Probabilistic Reasoning with LLMs for Privacy Risk Estimation

Jonathan Zheng[1] **Sauvik Das**[2] **lan Ritter**[1] **Wei Xu**[1]
[1]School of Interactive Computing, Georgia Institute of Technology
[2]Human-Computer Interaction Institute, Carnegie Mellon University
jzheng324@gatech.edu

bstract

Probabilistic reasoning is a key aspect of both human and artificial intelligence that allows for handling uncertainty and ambiguity in decision-making. In this paper, we introduce a new numerical reasoning task under uncertainty for large language models, focusing on estimating the privacy risk of user-generated documents containing privacy-sensitive information. We propose BRANCH, a new LLM methodology that estimates the k-privacy value of a text—the size of the population matching the given information. BRANCH factorizes a joint probability distribution of personal information as random variables. The probability of each factor in a population is estimated separately using a Bayesian network and combined to compute the final k-value. Our experiments show that this method successfully estimates the k-value 73% of the time, a 13% increase compared to o3-mini with chain-of-thought reasoning. We also find that LLM uncertainty is a good indicator for accuracy, as high variance predictions are 37.47% less accurate on average.

1 Introduction

Large language models (LLMs) have shown increasingly strong performance in mathematical and logical reasoning [84, 61, 16], enabling exploration of a broad suite of user-facing applications. One such application is helping users understand the magnitude of privacy risks when disclosing personal information online — a holy grail of usable privacy that remains elusive [59, 38]. For example, how much riskier is it to disclose one's precise age versus a general age category (22 vs. 18-25)? To that end, we introduce a task that instructs LLMs to estimate the k number of people in the world that match the personal attributes or experiences presented in user-written messages on pseudonymous online fora like Reddit [18] or anonymous user interactions with ChatGPT [88, 50]. This task serves both as a new challenge for LLMs' reasoning abilities beyond conventional math and logic tests [28, 45, 34] and as a practical security tool to inform users about online safety.

Traditionally, privacy research has focused on quantifiable properties such as k-anonymity [76, 67] applied by *dataset owners* to protect individual records, with privacy risk assessed by the success rate of re-identifying anonymized entries in a database (e.g., a person on Wikipedia) [52, 53, 85]. In this work, we shift the focus to *end-users*, who are neither dataset owners nor curators, by providing an interpretable risk estimate that can be computed *without* access to a comprehensive database containing personal records of all internet users. LLMs must rely on their internal knowledge of demographic statistics from census data to reason about and estimate the identification risk based on the multiple, potentially interconnected, first-person textual disclosures. For instance, LLMs may estimate the privacy risk of a post where a user mentions that she is from Italy, 26 years old, on the spectrum, and has social anxiety (see Figure 1 and 2 for more examples).

Reasoning on these joint attributes remains a significant challenge for language models. In Figure 1, Chain-of-Thought prompting estimates the prevalence rates of social anxiety and autism indepen-

Preprint. Under review.

Chain of Thoughts: **Instruction:** How many people share all the following attributes as this user?

Conditional Dependency Errors (25.64%)

The poster is from **Italy**, **female**, **26 years old**, **on the spectrum**, and has **social anxiety**.

1. Gender: The population of Italy is 59,000,000. ~50% are female. 29,500,000 × 0.5 = 29,500,000 ✓
2. Age: In a country with around 80 "age-buckets," 29,500,000 ÷ 80 = 368,750 are born in a single year. ✓
3. Autism: General prevalence for women is around 0.25%. 368,750 × 0.0025 ≈ 921.88 ✓
4. Social anxiety: Estimates for people with social anxiety are around 10%. 921.88 × 0.1 ≈ 92.19 ✗
(The prevalence of social anxiety in people on the Autism spectrum is roughly 50%)

After applying all these "filters" sequentially we get an estimate of roughly **93** people share these traits. ✗ (Correct: 461)

Personally Identifiable Information Errors (15.38%)

The poster is an **entrepreneur** of a **software development services company** called **Townbridge InfoTech**.

1. Occupation: There are 500 million native English speakers. 5% (25 million) of people are entrepreneurs. ✓
2. Business: Roughly 10% entrepreneurs are launching IT companies. 25,000,000 × 0.1 = 2,500,000. ✓
3. PII: We estimate that about 1 in 10,000 of entrepreneurs would choose the name "Townbridge InfoTech". ✗
2,500,000 × 0.0001 = 250 (Business names as specific as "Townsbridge InfoTech" are unique, i.e. 1 in 2,500,000)

Based on these self-disclosures, approximately **250** people would share all of these specific details. ✗ (Correct: 1)

Figure 1: Most common Chain-of-Thought reasoning error types (with occurrence rates) and examples for o3-mini on Privacy Risk Estimation. Errors and correct explanations are highlighted. Chain-of-Thought struggles to model PII and capture relationships between attributes for risk assessments.

dently, failing to account for the conditional dependency whereby being on the spectrum significantly increases the likelihood of social anxiety. To improve privacy risk estimation, we propose BRANCH (see Figure 2), a probabilistic reasoning framework for LLMs that represents each document as a joint distribution of personal attributes. BRANCH first factorizes this distribution by implicitly constructing a Bayesian network, capturing the interdependencies between all of the attributes. Each attribute is then transformed into a textual query and individually estimated using standalone LLMs. Finally, BRANCH reconstructs the joint probability following the structure of the Bayesian network to predict an integer k, representing the number of people worldwide who share the relevant personal attributes.

We empirically evaluate BRANCH and state-of-the-art baselines in privacy risk estimation using a new dataset of user posts with gold standard values. BRANCH accurately predicts k 72.61% of the time, outperforming baselines by 23.04%. On documents with four or more personal attributes, BRANCH is significantly better than Chain-of-Thought prompting due to its probabilistic model that estimates each attribute individually. We find that BRANCH excels at estimating single-attribute demographics—with low percentage errors compared to ground truth census records—so recombining these probabilities leads to superior privacy risk assessments. We further evaluate LLM uncertainty as an indicator for estimation accuracy and find that predictions with high variance result in 37.47% lower accuracy.

In summary, privacy risk estimation serves both as a method for evaluating the probabilistic reasoning capabilities of large language models and as an application to support user-centered privacy. Inspired by prior research in the field, we develop a general, human-interpretable value k that helps users understand privacy risks in text, even in the absence of a database of all online users. Our work is also motivated by human-computer interaction (HCI) user studies [38, 18] where participants expressed a desire for explanations on the severity of risks associated with personal disclosures. We envision BRANCH as a practical tool that can provide a number k with a reasoning chain to inform users about the potential identification risks, much like a password strength meter [36], for online privacy.

2 Related Work

Causal and Probabilistic Reasoning in Large Language Models. Prior research has developed benchmarks to assess LLMs on mathematical and logical reasoning [49, 45, 28, 34], causal reasoning [32, 82], general probabilistic reasoning in synthetic scenarios [54, 62, 58], and Bayesian reasoning for everyday scenarios [20]. Previous research has also evaluated LLM capabilities in constructing Bayesian Networks [3] and medical domain Causal Graphs [78]. Nevertheless, this field of Probabilistic Reasoning Large Language Models remains understudied, especially for real-world applications like modeling privacy risks. A key component of probabilistic reasoning is uncertainty quantification, which has traditionally been explored in Bayesian neural networks [17] and in deep learning models with dropout [23]. More recently, research has estimated uncertainty in autoregressive models and large language model generations with the negative log likelihood of sequences [48, 1]. In this work, we use the consistency of multiple LLM generations [80] as an estimate of uncertainty.