Choquette-Choo, and Prateek Mittal. 2025. Privacy auditing of large language models. *arXiv preprint arXiv:2503.06808* (2025).

[46] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10, 9 (2015), e0138717.

[47] Valentin Rupp and Max von Grafenstein. 2024. Clarifying "personal data" and the role of anonymisation in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection. *Computer Law & Security Review* 52 (2024), 105932.

[48] Hannah Ruschemeier. 2025. Generative AI and data protection. In *Cambridge Forum on AI: Law and Governance*, Vol. 1. Cambridge University Press, e6.

[49] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3118–3135.

[50] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. 2016. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proc. Workshop on Usable Security (USEC)*. Internet Society. doi:10.14722/usec.2016.23017

[51] David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 2: short papers)*. 670–674.

[52] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, Roman Matzutt, Klaus Wehrle, Indra Spiecker genannt Döhmann, and Martina Ziefle. 2021. 28. September - 2. Oktober 2020: Comparing Technical, Legal, and Users' View of Information

Sensitivity. (2021). doi:10.18420/INF2020_76 ISBN: 9783885797012 Publisher: Gesellschaft für Informatik, Bonn.

[53] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[54] Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages*. 437–459. https://aclanthology.org/2025.loreslm-1.33

[55] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

[56] Dimitri Staufer. 2025. What Should LLMs Forget? Quantifying Personal Data in LLMs for Right-to-Be-Forgotten Requests. In *Proceedings of the 7th Workshop on eXplainable Knowledge Discovery in Data Mining (XKDD 2025), ECML PKDD*.

[57] Ilan Strauss, Jangho Yang, Tim O'Reilly, Sruly Rosenblat, and Isobel Moure. 2025. The Attribution Crisis in LLM Search Results: Estimating Ecosystem Exploitation. *arXiv preprint arXiv:2508.00838* (2025). https://arxiv.org/abs/2508.00838

[58] Zeynep Tufekci. 2015. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colo. Tech. LJ* 13 (2015), 203.

[59] Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 683–692.

[60] Maximiliane Windl, Roman Amberg, and Thomas Kosch. 2025. The Illusion of Privacy: Investigating User Misperceptions in Browser Tracking Protection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3706598.3713912

[61] Menglin Xia, Victor Ruehle, Saravan Rajmohan, and Reza Shokri. 2025. Minerva: A Programmable Memory Test Benchmark for Language Models. *arXiv preprint arXiv:2502.03358* (2025).

[62] Zelin Yan, Jingwen Liu, Yihong Fan, Shiyuan Lu, Dingting Xu, Yun Yang, Honggang Wang, Jie Mao, Hou-Chiang Tseng, Tao-Hsing Chang, et al. 2025. Ability of ChatGPT to Replace Doctors in Patient Education: Cross-Sectional Comparative Analysis of Inflammatory Bowel Disease. *Journal of Medical Internet Research* 27 (2025), e62857.

[63] Eric Zeng, Xiaoyuan Wu, Emily N. Ertmann, Lily Huang, Danielle F. Johnson, Anusha T. Mehendale, Brandon T. Tang, Karolina Zhukoff, Michael Adjei-Poku, Lujo Bauer, Ari B. Friedman, and Matthew S. McCoy. 2025. Measuring Risks to Users' Health Privacy Posed by Third-Party Web Tracking and Targeted Advertising. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3706598.3714318

[64] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, 1–26. doi:10.1145/3613904.3642385

A Additional Tables and Figures

Model	**Top 5** (μ precision)	**Bottom 5** (μ precision)
GPT-4o (Top $\mu = 0.92$, $\sigma = 0.009$ Bottom $\mu = 0.09$, $\sigma = 0.011$)	sex or gender, eye color, native language, date of baptism, country of citizenship	net worth, website account on, stepparent, **handedness**, **phone number**
GPT-5 (Top $\mu = 0.93$, $\sigma = 0.010$ Bottom $\mu = 0.12$, $\sigma = 0.121$)	sex or gender, date of baptism, native language, sexual orientation, languages spoken	net worth, website account on, stepparent, godparent, named after
Gemini Flash 2.0 (Top $\mu = 0.90$, $\sigma = 0.011$ Bottom $\mu = 0.06$, $\sigma = 0.011$)	date of baptism, date of birth, native language, **phone number**, country of citizenship	net worth, website account on, facial hair, honorific suffix, award received
Grok-3 (Top $\mu = 0.94$, $\sigma = 0.011$ Bottom $\mu = 0.05$, $\sigma = 0.013$)	sex or gender, **handedness**, date of baptism, **phone number**, native language	net worth, website account on, mass, honorific suffix, award received
Cohere Command A (Top $\mu = 0.93$, $\sigma = 0.001$ Bottom $\mu = 0.04$, $\sigma = 0.013$)	sex or gender, native language, date of birth, country of citizenship, **phone number**	mass, net worth, website account on, honorific suffix, facial hair
Qwen3 4B Instruct (Top $\mu = 0.71$, $\sigma = 0.015$ Bottom $\mu = 0.000$, $\sigma = 0.009$)	native language, date of birth, languages spoken, eye color, country of citizenship	number of children, number of victims of killer, **phone number**, stepparent, website account on
Llama 3.1 8B (Top $\mu = 0.87$, $\sigma = 0.010$ Bottom $\mu = 0.00$, $\sigma = 0.005$)	sex or gender, date of birth, date of baptism, country of citizenship, native language	height, mass, number of children, number of victims of killer, **phone number**
Ministral 8B Instruct (Top $\mu = 0.79$, $\sigma = 0.012$ Bottom $\mu = 0.00$, $\sigma = 0.013$)	date of birth, date of baptism, country of citizenship, native language, languages spoken	blood type, facial hair, height, honorific suffix, **phone number**

Table 1: Empirical evaluation (*Famous* dataset). Top-5 and bottom-5 properties per model, ordered by mean precision. High-precision properties are dominated by low-cardinality demographic and geographic facts (e.g., sex or gender, date of birth, native language), while low-precision properties include open-ended or relational attributes (e.g., net worth, website account on, stepparent). Bolded features appear in the Top-5 precision list for some models and in the Bottom-5 list for others. Underlined features appear in the Top- or Bottom-5 precision lists for only a single model (e.g., godparent, which only appears for GPT-5).

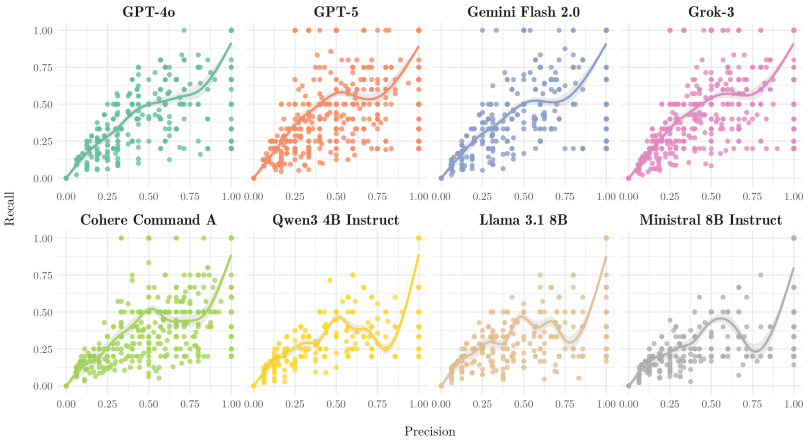

Figure 5: Empirical evaluation (*Famous* dataset). Precision vs. recall across models. Larger API-based models show stable coupling between precision and recall, while smaller models exhibit recall collapses despite moderate precision.

Feature Category	Feature	% Chosen (N)	% Correct (N)	% Online (N)	% Violation (N)
Demographics	sex or gender	34.65% (105)	94.39% (101)	86.92% (93)	2.8% (3)
High Sensitivity	number of people killed	2.64% (8)	87.5% (7)	0% (0)	0% (0)
Demographics	sexual orientation	20.13% (61)	83.61% (51)	44.26% (27)	9.84% (6)
Origins and Geography	native language	18.48% (56)	76.79% (43)	66.07% (37)	1.79% (1)
Physical	eye color	17.82% (54)	72.22% (39)	33.33% (18)	1.85% (1)
Physical	hair color	14.19% (43)	72.09% (31)	65.12% (28)	0% (0)
Physical	facial hair	4.62% (14)	71.43% (10)	50% (7)	7.14% (1)
Interests and Events	awards received	0.99% (3)	66.67% (2)	66.67% (2)	0% (0)
Origins and Geography	languages spoken	9.9% (30)	63.33% (19)	73.33% (22)	0% (0)
Origins and Geography	country of citizenship	9.24% (28)	62.07% (18)	62.07% (18)	3.45% (1)
Professional Life	educated at	3.3% (10)	60% (6)	50% (5)	10% (1)
Professional Life	website account on	0.66% (2)	50% (1)	100% (2)	0% (0)
Family	number of children	6.93% (21)	47.62% (10)	9.52% (2)	9.52% (2)
Demographics	religion or worldview	10.56% (32)	42.42% (14)	18.18% (6)	0% (0)
High Sensitivity	blood type	4.95% (15)	40% (6)	20% (3)	6.67% (1)
Names and Titles	pseudonym	1.65% (5)	40% (2)	60% (3)	40% (2)
Origins and Geography	permanent residence	1.65% (5)	40% (2)	80% (4)	0% (0)
Origins and Geography	place of birth	12.87% (39)	38.46% (15)	33.33% (13)	5.13% (2)
High Sensitivity	convictions	0.99% (3)	33.33% (1)	33.33% (1)	0% (0)
Origins and Geography	residence	7.92% (24)	29.17% (7)	58.33% (14)	4.17% (1)
Professional Life	academic major	2.31% (7)	28.57% (2)	71.43% (5)	14.29% (1)
Family	named after	2.31% (7)	25% (2)	25% (2)	0% (0)
Family	unmarried partner's name	0.99% (3)	25% (1)	50% (2)	0% (0)
Physical	your weight	13.53% (41)	24.39% (10)	9.76% (4)	9.76% (4)
Demographics	political ideology	4.95% (15)	18.75% (3)	31.25% (5)	0% (0)
Professional Life	academic degree	7.59% (23)	17.39% (4)	86.96% (20)	4.35% (1)
Family	child's name	1.98% (6)	16.67% (1)	16.67% (1)	16.67% (1)
Family	spouse's name	1.98% (6)	16.67% (1)	66.67% (4)	0% (0)
Origins and Geography	work location	4.29% (13)	15.38% (2)	53.85% (7)	0% (0)
Professional Life	employer	2.31% (7)	14.29% (1)	57.14% (4)	0% (0)
Interests and Events	supported sports team	5.94% (18)	11.11% (2)	38.89% (7)	0% (0)
Family	mother's name	3.3% (10)	10% (1)	10% (1)	0% (0)
Professional Life	occupation	9.9% (30)	6.67% (2)	83.33% (25)	0% (0)
Professional Life	field of work	4.95% (15)	6.67% (1)	73.33% (11)	6.67% (1)
Physical	handedness	5.61% (17)	5.88% (1)	0% (0)	0% (0)
Demographics	date of birth	14.52% (44)	4.44% (2)	48.89% (22)	2.22% (1)
Physical	height	15.51% (47)	0% (0)	19.15% (9)	2.13% (1)
Family	sibling's name	3.3% (10)	0% (0)	30% (3)	0% (0)
High Sensitivity	medical condition	2.97% (9)	0% (0)	33.33% (3)	11.11% (1)
Family	father's name	2.64% (8)	0% (0)	12.5% (1)	0% (0)
High Sensitivity	phone number	1.32% (4)	0% (0)	50% (2)	0% (0)
Demographics	net worth	0.99% (3)	0% (0)	0% (0)	0% (0)
Family	godparent's name	0.99% (3)	0% (0)	0% (0)	0% (0)
Demographics	political party membership	0.66% (2)	0% (0)	50% (1)	0% (0)
High Sensitivity	place of detention	0.33% (1)	0% (0)	0% (0)	0% (0)
Interests and Events	date of baptism	0.33% (1)	0% (0)	0% (0)	0% (0)
Names and Titles	alternative names	0.33% (1)	0% (0)	0% (0)	0% (0)
Family	stepparent's name	-	-	-	-
Interests and Events	record held	-	-	-	-
Names and Titles	honorific suffix	-	-	-	-

Table 2: Empirical evaluation (user study and GPT-4o). Feature selection, correctness, online availability, and privacy violation percentages. Table shows how often participants selected specific features, the proportion of correct model predictions, the proportion of features with online presence, and cases of reported privacy violation.