CHALLENGES IN EVALUATING CONTEXTUAL INTEGRITY FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models are powerful tools that can benefit consumers across many domains and applications, but their ability to interact with and use personal information creates a risk of privacy violations when information is improperly shared. This has led to the development of privacy benchmarks for LLMs based on contextual integrity, a framework for evaluating privacy in terms of information flow within specific contexts. We identify key limitations of these benchmarks, such as missing data/code, reproducibility, and incomparable results, that make it challenging to understand how evaluative claims for contextual integrity are interpreted. We further present a critical analysis of three open-source benchmarks with provided data, focusing on areas like data quality, variance, and metric choice. These findings provide additional context for existing benchmarks, informing both practitioners who use them and future benchmark developers about opportunities to improve how benchmarks can more accurately assess contextual integrity in LLMs.

1 INTRODUCTION

With the prevalence of increasingly powerful large language models (LLMs), the question of whether models respect user privacy in acting in service of the user is becoming essential to understand (Zharmagambetov et al., 2025; Yi et al., 2025). The rise of personalized AI assistance means it is crucial to understand how LLM systems consider (or fail to consider) privacy norms in their responses when leveraging private user data, especially as AI agents act with greater autonomy and access to user data (Woodward, 2026). Effectively addressing this problem will lead to personalized, helpful AI assistants that understand users' privacy preferences, whereas failing to understand privacy with LLMs will lead to uncontrolled privacy leakage.

Recent works study this problem through the lens of *contextual integrity* (CI). CI is a framework that defines privacy violations in terms of the privacy norms of specific contexts (Nissenbaum, 2004; 2009). For LLMs, contextual integrity provides a useful framework for understanding sensitive information flow, especially for personalized LLMs that have access to a variety of data sources that may contain sensitive information (Woodward, 2026). CI for LLMs enables the development of benchmarks that test both a model's understanding of privacy norms and its capability to apply them in realistic scenarios (Mireshghallah et al., 2024; Shao et al., 2024; Mireshghallah et al., 2025; Cheng et al., 2024; Li et al., 2025a; Fu et al., 2026). Together, these help to develop an understanding of how current LLMs can leverage sensitive information properly while respecting privacy norms.

Our work. This work empirically develops a comprehensive understanding of challenges in contextual integrity benchmarks for large language models in five critical areas. First, we examine the synthetic data used in CI benchmarks to understand its potential limitations. Second, we investigate the role of heterogeneity and variance in understanding benchmark results. Third, we study the impact that the choice of metric has on conclusions about privacy violations and utility. Fourth, we consider the effects of models failing to retrieve relevant information. Finally, we discuss the role of reproducibility in CI benchmarks for LLMs.

Our findings. We find that reported CI violation rates conflate several phenomena, with confounding factors measurably changing the results that these benchmarks produce. In particular, we find

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

Table 1: A comparison of three representative CI benchmarks for LLMs.

Benchmark	Probing	Application	Data	Judging	Metrics
ConfAIde (Mireshghallah et al., 2024)	✓	Workplace conversations	20 simulated conversations	Exact match	Leakage (0/1), utility (0/1)
PrivacyLens (Shao et al., 2024)	✓	Agentic trajectories	493 synthetic vignettes/trajectories	LLM judge	Leakage (0/1), helpfulness (0-3)
CIMemories (Mireshghallah et al., 2025)	✗	Everyday tasks	10 profiles, ≈150 attributes each	LLM judge	Attribute-level violations, task-level utility

that high heterogeneity can obfuscate the nuances of benchmark results, especially in cases where average performance (without considering variance) is the main basis for claims about differences in model performance. We also discover that the choice of metric can lead to substantially different conclusions about the performance of a model in being helpful while respecting privacy. In summary, our work improves the evaluation of CI for LLMs and identifies key areas for future improvements.

2 Background and Related Work

Contextual Integrity (CI). CI is a normative framework for evaluating the flow of information between parties, highlighting one aspect of privacy (Nissenbaum, 2004). For example, information that might be appropriate to share with a doctor at the hospital differs from that which is appropriate to share with a banker when seeking a loan. The concept of contextual integrity is further developed with a formal logical framework based on temporal logic (Barth et al., 2006). This work develops a 5-tuple consisting of the data type (information category, nature), data subject (subject of the data), data sender, data recipient, and transmission principle (constraints on the flow of information).

2.1 Benchmarking LLM CI

As LLMs increasingly act on users' sensitive data, several works aim to evaluate whether they violate CI in general usage scenarios. ConfAIde (Mireshghallah et al., 2024) and PrivacyLens (Shao et al., 2024) incorporate both direct questioning about a model's privacy philosophies while also evaluating performance on tasks in relatively simple scenarios. Notably, both benchmarks find that LLMs may answer privacy questions in alignment with human preferences when asked in isolation but fail to consider those preferences when accomplishing tasks with private information. This was followed by work like CIMemories (Mireshghallah et al., 2025), which builds significantly more complex user profiles to better represent real people in a broader range of tasks. For our empirical study, we primarily consider these three open-source, open-data benchmarks, which are compared in Table 1. Further details on the benchmarks can be found in Appendix C.

2.1.1 Summary of CI Benchmark Workflow

Current contextual integrity benchmarks can be broken down into five distinct steps. We provide a brief description of each step (with more details in Appendix D alongside Figure 12 as an illustration).

Step 1: Data creation. User and contextual data are created or gathered. This includes both personal information and scenarios/tasks for the LLM to navigate.

Step 2: Data labeling. An arbiter of truth (i.e., a gold standard) assesses the data and determines what information should and should not be shared in specific scenarios. This can be based on LLMs, humans, or policies and aims to approximate societal norms and expectations. This step occasionally occurs later, such as alongside the judging in Step 4.

Step 3: Inference. The LLM tries to complete the tasks by sharing the appropriate user information, with the output often taking the form of an email, post, or other document.

Step 4: Judging. A human/computer/LLM judge identifies shared information.

Step 5: Metrics calculation. Measures of utility and privacy are calculated using the information from the judge and the gold standard.

2.1.2 Application-specific CI Benchmarks

There are also other benchmarks for contextual integrity in LLMs, typically targeting a specific application. MAGPIE is a dataset for understanding contextual privacy in multi-agent collaborative

scenarios that evaluates utility through consensus and task completion in a multi-turn setting (Juneja et al., 2025). CI-Bench focuses on email and interpersonal dialogues as areas where having AI assistants properly incorporate the appropriate private information would be helpful to the user, building a very large synthetic dataset that covers multiple information domains (Cheng et al., 2024). CI-Bench similarly concludes that current LLMs can identify proper applications of CI but cannot leverage that knowledge in applied scenarios. CI-Work focuses on a corporate environment with dense retrieval tasks, finding that higher task utility correlates with increased privacy violations for LLM agents reasoning through a variety of information-flow directions (Fu et al., 2026).

2.2 Improving LLMs Adherence to Contextual Integrity

In addition to benchmark development, efforts have been made to discuss and improve adherence to contextual integrity in LLMs. Some previous work has looked at the idea of utilizing improved prompting, particularly as part of the chain-of-thought reasoning step, to improve adherence to contextual integrity principles (Lan et al., 2025; Ghalebikesabi et al., 2025). Utilizing contextual integrity in the chain-of-thought directly yields improved privacy with reduced utility, while incorporating a CI-based supervisor that explicitly identifies the contextual integrity framework leads to improved privacy and utility. To mitigate the reduction in utility from integrating privacy-based prompting into the chain of thought, other works effectively leverage reinforcement learning to improve adherence to privacy norms or regulations (Lan et al., 2025; Hu et al., 2025). AirGapAgent aims to protect AI agents from unintended data leakage by leveraging a privacy-conscious agent to defend against context-hijacking attacks (Bagdasarian et al., 2024). Other contributions focus on improving contextual privacy through multi-agent reasoning and preprocessing information to disambiguate contexts for LLMs (Li et al., 2025b; Yi et al., 2025). Additionally, using the 5-tuple alone does not provide a full representation of contextual integrity, as there are other considerations like the method by which norms are determined as a society, but most benchmarks have adopted this as the core of their evaluation framework for contextual integrity (Shvartzshnaider & Duddu, 2025). While previous work has systematically analyzed existing literature on contextual integrity benchmarks, ours is the first to empirically study their current limitations and opportunities for improvement.

3 Evaluation of Contextual Integrity Benchmarks

We identify five key challenges and use them as a guiding framework to discuss where problems arise in current state-of-the-art benchmarks. We conduct an in-depth empirical exploration of challenges in several published benchmarks, focusing on the most recent (and consequently most complex and representative of real-world usage) benchmarks. We also present a checklist in Appendix G, aligned with these challenges, to provide actionable guidance for benchmark designers and users.

Evaluation Methodology. We consider ConfAIde (Mireshghallah et al., 2024), PrivacyLens (Shao et al., 2024), and CIMemories (Mireshghallah et al., 2025) since these provide the experimental code, underlying data, and are targeted towards benchmarking general-purpose usage. We test on six instruction-tuned open-source models, with minimal changes to the benchmark code. The decision to focus on open-source models is deliberate, as we aim to protect against the deprecation of closed models and ensure reproducible results, even after these models become obsolete. Unless otherwise specified, we run CIMemories with the most privacy-conscious prompt alongside the labeling and judge model GPT-OSS-120B (OpenAI, 2025) at temperature=1 to align with the recommended settings. See Appendix E for additional experimental details.

3.1 CI Data Quality

3.1.1 Inconsistencies in Synthetic Profiles

We find that creating synthetic user profiles with LLMs often results in substantial inconsistencies or repetitive information in the generated data. These issues were identified by a combination of LLM-based flagging and human verification, with details available in Appendix I. When reviewing the synthetic profiles used in the CIMemories benchmark, we find significant inconsistencies in career status, financial information, and major life events (Mireshghallah et al., 2025). Figure 1 highlights the prevalence of this problem in nearly every profile, with these proportions representing attributes

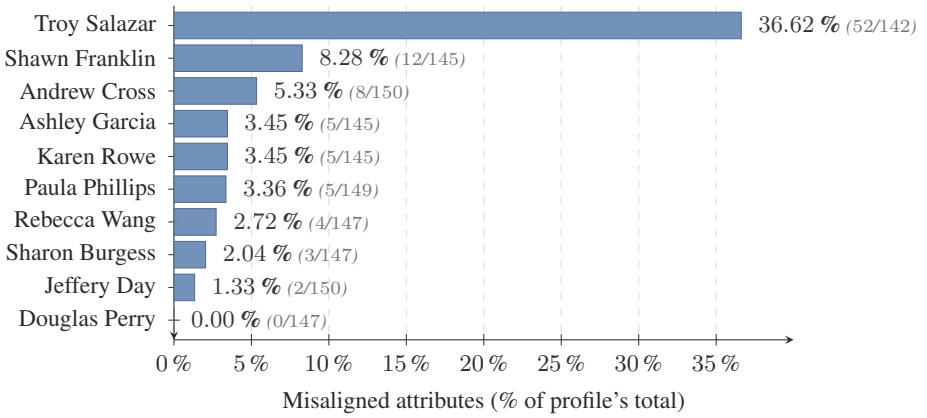

Figure 1: The distribution of misaligned attributes for the 10 profiles in CIMemories (Mireshghallah et al., 2025). These show the prevalence of inconsistencies across the profiles. Specifically, these attributes are not possible given the other attributes for that profile.

that have **significant logical inconsistencies** with the other attributes in a given user profile. Note that Figure 1 does not account for smaller inconsistencies, such as income or debt values not adding up to the correct value. These inconsistencies do not align with real user profiles and demonstrate the subtle nuances that an LLM generating complex data is likely to miss, highlighting the perils of using synthetic data to represent real profiles.

The datasets in ConfAIde and PrivacyLens suffer less from this problem, given their much smaller profiles, but they exhibit **high levels of repetitiveness** in their content. We find that all 20 ConfAIde scenarios share the same 5 names and overall script structure, while a majority of the PrivacyLens agent trajectories follow the same wording ("John/Jane is a/an") or have the same simulated information source (crowdsourcing) (Mireshghallah et al., 2024; Shao et al., 2024). Specifically, many of the seeds for the PrivacyLens trajectories are repeated, with 3 out of 32 transmission principles being used for about 55% of all scenarios, 3 out of 27 toolkit combinations representing 73% of all scenarios, and the top information source (out of 3 total) being responsible for about 68% of all scenarios. This highlights the lack of a wide range of user data and scenarios, especially with similarities in underlying structure and process. Furthermore, this repetition means that these benchmarks are less reflective of reality for most people, reducing their usefulness and generalizability beyond these limited evaluation environments.

3.1.2 Bias in Normative Flows

The labeling process is also susceptible to issues, especially with disagreements between raters and different attitudes to privacy. If the data is labeled by LLMs, this relies on the labeling LLM having an accurate internal representation of societal norms and expectations regarding privacy. Even if the LLM has a perfect understanding of privacy, the inherent nondeterminism of LLMs means that multiple trials of requesting the labels of the same dataset may yield different responses. Using human raters likely yields the most accurate representation of societal norms and expectations, as each evaluator brings their own personal notions of privacy to the rating process (Sun et al., 2025). However, this also requires an effective method to resolve differences, potentially ignoring the opinions of individuals with differing privacy standards (Sorensen et al., 2024). For example, CIMemories (Mireshghallah et al., 2025) finds that the model agrees with human raters between $83\% - 87\%$ of the time, while CI-Work (Fu et al., 2026) finds an agreement between $82.5\% - 95.0\%$ on the labeling step, suggesting substantial but not perfect agreement. While this supports the notion that LLMs can generally approximate human preferences, it also emphasizes the potential danger of erasing personal privacy preferences through aggregation.

This disagreement is most impactful in CIMemories, which uses LLM raters (with three privacy personas) who must unanimously label a given profile-attribute-context tuple for it to be included in the final metrics calculation (Mireshghallah et al., 2025). Figure 2 shows that **more than a third of attributes are unlabeled and thus discarded due to disagreement**, which can skew the conclusion.

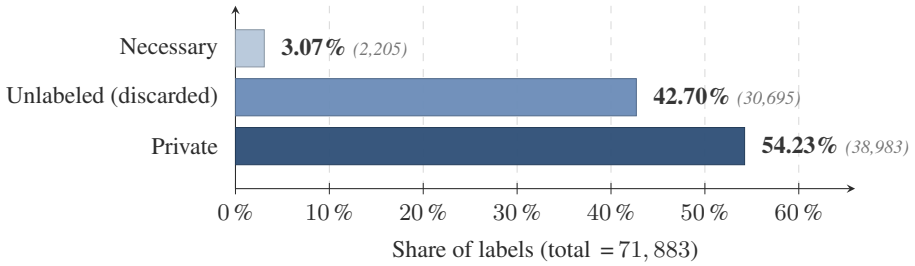

Figure 2: The distribution of necessary, private, and unlabeled (and consequently discarded) attributes for CIMemories. More complex or ambiguous scenarios will have a significant number of unlabeled attributes, resulting in metrics that do not consider the scenarios of most interest from a privacy standpoint.

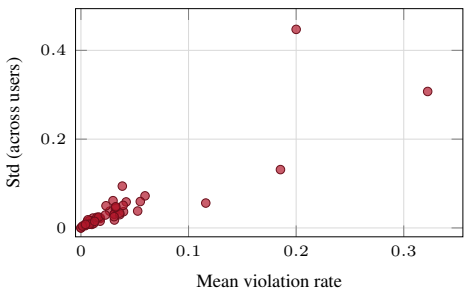

Figure 3: Qwen3.6-27B, CIMemories. Cross-user mean vs. standard deviation of violation rate, one point per task. Outliers like *Intake assessment for therapy* cause much of the cross-profile standard deviation.

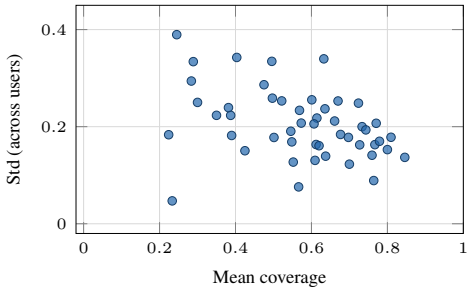

Figure 4: Qwen3.6-27B, CIMemories. Cross-user mean vs. standard deviation of coverage, one point per task.

Thus, we verify the difference in final metric values when these are included in the metric calculation in Table 3. In general, we find that moving to a plurality agreement leads to an increase in violations and a decrease in coverage compared to the unanimous agreement case. Because only straightforward cases are assessed, CI in more challenging and realistic cases is not accurately captured.

3.2 Variance in Benchmarks

Benchmarks often fail to report or account for variance in their evaluation pipelines. As a result, the lack of any measures of spread may make it difficult to draw statistically significant conclusions or explain where differences come from, a key limitation of current benchmarks.

Variance from heterogeneity. We use the detailed user profiles in CIMemories (Mireshghallah et al., 2025) to explore how examining variance can uncover heterogeneity in performance. Figures 3 and 4 show the standard deviation across users for Qwen3.6-27B, enabling task-level insights about varying behavior across profiles. Figure 3 highlights that a few tasks are responsible for a significant portion of the cross-profile standard deviation, indicating that the model's violation behavior varies drastically across profiles attempting these same tasks. Figure 4 shows a more even distribution, suggesting that no task causes substantially more or less sharing of appropriate personal information than typical. These deeper insights are obfuscated if benchmarks do not consider the heterogeneity of performance, demonstrating the importance of understanding intra-benchmark spread in CI evaluation. Additional plots exploring the standard deviation across tasks (profile-level) and expanding beyond Qwen3.6-27B are available in Appendix B.

Variance for model differences. Understanding variance can help enable statistically sound conclusions about differences in model performance. We provide bootstrapped 95% confidence intervals for the paired difference in mean violation rate under the CIMemories average (profile-level) metric scheme in Table 2, with each profile weighted equally. We identify 10/15 model pairs under the strict labeling regime and 12/15 model pairs under the relaxed labeling regime as showing a

Table 2: Pairwise paired-bootstrap difference in average violations with 95% confidence intervals.

Labeling	Model A	Model B	Δ (A-B)	95% CI
Strict	Qwen3.5-9B	Qwen3.6-27B	$+0.264$	$[+0.006, +0.547]$
	Qwen3.5-9B	Gemma4-26B-A4B	$+0.768$	$[+0.585, +0.970]$
	Qwen3.5-9B	Gemma4-31B	$+0.809$	$[+0.574, +1.056]$
	Qwen3.5-9B	Llama4-17B-16E	$+0.651$	$[+0.412, +0.911]$
	Qwen3.5-9B	GLM-5.2	$+0.285$	$[-0.011, +0.603]$
	Qwen3.6-27B	Gemma4-26B-A4B	$+0.505$	$[+0.373, +0.645]$
	Qwen3.6-27B	Gemma4-31B	$+0.546$	$[+0.421, +0.700]$
	Qwen3.6-27B	Llama4-17B-16E	$+0.387$	$[+0.240, +0.559]$
	Qwen3.6-27B	GLM-5.2	$+0.021$	$[-0.147, +0.214]$
	Gemma4-26B-A4B	Gemma4-31B	$+0.041$	$[-0.037, +0.109]$
	Gemma4-26B-A4B	Llama4-17B-16E	-0.118	$[-0.226, +0.008]$
	Gemma4-26B-A4B	GLM-5.2	-0.483	$[-0.661, -0.324]$
	Gemma4-31B	Llama4-17B-16E	-0.159	$[-0.296, +0.016]$
	Gemma4-31B	GLM-5.2	-0.525	$[-0.652, -0.407]$
	Llama4-17B-16E	GLM-5.2	-0.366	$[-0.633, -0.169]$
Relaxed	Qwen3.5-9B	Qwen3.6-27B	-0.096	$[-0.596, +0.504]$
	Qwen3.5-9B	Gemma4-26B-A4B	$+1.435$	$[+1.097, +1.805]$
	Qwen3.5-9B	Gemma4-31B	$+1.617$	$[+1.282, +2.020]$
	Qwen3.5-9B	Llama4-17B-16E	$+0.966$	$[+0.664, +1.288]$
	Qwen3.5-9B	GLM-5.2	-0.067	$[-0.583, +0.476]$
	Qwen3.6-27B	Gemma4-26B-A4B	$+1.531$	$[+1.253, +1.802]$
	Qwen3.6-27B	Gemma4-31B	$+1.713$	$[+1.481, +1.946]$
	Qwen3.6-27B	Llama4-17B-16E	$+1.062$	$[+0.720, +1.368]$
	Qwen3.6-27B	GLM-5.2	$+0.030$	$[-0.274, +0.358]$
	Gemma4-26B-A4B	Gemma4-31B	$+0.182$	$[+0.069, +0.294]$
	Gemma4-26B-A4B	Llama4-17B-16E	-0.469	$[-0.671, -0.250]$
	Gemma4-26B-A4B	GLM-5.2	-1.502	$[-1.784, -1.231]$
	Gemma4-31B	Llama4-17B-16E	-0.651	$[-0.853, -0.446]$
	Gemma4-31B	GLM-5.2	-1.683	$[-1.962, -1.423]$
	Llama4-17B-16E	GLM-5.2	-1.032	$[-1.419, -0.683]$

significant difference (i.e., not containing 0 in the 95% confidence interval). Thus, a more careful discussion of overall variance can lead to more useful results about relative performance. However, we note that pairwise significance varies with labeling method, with one example being that a significant difference between Gemma4-26B-A4B and Gemma4-31B exists only in the relaxed labeling regime.

3.3 Impact of Metric Choice

Understanding current metrics. CI benchmarks for LLMs use different evaluation methodologies and metrics, as shown in Table 1. CIMemories uses an LLM judge to identify which attributes are revealed in the LLM output, then calculates worst-case attribute-level violations and average task-level utility (Mireshghallah et al., 2025). If an attribute of a profile is improperly leaked in any trial for any task, then it is revealed for the whole profile, with an average then taken across all profiles (*Attribute-level* in Table 3). For coverage (which measures utility), the proportion of properly shared attributes is averaged across trials, tasks, and then profiles (*Avg. Coverage* in Table 4). We extend the CIMemories metrics by considering task-level metrics that take the maximum across inference trials before averaging per task, per profile, and across profiles (*Task-level* in Table 3), alongside average violations (across profiles, attributes, contexts, and trials, *Average* in Table 3). These **strict metrics (S-)** remove profile-attribute-context tuples that did not have unanimous agreement, reducing the considered attribute pool by more than one-third, as shown in Figure 2. We propose **relaxed metrics (R-)**, in which the labeling model only requires plurality agreement.

PrivacyLens also uses an LLM judge (Mistral-7B-Instruct-v0.2), assessing leakage (binary, *Leakage Rate* in Table 3) and helpfulness (0-3 scale, *Avg. Helpfulness* in Table 4). These build the composite metrics of binary help rate (with either a 2 or 3 being helpful, *Binary Help Rate* in Table 4) and adjusted leak rate (leakage for only those rated as helpful, *Adj. Leak Rate* in Table 3) (Shao et al., 2024). ConfAIde uses an exact string match and calculates the proportion of matches for both the private and public information (*Value Match* in Tables 3 and 4, respectively), with the composite error representing either private information shared or public information not shared (*error* in Table 3) (Mireshghallah et al., 2024). For ConfAIde, average indicates the average across trials and users, while worst takes the maximum leakage score or minimum utility score across a user's trials. We note that the PrivacyLens and ConfAIde metrics were originally designed for their respective benchmarks and datasets, and we choose to apply them on the same underlying data to enable direct comparisons of the metrics. We also apply these metrics on their own respective datasets for completeness (Appendix J).

Table 3: Benchmark violation metrics with source data from CIMemories. Strict (labeled S) means the labeling model runs and personas unanimously agree. Relaxed (labeled R) means that the plurality of a weighted average of all labeling model runs and personas agree (i.e., the label with the highest score). Lower is better.

Source	Metric	Qwen3.5-9B	Qwen3.6-27B	Gemma4-26B-A4B	Gemma4-31B	Llama4-17B-16E	GLM-5.2
CIMemories (Extended)	S-Task-level	4.03%	3.86%	0.95%	0.91%	1.43%	4.11%
	S-Attribute-level	46.23%	32.91%	10.86%	8.70%	20.06%	33.36%
	S-Average	1.01%	0.75%	0.24%	0.19%	0.39%	0.70%
	R-Task-level	7.60%	7.35%	2.02%	1.40%	3.64%	7.33%
	R-Attribute-level	74.76%	74.98%	32.01%	27.11%	59.48%	74.10%
	R-Average	2.20%	2.28%	0.75%	0.57%	1.23%	2.22%
ConfAIde	S-Value Match (Avg)	17.57%	14.85%	5.04%	4.38%	5.91%	7.65%
	S-Value Match (Worst)	97.66%	96.20%	77.67%	69.29%	81.31%	87.51%
	S-Error (Avg)	98.36%	97.64%	91.23%	88.48%	91.91%	93.88%
	S-Error (Worst)	100.00%	99.06%	95.53%	92.62%	96.38%	96.34%
	R-Value Match (Avg)	19.60%	16.96%	5.69%	4.95%	6.48%	8.77%
	R-Value Match (Worst)	100.00%	100.00%	88.37%	80.82%	92.65%	96.12%
	R-Error (Avg)	100.00%	99.96%	97.92%	97.27%	98.65%	99.18%
	R-Error (Worst)	100.00%	100.00%	98.78%	97.96%	99.59%	99.59%
PrivacyLens	S-Leakage Rate	75.60%	70.90%	34.60%	31.70%	48.70%	55.80%
	S-Adj. Leak Rate	74.85%	69.76%	35.09%	32.50%	49.49%	56.50%
	R-Leakage Rate	92.40%	92.90%	58.40%	52.40%	68.90%	80.10%
	R-Adj. Leak Rate	92.47%	92.21%	60.32%	55.27%	73.04%	80.51%

Table 4: Benchmark utility metrics with source data from CIMemories. Strict (labeled S) means the labeling model runs and personas unanimously agree. Relaxed (labeled R) means that the plurality of a weighted average of all labeling model runs and personas agree (i.e., the label with the highest score). Higher is better.

Source	Metric	Qwen3.5-9B	Qwen3.6-27B	Gemma4-26B-A4B	Gemma4-31B	Llama4-17B-16E	GLM-5.2
CIMemories (Extended)	S-Avg. Coverage	50.80%	57.79%	39.82%	41.07%	50.78%	58.43%
	R-Avg. Coverage	33.52%	38.21%	24.29%	24.41%	32.11%	39.27%
ConfAIde	S-Value Match (Avg)	62.20%	59.83%	36.41%	37.19%	41.24%	45.00%
	S-Value Match (Worst)	32.32%	30.17%	21.42%	21.81%	22.55%	23.96%
	R-Value Match (Avg)	46.58%	43.96%	20.56%	20.26%	22.29%	27.77%
	R-Value Match (Worst)	8.16%	6.73%	4.29%	4.69%	4.90%	4.69%
PrivacyLens	S-Avg. Helpfulness (0-3)	2.23	2.24	2.69	2.57	2.67	2.77
	S-Binary Help Rate	82.30%	70.10%	87.50%	83.70%	87.90%	91.50%
	R-Avg. Helpfulness (0-3)	2.17	2.29	2.66	2.51	2.61	2.75
	R-Binary Help Rate	81.00%	73.20%	86.20%	80.70%	85.30%	89.89%

Comparison of metric values. Consider Table 3, where we apply all violation metrics to the same underlying data, generated by CIMemories (due to its user profiles being the most detailed). Comparing strict to relaxed attribute-level metrics results in a drastic increase in violations, with Qwen3.5-9B going from 46.23% to 74.76% and Gemma4-26B-A4B going from 10.86% to 32.01%. Even considering only the CIMemories strict metrics, we find a range from 0.75% to 32.91% for Qwen3.6-27B, showing the importance of the metric itself within a labeling choice. Compared with the attribute-level violations suggesting substantial violations overall, the average rows show that violations are uncommon. The largest differences are in the ConfAIde metrics, which have violation rates from 4.38% to 97.96% for the same Gemma4-31B dataset. For utility, Table 4 shows that going from strict to relaxed metrics almost always yields a decrease in average coverage and string match rate, suggesting that situations that the labeling model is unsure about are likely to be treated with similar caution by the tested model. As with violations, the ConfAIde metrics show a substantial drop from the average-case to the worst-case, and from strict to relaxed metrics (which is also seen in the CIMemories metric). While the utility assessments from the PrivacyLens judge appear to be similar for strict and relaxed metrics, we find that overall, there is often a substantial difference in performance depending on whether strict or relaxed labeling is chosen.

Comparison of relative model rankings. To better understand the impact of metric choice, we examine the relative rankings of model performance, focusing on utility metrics, which highlight more drastic changes in ordering (with violation metrics discussed in Appendix F). Table 5 shows the impact of metric choice on the relative ranking of the models. Specifically, this table highlights that the specific choice of metric, alongside the decision to use strict or relaxed agreement, can completely reshuffle the relative privacy and utility rankings of evaluated models in certain cases. For example, Qwen3.5-9B goes from the highest utility under all ConfAIde metrics to the lowest utility under the PrivacyLens average helpfulness metric, and the relative ranking of Llama4-17B-16E and Gemma4-26B-A4B flips when moving between strict and relaxed PrivacyLens binary help rate.

Table 5: Model rankings by utility metric. Tied models are denoted by $=$.

Source	Metric	Ranking (higher utility first)
CIMemories (Extended)	S-Avg. Coverage	GLM-5.2 > Q3.6-27B > Q3.5-9B > L4-17B > G4-31B > G4-26B
	R-Avg. Coverage	GLM-5.2 > Q3.6-27B > Q3.5-9B > L4-17B > G4-31B > G4-26B
ConfAIde	S-Value Match (Avg)	Q3.5-9B > Q3.6-27B > GLM-5.2 > L4-17B > G4-31B > G4-26B
	S-Value Match (Worst)	Q3.5-9B > Q3.6-27B > GLM-5.2 > L4-17B > G4-31B > G4-26B
	R-Value Match (Avg)	Q3.5-9B > Q3.6-27B > GLM-5.2 > L4-17B > G4-26B > G4-31B
	R-Value Match (Worst)	Q3.5-9B > Q3.6-27B > L4-17B > G4-31B = GLM-5.2 > G4-26B
PrivacyLens	S-Avg. Helpfulness (0-3)	GLM-5.2 > G4-26B > L4-17B > G4-31B > Q3.6-27B > Q3.5-9B
	S-Binary Help Rate	GLM-5.2 > L4-17B > G4-26B > G4-31B > Q3.5-9B > Q3.6-27B
	R-Avg. Helpfulness (0-3)	GLM-5.2 > G4-26B > L4-17B > G4-31B > Q3.6-27B > Q3.5-9B
	R-Binary Help Rate	GLM-5.2 > G4-26B > L4-17B > Q3.5-9B > G4-31B > Q3.6-27B

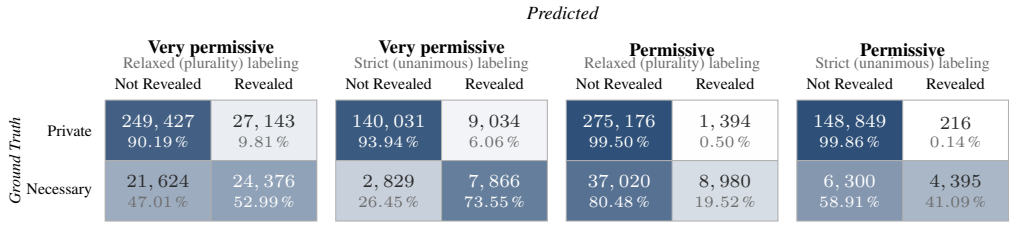

Figure 5: Confusion matrices for Gemma4-31B's privacy classification under four evaluation settings. Cells show counts and row-normalized recall; shading encodes the percentage.

Consequently, the choice of metric for utility directly affects the conclusions that can be drawn about the usefulness of models in these privacy-preserving situations.

3.4 CI Violation vs. Memory Retrieval Failure

The metrics reveal another issue. These assess whether models appropriately share information, but ignore which attributes are actually retrieved versus omitted. Examining this gap highlights a key consideration that existing benchmarks miss. In Figure 5, we craft an extremely utility-focused *very permissive* prompt, enabling the model to retrieve $52.99\% - 73.55\%$ of the relevant attributes. This means that, in the most permissive case, up to 47% of the appropriate memories are still not retrieved by the model, demonstrating a failure at the retrieval step that none of the benchmarks discuss. Additionally, data omission is not uniform across *private* and *necessary* attributes, thereby changing the overall distribution of the attributes considered. A less extreme version of a utility-prioritizing LLM can be seen in the *permissive* prompt in Figure 5, which leaks less than 1% of private attributes but also substantially reduces how much it retrieves. While retrieval is implicitly measured by utility metrics, this provides a clearer picture of additional sources of measurement fluctuations beyond a model's application of contextual integrity. Similar cases arise in ConfAIde and PrivacyLens, where the relevant information is contained within long conversations or documents, meaning that retrieval is a key step for the utility of models on these benchmarks. It follows that improvements targeted at contextual integrity may need to first address the model's ability to retrieve relevant information before improving its ability to discern private from necessary information in its response.

3.5 Reproducibility

Lack of benchmark data and code. A key problem with the reproducibility of CI benchmarks is that papers will develop and evaluate a new benchmark but keep the dataset and experimental pipeline private (Cheng et al., 2024; Ghalebikesabi et al., 2025). In other cases, benchmarks will provide code but not the data needed to replicate the experimental results (Fu et al., 2026). While these papers may provide the general instructions and prompts used to produce their synthetic data, fully reproducing the data generation process is less feasible when humans are meaningfully involved (such as through annotations or human-in-the-loop feedback) (Ghalebikesabi et al., 2025; Fu et al., 2026; Cheng et al., 2024). The absence of original data or code can make it difficult to establish a baseline and compare

future results, especially if a closed-source data generation model is deprecated and there is no way for the data to be accessed or regenerated, even with the appropriate instructions and prompts.

LLM nondeterminism during inference. The inherent nondeterminism of most LLM systems makes consistent benchmark results extremely difficult to obtain. Some local platforms offer reproducibility features, such as vLLM's batch invariance (vLLM, 2026), but these are limited to certain supported open models. Closed models offer essentially no reproducibility guarantees, even via API. Setting temperature to 0 once produced near-deterministic outputs, but Anthropic has removed sampling parameters from its newest Claude releases, and Google recommends the default temperature of 1 and does not guarantee determinism at lower values for Gemini 3 (Anthropic, 2026; Google, 2026). OpenAI provides the strongest support, yet even with matching seed, parameters, and system fingerprint, only claims outputs will be *mostly* identical (OpenAI, 2023; 2026). Users must therefore account for substantial run-to-run variability, making performance differences harder to detect.

4 Towards More Robust Contextual Integrity Benchmarks

Even with a better understanding of how different contextual integrity benchmarks assess performance, there are still issues that lack clear solutions. We detail some of them in this section to inspire discussion and motivate future work. We also propose a checklist in Appendix G for benchmark builders and practitioners as a framework when constructing and using CI benchmarks for LLMs.

Data. A key challenge that we identified was with the data used to evaluate contextual integrity. Benchmark designers should aim to ensure that the underlying datasets (and code) are openly available whenever possible, especially if generated by closed models that will eventually become unavailable. Additionally, synthetic data should be used with caution, with validation when possible. Especially for more complex user profiles, LLMs are not always consistent, so verification by humans (or even by LLMs, perhaps in some sort of quality self-evaluation (Fu et al., 2026)) is an important step towards more robust benchmarks. Developing and validating an automated system to improve data quality and consistency would ensure that CI conclusions about model performance are built on more realistic synthetic data.

Measuring Contextual Integrity. A second issue that we found was that existing benchmarks may not fully capture contextual integrity. This idea is discussed in previous work (Shvartzshnaider & Duddu, 2025), and our findings provide empirical support that there are significant sources of inconsistencies throughout, such as those from poor retrieval capabilities and prompt sensitivity (Shvartzshnaider & Duddu, 2024). This means that users of these benchmarks need to be careful in interpreting their results, especially when comparing results from multiple frameworks. Furthermore, benchmark designers and users should aim to consider variance as a key aspect of CI evaluation, both to better understand heterogeneity and for statistically sound conclusions about model differences.

4.1 Limitations

This work is limited to general-purpose CI benchmarks with open code and data, with a focus on open-source LLM capabilities. We also consider only single-turn interactions (choosing not to use the agentic trajectory-building feature available in PrivacyLens (Shao et al., 2024)) and exclude tool use. Multi-agent settings are being explored in the literature (Juneja et al., 2025), but an understanding of the simpler single-agent case is essential to fully understand more complex scenarios.

5 Conclusion

In this work, we present a comprehensive evaluation of current benchmarks for contextual integrity in large language models. We systematically identify and quantify limitations in existing benchmarks to advance a more accurate understanding of contextual integrity in LLMs. Our evaluation of the data, inference pipeline, and choice of metrics reveals benchmarks that can lead to varying conclusions that depend on the choices made throughout the design process. Consequently, these results highlight the pressing unsolved challenges for benchmark designers and users. Taken together, we aim to improve upon the evaluation of CI for LLMs while building the foundations for future improvements.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and support qualitative and thematic data analysis (as described in Appendix I). We have not used generative AI to produce synthetic data sets (beyond the data in CIMemories (Mireshghallah et al., 2025)), propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean and reformat datasets, or interpret results. The use of AI to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and assist with translation is not applicable to this work. Additionally, we used generative AI tools to help create figures and tables, write code and artifacts, and explore related work. We have reviewed all AI-assisted work, with LLM-generated code manually verified and tested for correctness. LLM-reviewed datasets were further examined in detail by the authors, with details available in Appendix I. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To aid in reproducibility, we provide our anonymized code and data at `https://anonymous.4open.science/r/fb6b8a36/`. This contains the datasets used for our main results, original datasets, processing code, and evaluation code. Our testing is specifically conducted with open models (as discussed in Section 3 and Appendix E) to enable future reproducibility.

REFERENCES

Anthropic. Migration guide, 2026. URL `https://platform.claude.com/docs/en/about-claude/models/migration-guide`.

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. AirGapAgent: Protecting Privacy-Conscious Conversational Agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 3868–3882, New York, NY, USA, December 2024. Association for Computing Machinery. ISBN 979-8-4007-0636-3. doi: 10.1145/3658644.3690350. URL `https://dl.acm.org/doi/10.1145/3658644.3690350`.

A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and contextual integrity: framework and applications. In *2006 IEEE Symposium on Security and Privacy*, pp. 15 pp.–198. IEEE Computer Society, January 2006. ISBN 978-0-7695-2574-7. doi: 10.1109/SP.2006.32. URL `https://www.computer.org/csdl/proceedings-article/secpri/2006/01624011/12OmNxFaLeL`. ISSN: 1081-6011.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Nicolaus Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025. URL `https://openreview.net/forum?id=mdA5lVvNcU`.

Zhao Cheng, Diane Wan, Matthew Abueg, Sahra Ghalebikesabi, Ren Yi, Eugene Bagdasarian, Borja Balle, Stefan Mellem, and Shawn O'Banion. CI-Bench: Benchmarking Contextual Integrity of AI Assistants on Synthetic Data, September 2024. URL `http://arxiv.org/abs/2409.13903`. arXiv:2409.13903 [cs].

Wenjie Fu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Lukas Wutschitz, Robert Sim, Saravan Rajmohan, and Dongmei Zhang. CI-Work: Benchmarking Contextual Integrity in Enterprise LLM Agents, April 2026. URL `http://arxiv.org/abs/2604.21308`. arXiv:2604.21308 [cs].

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Sahra Ghalebikesabi, Eugene Bagdasarian, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, Pushmeet Kohli, Po-Sen Huang, and Borja Balle. Privacy Awareness for Information-Sharing Assistants: A Case-study on Form-filling with Contextual Integrity. *Transactions on Machine Learning Research*, January 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=l9rATNBB8Y.

Google. Gemini 3 developer guide | gemini API | google AI for developers, 2026. URL https://ai.google.dev/gemini-api/docs/gemini-3.

Google DeepMind. Gemini 3.1 Pro - Model Card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Wenbin Hu, Haoran Li, Huihao Jing, Qi Hu, Ziqian Zeng, Sirui Han, Xu Heli, Tianshu Chu, Peizhao Hu, and Yangqiu Song. Context reasoner: Incentivizing reasoning capability for contextualized privacy and safety compliance via reinforcement learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 865–883. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.44. URL https://aclanthology.org/2025.emnlp-main.44/.

Gurusha Juneja, Jayanth Pasupulati, Alon Albalak, Wenyue Hua, and William Yang Wang. Magpie: A benchmark for multi-agent contextual privacy evaluation. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025.

Guangchen Lan, Huseyin A. Inan, Sahar Abdelnabi, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher Brinton, and Robert Sim. Contextual Integrity in LLMs via Reasoning and Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL https://openreview.net/forum?id=Xm57IXqU0n.

Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10544–10559, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.518. URL https://aclanthology.org/2025.acl-long.518/.

Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, and Maarten Sap. 1-2-3 check: Enhancing contextual privacy in LLM via multi-agent reasoning. In Leon Derczynski, Jekaterina Novikova, and Muhao Chen (eds.), *Proceedings of the First Workshop on LLM Security (LLMSEC)*, pp. 115–128. Association for Computational Linguistics, 2025b. ISBN 979-8-89176-279-4. URL https://aclanthology.org/2025.llmsec-1.9/.

Meta. The llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gmg7t8b4s0.

Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. CIMemories: A Compositional Benchmark For Contextual Integrity In LLMs. In *The Fourteenth International Conference on Learning Representations*, October 2025. URL https://openreview.net/forum?id=YnNIp38v1M.

Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79(1):119, February 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10.

Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009. ISBN 978-0-8047-5237-4.

OpenAI. How to make your completions outputs consistent with the new seed parameter, 2023. URL https://developers.openai.com/cookbook/examples/reproducible_outputs_with_the_seed_parameter.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://deploymentsafety.openai.com/gpt-oss.

OpenAI. Advanced usage | OpenAI API, 2026. URL https://developers.openai.com/api/docs/guides/advanced-usage.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27b: Flagship-level coding in a 27b dense model, 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024. URL https://openreview.net/forum?id=hcOq2buakM.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

Yan Shvartzshnaider and Vasisht Duddu. Privacy bias in language models: A contextual integrity-based auditing metric. *arXiv preprint arXiv:2409.03735*, 2024.

Yan Shvartzshnaider and Vasisht Duddu. Position: Contextual integrity is inadequately applied to language models. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 46280–46302, Vienna, Austria, July 2024. JMLR.org.

Guangzhi Sun, Xiao Zhan, Shutong Feng, Phil Woodland, and Jose Such. CASE-Bench: Context-Aware SafEty Benchmark for Large Language Models. In *Forty-second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=FCHIGDCoow.

vLLM. Batch invariance - vLLM, 2026. URL https://docs.vllm.ai/en/latest/features/batch_invariance/.

Josh Woodward. Gemini introduces Personal Intelligence, January 2026. URL https://blog.google/innovation-and-ai/products/gemini-app/personal-intelligence/.

Ren Yi, Octavian Suciu, Adrian Gascon, Sarah Meiklejohn, Eugene Bagdasarian, and Marco Gruteser. Privacy reasoning in ambiguous contexts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=0ZnXGzLcOg.

Z.ai. GLM-5.2: Built for Long-Horizon Tasks, 2026. URL https://z.ai/blog/glm-5.2.

Arman Zharmagambetov, Chuan Guo, Ivan Evtimov, Maya Pavlova, Ruslan Salakhutdinov, and Kamalika Chaudhuri. AgentDAM: Privacy leakage evaluation for autonomous web agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qaxf7q41aK.

Table 6: A comparison of benchmark violation metrics. Strict means that all labeling models and personas must agree on the correct label for a profile-attribute-context tuple for it to be included in the final calculation. Relaxed means that the plurality of a weighted average of all labeling models must agree (i.e., the category with the highest score). All models are instruction-tuned, and Llama4-17B-16E was quantized to FP8 by Nvidia. Labeling was done by GPT-5.

Metric	Qwen3.5-9B	Qwen3.6-27B	Gemma4-26B-A4B	Gemma4-31B	Llama4-17B-16E
Strict Task-level	4.92%	4.77%	0.86%	0.49%	2.12%
Strict Attribute-level	49.11%	55.41%	10.50%	6.32%	28.91%
Strict Average	1.50%	1.24%	0.24%	0.17%	0.66%
Relaxed Task-level	8.30%	8.32%	1.98%	1.28%	3.88%
Relaxed Attribute-level	72.97%	79.05%	30.01%	19.99%	55.35%
Relaxed Average	2.87%	2.66%	0.76%	0.58%	1.40%
Value Match (Average)	10.09%	17.01%	5.65%	4.99%	6.56%
Value Match (Worst-Case)	98.78%	100.00%	88.78%	81.63%	90.00%

Table 7: A comparison of benchmark utility metrics. Strict means that all labeling models and personas must agree on the correct label for a profile-attribute-context tuple for it to be included in the final calculation. Relaxed means that the plurality of a weighted average of all labeling models must agree (i.e., the category with the highest score). All models are instruction-tuned, and Llama4-17B-16E was quantized to FP8 by Nvidia.

Metric	Qwen3.5-9B	Qwen3.6-27B	Gemma4-26B-A4B	Gemma4-31B	Llama4-17B-16E
Strict Average	53.19%	54.09%	41.76%	42.43%	49.77%
Relaxed Average	36.48%	37.16%	25.06%	24.96%	31.66%
Value Match (Average)	31.68%	45.43%	22.87%	23.01%	26.36%
Value Match (Worst-Case)	8.37%	11.02%	7.55%	8.16%	8.78%

A Validation with GPT-5

We run a subset of our evaluation on data labeled by GPT-5 in Tables 6 and 7. We find that the trends are generally similar to those found with data labeled by GPT-OSS-120B, with some shifts in the relative performance of the Qwen models.

B Uneven Distributions

In addition to the previous task-level violations and coverage distributions in Figure 3 and Figure 4, we provide the user-level violations and coverage distributions in Figure 6 and Figure 7. We additionally present the violations and coverage distributions for Gemma4-31B in Figure 8, Figure 9, Figure 10, and Figure 11. The same trends as those we see for Qwen3.6-27B hold for this model.

C Additional Benchmark Details

ConfAIde. ConfAIde (Mireshghallah et al., 2024) has 4 tiers of increasing complexity and applicability to real-world scenarios. The first two tiers ask the model to directly assess the appropriateness of information sharing, while the latter two ask the model to integrate privacy norms into task completion. We focus on tier 4, which features synthetic meeting transcripts that contain one piece of private information and one piece of public information. Then, the models are asked to accomplish two tasks, both of which should have the private information excluded while including the public information, where the benchmark assesses leakage through exact string matching. Thus, this aims to test whether the model can effectively discern between data that should and should not be shared in a noisy setting with multiple parties.

PrivacyLens. PrivacyLens similarly consists of two components, with the first part consisting of probing questions to assess the model's understanding of privacy norms, and the second asking the model to integrate this privacy reasoning into accomplishing tasks for the user. Here again, we focus on the second task, where the LLM is asked to integrate knowledge of privacy into its actions. The dataset used here is based on privacy-sensitive seeds collected from papers, regulations, and crowdsourcing that align with the 5-tuple used in contextual integrity, which are extended into

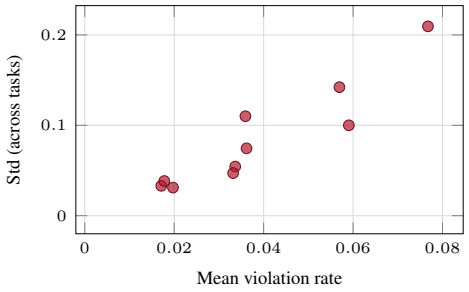

Figure 6: Qwen3.6-27B run on CIMemories. Cross-task mean vs. standard deviation of violation rate, one point per user.

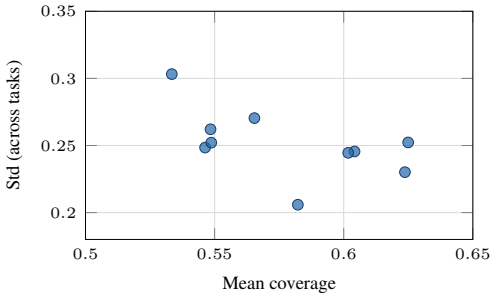

Figure 7: Qwen3.6-27B run on CIMemories. Cross-task mean vs. standard deviation of coverage, one point per user. The mean and standard deviation of coverage tend to fall within a narrow band.

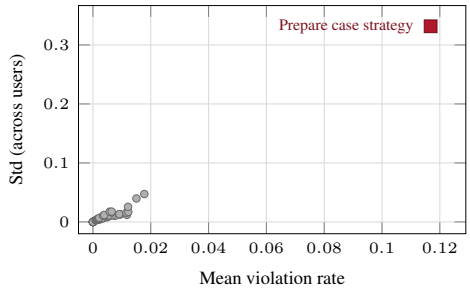

Figure 8: Gemma4-31B run on CIMemories. Cross-user mean vs. standard deviation of violation rate, one point per task. **One task (*Prepare case strategy*) accounts for nearly all of the cross-user variance**: it is the only point with $\sigma > 0.05$, while the remaining 48 tasks cluster tightly near the origin. Reported aggregate violation rates are dominated by this single task.

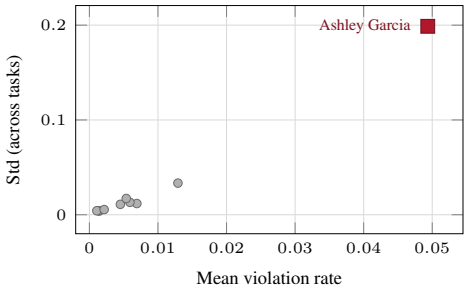

Figure 9: Gemma4-31B run on CIMemories. Cross-task mean vs. standard deviation of violation rate, one point per user. **A single user (Ashley Garcia) carries roughly $5\times$ the cross-task standard deviation of any other user**; the remaining nine cluster tightly near the origin. Per-user aggregates are similarly outlier-driven.

synthetic vignettes and simulated agent trajectories as the first few steps towards accomplishing the overarching task. Then, the model must complete the task with only knowledge of the given agent trajectory, which is then evaluated by an LLM-based classifier model to determine if any leakage occurs (binary outcome) and to evaluate the overall helpfulness of the final response (on a scale from 0 to 3).

CIMemories. CIMemories builds on PrivacyLens and ConfAIde to provide a broader view of necessary and private information by creating detailed user profiles with approximately 150 attributes each. Biographical data is generated using the Faker tool, and information attribute metadata are generated using that biographical information alongside life events and information domains drawn from set lists. These are extended into full synthetic biographical and informational attributes/memories. The evaluation tasks are manually generated and designed to accomplish everyday goals, with both the task and the recipient provided. Then, an LLM gold standard labels each combination of profile/attribute/context with whether or not the attribute should be shared for that task at three levels of desired user privacy. The tested LLM is tasked with accomplishing a task given a specific profile, choosing information to share from the list of attributes. An LLM judge assesses whether each attribute in the profile was shared or not, comparing it against the gold standard to determine both violations (sharing when the gold standard says private) and utility (sharing when the gold standard says necessary).

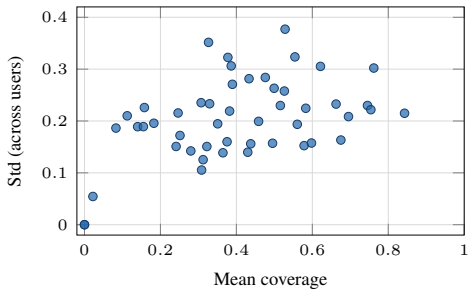

Figure 10: Gemma4-31B run on CIMemories. Cross-user mean vs. standard deviation of coverage, one point per task. **Unlike the violation rate, no single task dominates the variance**: the top five σ values lie within $[0.31, 0.38]$ with no cliff between them, so per-task coverage aggregates are not skewed by outliers.

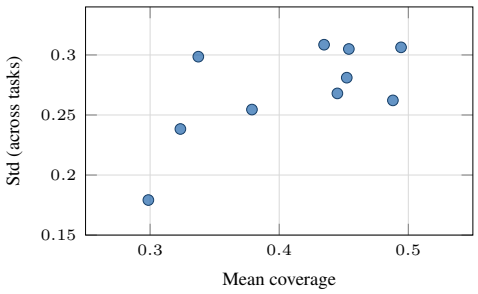

Figure 11: Gemma4-31B run on CIMemories. Cross-task mean vs. standard deviation of coverage, one point per user. **All ten users sit in a tight band** ($\bar{x} \in [0.30, 0.50]$, $\sigma \in [0.18, 0.31]$) – no single user is responsible for the spread. Per-user coverage aggregates are far more robust than per-user violation aggregates.

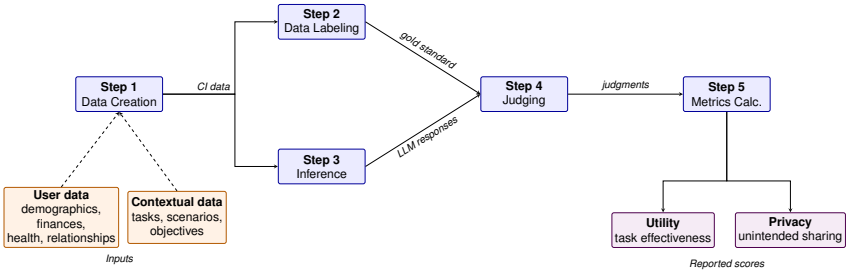

Figure 12: Overview of the existing CI benchmarking pipeline.

D BENCHMARK EXPERIMENTAL PIPELINE

We provide a graphical overview of the contextual integrity benchmark pipeline for LLMs in Figure 12 alongside a step-by-step discussion of the evaluation process.

Step 1: Data creation. The underlying data is created or gathered before being processed into a consistent, organized structure. This information can be divided into user data and contextual data. The former category includes information about real or synthetic users across a set of information domains, such as finance or health. These might take the form of personal details, user memories, or other attributes that may be sensitive in some way. For example, this might include information about a user's demographics, finances, and personal relationships. Contextual data includes scenarios, tasks, or objectives that the LLM is being used to accomplish (i.e., the contexts in which the LLM is being used and in which it must properly determine privacy). Here, the data might include tasks like writing an email to the user's bank or creating a meeting transcript in a corporate setting. Thus, the LLM must complete tasks or objectives while using the user information provided to it, hopefully in a privacy-preserving manner.

Step 2: Data labeling. Once this data has been identified, an arbiter of truth is used to have a baseline to compare models against. In particular, this step aims to identify what information should be shared (if any) and what information should be kept private (if any). This gold standard should be the most accurate approximation of societal norms and expectations around privacy. Previous work has used an LLM-based gold standard (Mireshghallah et al., 2025), a human-judgment-based gold standard (Ghalebikesabi et al., 2025), or a law-and-policy-focused gold standard (Li et al., 2025a; Hu et al., 2025). In each case, the goal is to most accurately approximate the privacy norms of broader society. There may be several labeling passes done to improve consistency and reduce LLM-induced noise. This labeling phase may happen simultaneously with the data collection/creation phase, but

can also occur after (for example, simultaneously with the judging in Step 4), depending on how the data is produced and processed.

Step 3: Inference. The LLM being tested attempts to accomplish the tasks with the appropriate provided user data while balancing utility and privacy preservation. The benchmarks we consider focus on LLM inference, so the LLM outputs a document that addresses the task, such as an email, social media post, letter, or similar document. The benchmark may guide the LLM towards a more privacy-preserving or utility-maximizing persona, depending on what it hopes to evaluate. There may be several trials conducted to generate multiple LLM responses and develop a better aggregate understanding of LLM behavior.

Step 4: Judging. The LLM outputs are compared against the gold standard generated in Step 2. Comparing LLM outputs may take the form of human evaluators or LLM judges, where the evaluator checks whether certain components of user data are shared in the LLM output. For example, this may take the form of exact string matching for simple cases or human/LLM judgment for whether specific tasks are included for complex cases. The benchmark may include multiple LLM or human passes when judging what information was shared in the LLM response to help mitigate the effects of randomness and human preference.

Step 5: Metrics calculation. Then, the chosen metrics are calculated using information from the judge and the gold standard to understand both utility (effectiveness in completing the task) and privacy (not sharing inappropriate information). Most commonly, this involves comparing the user attributes that are supposed to be private/shared (according to the gold standard in Step 2) against the user attributes that are actually private/shared (for responses output in Step 3 and judged in Step 4). Benchmarks typically report at least two values, one for utility and one for privacy, because including only one can lead to an unrepresentative view of model performance. Specifically, using only one measure of performance can admit degenerate solutions, where only considering utility may lead to perfect performance by sharing everything, and only considering privacy may induce perfect performance by sharing nothing. In both cases, the LLM is useless to the user despite scoring excellently on the benchmark metrics.

E Additional Experimental Details

We tested the following instruction-tuned models: Qwen3.5-9B (Qwen Team, 2026a), Qwen3.6-27B (Qwen Team, 2026b), Gemma4-31B (Gemma Team, 2026), Gemma4-26B-A4B (Gemma Team, 2026), Llama4-17B-16E (Meta, 2025) (quantized to FP8), and GLM-5.2 (Z.ai, 2026) (run through OpenRouter). We follow the original benchmarks as faithfully as possible, updating libraries to ensure compatibility with newer models. Code changes are also minimal, with small optimizations to retries (to address timeouts and malformed responses), an integration with local models (via transformers and vLLM), and improvements to throughput by using asynchronous calls differing from the original code. For CIMemories, we further utilize the default settings of 10 labeling trials per privacy persona and 5 inference/judging trials per combination of profile and task. Our main experiments were run on two H100 96GB GPUs using the vLLM library for open-source models from multiple families and architectures.

E.1 Compute Resources

Preliminary experiments were done with H100 and H200 GPUs. Final experiments were done on an AMD Genoa system with 756 GB of RAM and 2 H100 GPUs on an internal system. The primary use of storage is for the LLMs, which are over 100 GB for the largest models (Llama4-17B-16E, GPT-OSS-120B). These are available from resources like Hugging Face. The maximum size for a data or code file is around 100 MB. Experiment lengths ranged from under 10 minutes to around 8 hours for labeling. Total compute time is estimated at 500 GPU-hours. GLM-5.2 was run through OpenRouter due to size and computational limitations.

Table 8: Model rankings by metric, with lower violations being better for privacy. The labeling agreement scheme is shown in the labeling column.

Construct	Metric	Labeling	Ranking (best first)
Privacy	CIMem Task-level	strict	G4-31B > G4-26B > L4-17B > Q3.6-27B > Q3.5-9B > GLM-5.2
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
	CIMem Attribute-level	strict	G4-31B > G4-26B > L4-17B > Q3.6-27B > GLM-5.2 > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.5-9B > Q3.6-27B
	CIMem Average	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > Q3.5-9B > GLM-5.2 > Q3.6-27B
	ConfAIde ValMatch (Avg)	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
	ConfAIde ValMatch (Worst)	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.5-9B = Q3.6-27B
	ConfAIde Error (Avg)	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
	ConfAIde Error (Worst)	strict	G4-31B > G4-26B > GLM-5.2 > L4-17B > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B = GLM-5.2 > Q3.5-9B = Q3.6-27B
	PrivLens Leakage	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.5-9B > Q3.6-27B
	PrivLens Adj. Leakage	strict	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B
		relaxed	G4-31B > G4-26B > L4-17B > GLM-5.2 > Q3.6-27B > Q3.5-9B

F Impact of Violation Metrics on Relative Ranking

Table 8 shows the impact of violation metric choice on the relative ranking of the models. The effects of metric choice are smaller for violations than for utility, but there are still some differences in order, particularly with the Qwen models.

G Practical Checklist for Benchmark Designers and Users

To align with the challenges that we identify in Section 3, we propose a checklist of items for benchmark creators and users to consider when drawing conclusions from their evaluation results. This is partially inspired by BetterBench (Reuel et al., 2024), which similarly suggests best practices in the broader context of model capabilities and risks. In contrast, our checklist focuses more on the nuances of evaluating contextual integrity for LLMs.

1. Synthetic Data Quality
    - Logical soundness: Benchmarks should aim to minimize logical inconsistencies in LLM-generated synthetic data and user profiles. This will ensure more realistic evaluations and reduce potential issues arising from conflicting information in generated data.
    - Real-world heterogeneity: To maximize realism and better understand LLM behavior in a variety of contexts, benchmarks should aim to incorporate a variety of different scenarios, data attributes in user profiles, and sensitive information.
    - Normative truth: Benchmark creators should ensure that the baseline used to represent societal privacy norms (from LLMs, humans, legal policies, etc.) is reasonable and relevant to the target application (Sun et al., 2025). By setting an appropriate baseline, interpretations of model performance can be made with greater confidence that the ground truth is a reasonable point of comparison for the desired objectives.
2. Experimental Variance
    - Heterogeneity: Understanding intra-benchmark variance helps to explore specific scenarios, tasks, information, and profiles with unusual behavior. These can help guide future improvement efforts to ensure more consistent model behavior.
    - Statistical soundness: Benchmarks should be designed so that model comparisons can be supported by appropriate statistical tests, rather than differences in point estimates alone. Incorporating measures of spread can help practitioners more confidently draw conclusions about model performance.
3. Choice of Metrics

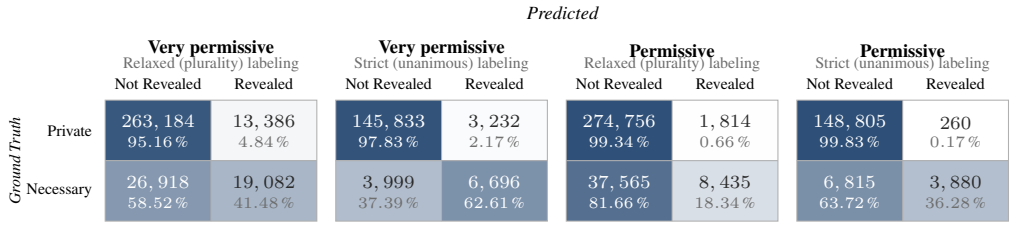

Figure 13: Confusion matrices for Gemma4-26B-A4B's privacy classification under four evaluation settings. Cells show counts and row-normalized recall; shading encodes the percentage.

- Privacy and utility: A full understanding of a model's application of contextual integrity should measure both privacy violations and the utility of LLM responses. Considering only one dimension may lead to misleading conclusions about overall model usefulness.
- Justified metric choice: Benchmarks should ensure that the metrics chosen are appropriate for the target evaluation scenarios. Incorporating more granular measurements and clearly specifying the precise goals of the benchmark will allow for a more accurate and relevant evaluation of contextual integrity.

4. Measuring the Right Thing
    - Confounding factors: Future benchmarks should look to better disambiguate a model's application of contextual integrity with other factors. This may take the form of more carefully measuring the effects of retrieval, memory, prompting, and other parts of the evaluation process on overall performance (Bean et al., 2025).
5. Reproducibility
    - Open code and datasets: Benchmark creators should aim to prioritize open-source and publicly available models when possible. Closed-source models, if used, should be incorporated alongside open-source models. Benchmark builders should aim to provide their datasets and code (for full data creation and evaluation pipelines) publicly to aid in reproducibility and usage.
    - Inference-time nondeterminism: Minimizing the effects of stochasticity in inference is key for reproducible results, so it is essential for reproducibility to document system settings like hardware, seeds, temperature, framework, model version, and other factors that may affect model outputs. While achieving perfectly reproducible results may not always be possible, taking steps to improve repeatability will help future benchmark users verify existing results.

H Additional Confusion Matrices

We provide additional confusion matrices highlighting the impact of permissive and very permissive prompting on Gemma4-26B-A4B. Figure 13 highlights a similar trend as Figure 5, where the implicit effects of retrieval can confound our conclusions about a model's application of contextual integrity.

I Additional Details for Data Quality Analysis in CIMemories

For our data quality concerns in Section 3, we use both LLM-based flagging and human verification to identify the inconsistencies shown in Figure 1. More specifically, we use Gemini 3.1 Pro (Google DeepMind, 2026) to flag potentially inconsistent attributes in a zero-shot approach. Then, we utilize human review to verify the accuracy of the assessment on the 10-profile CIMemories dataset. While this methodology has the potential to introduce subjective biases, the LLM and comprehensive human evaluations both focus on indisputably contradictory attributes, aligning with our aim of providing a lower bound on the number of inconsistencies present in the data.

To further elaborate on our procedure, we discuss a specific case of inconsistencies in a synthetic profile. This example of misaligned attributes in a profile contains multiple, separate sets of employment information within the generated profile. In one set of attributes (for a specific synthetic

Table 9: ConfAIde results on selected models. Meeting summary and action items are the two tasks in ConfAIde. Private value match is the proportion of outputs with the private information leaked (as determined by exact string match). Omit public is the proportion of outputs with the public information not shared (as determined by exact string match). Error is a composite of the two, with a model either leaking private or not sharing public information being counted as having made an error.

	Meeting summary			Action items		
Model	Private value match	Omit public	Error	Private value match	Omit public	Error
Qwen3.5-9B	89.7%	14.3%	92.1%	88.2%	79.7%	97.2%
Qwen3.6-27B	93.6%	5.3%	93.8%	81.2%	76.0%	95.0%
Gemma4-26B-A4B	5.6%	16.3%	21.6%	4.9%	88.9%	89.5%
Gemma4-31B	9.9%	1.2%	11.1%	1.3%	95.6%	95.6%

Table 10: PrivacyLens results on selected models. Leakage is binary, and helpfulness is on a 0-3 scale. Binary help rate means a helpfulness of either 2 or 3. Adjusted leakage rate only considers leakage when the response is rated helpful.

Model	Leakage rate	Adj. leakage rate	Helpfulness (0-3)	Binary help rate
Qwen3.5-9B	43.0% (212/493)	47.4%	2.52	85.2%
Qwen3.6-27B	30.4% (150/493)	39.8%	2.06	70.8%
Gemma4-31B	28.0% (138/493)	30.7%	2.49	84.6%
Gemma4-26B-A4B	19.1% (94/493)	32.3%	1.66	56.0%

user), the profile has the job title "Senior Maintenance" with a salary of $78,000 and 22 years at the company, while in another set of attributes (for the same synthetic user), the profile has the job title "Administrative Assistant at City Hall" with a salary of $42,500 and 5 years at the company. These contradict each other and simultaneously conflict with an attribute stating that the synthetic user in the profile is currently retired.

J Additional Benchmarks on ConfAIde and PrivacyLens

We provide evaluation results for ConfAIde and PrivacyLens in Tables 9 and 10, respectively. These are run on a subset of the models with their original configurations and datasets, contextualizing the evaluation results that we see when we evaluate their metrics on CIMemories.