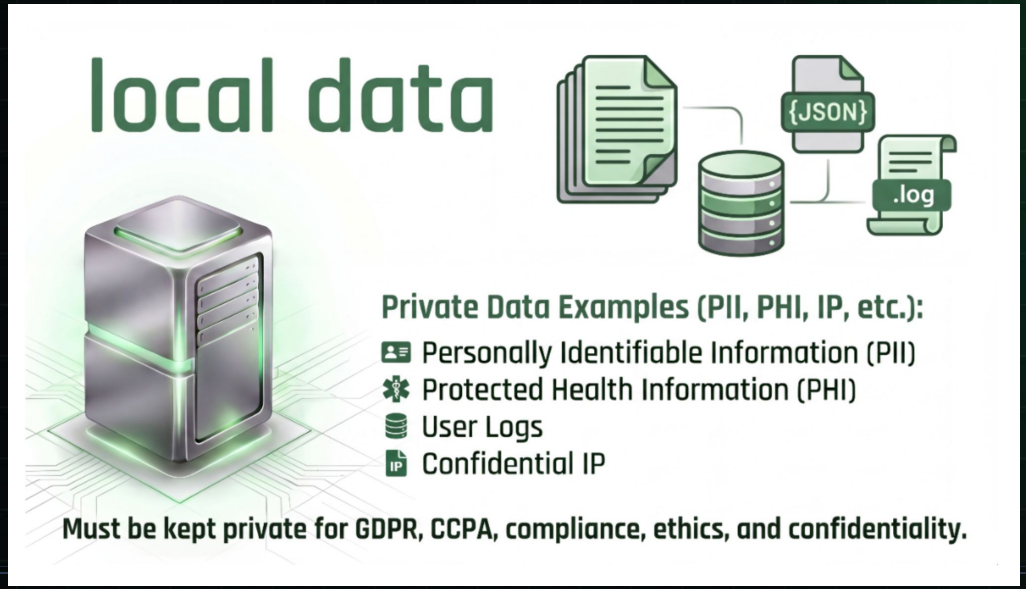
Mohamed bin Zayed University of Artificial Intelligence | Massachusetts Institute of Technology | ICLR International Conference On Learning Representations

DP-FUSION

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, Nils Lukas
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Massachusetts Institute of Technology (MIT),
first_name.last_name@mbzuai.ac.ae

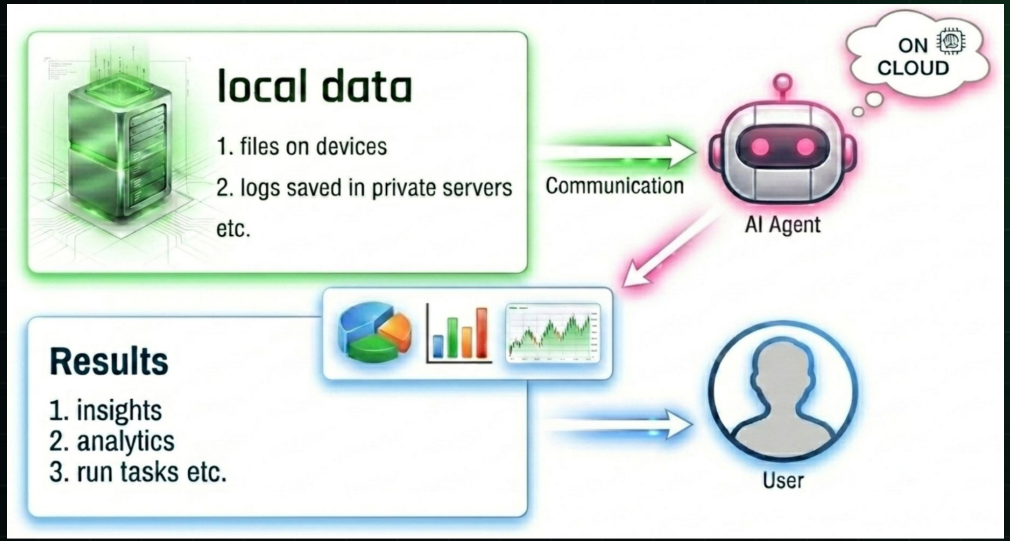
Local Data

Private data stays on-prem traditionally.



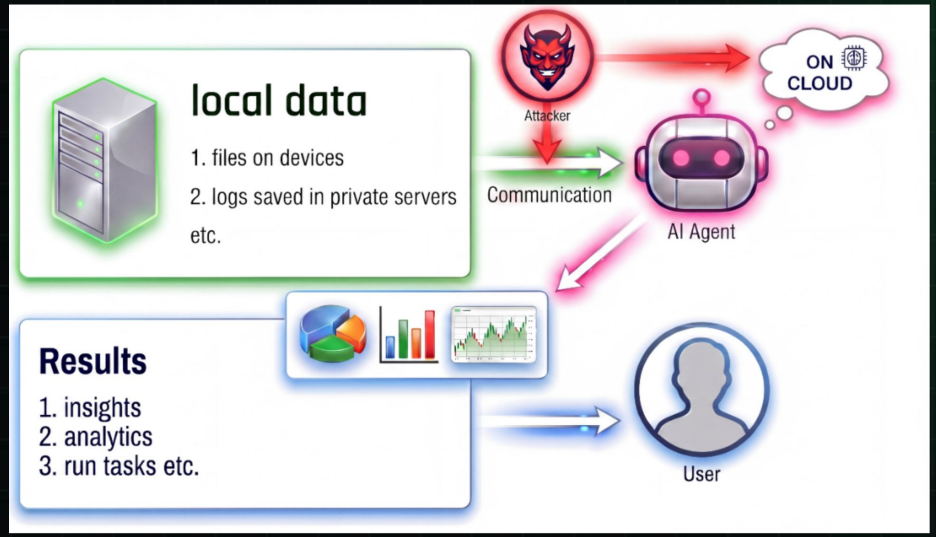
AI Agent Workflows

Local data sent to AI Agent on Cloud.



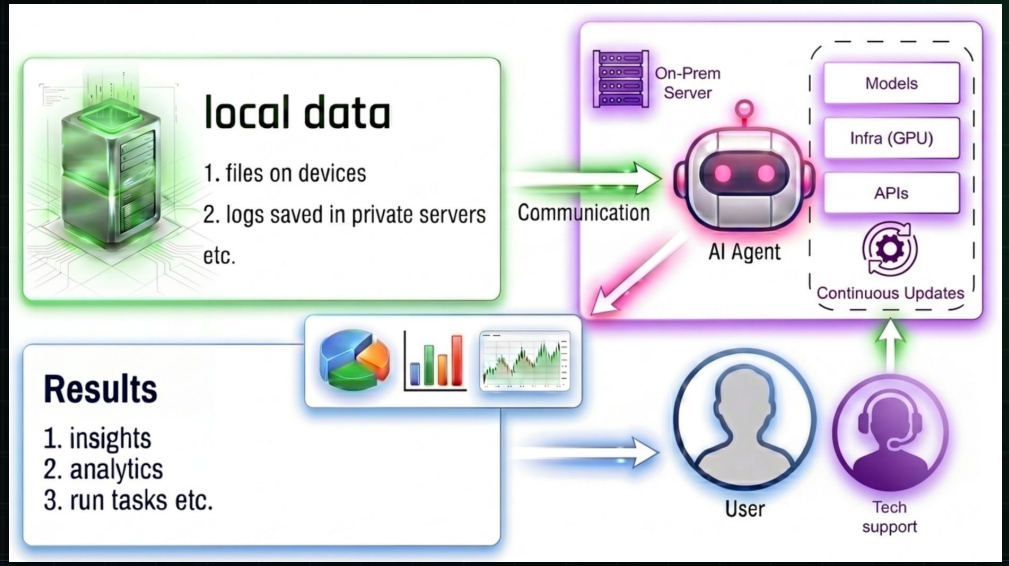
AI Agent Workflows

This creates security & compliance risks, an attacker might steal data from the cloud or communication stage.



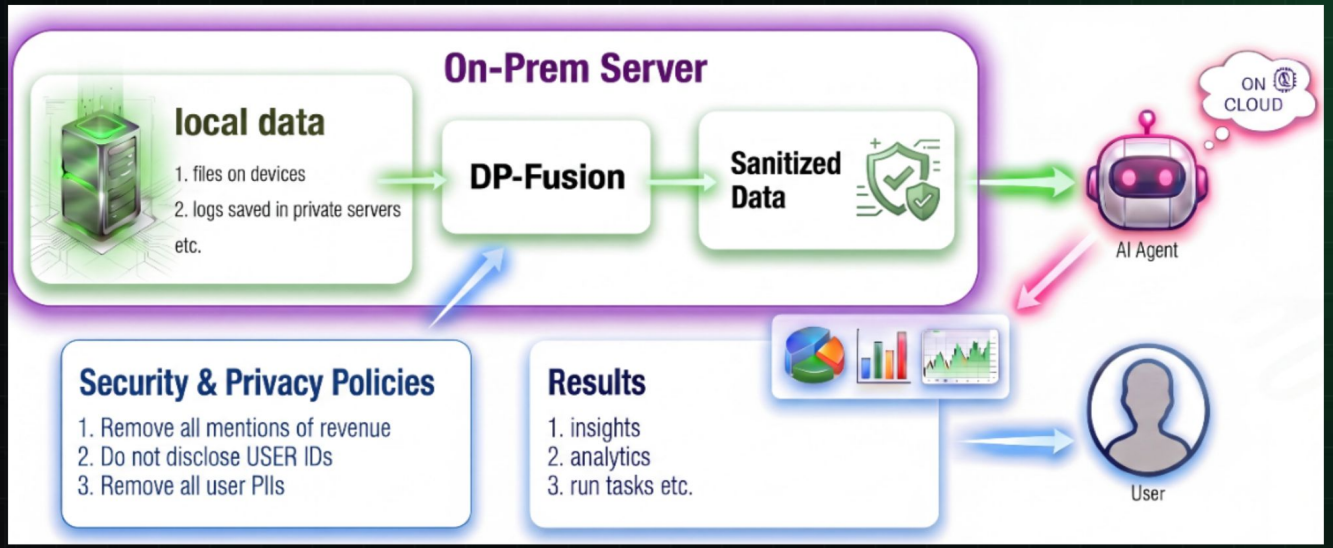
Local AI as a solution

Local AI is a viable solution, but it is costly to maintain and often comes with a performance trade-off compared to frontier closed-source models



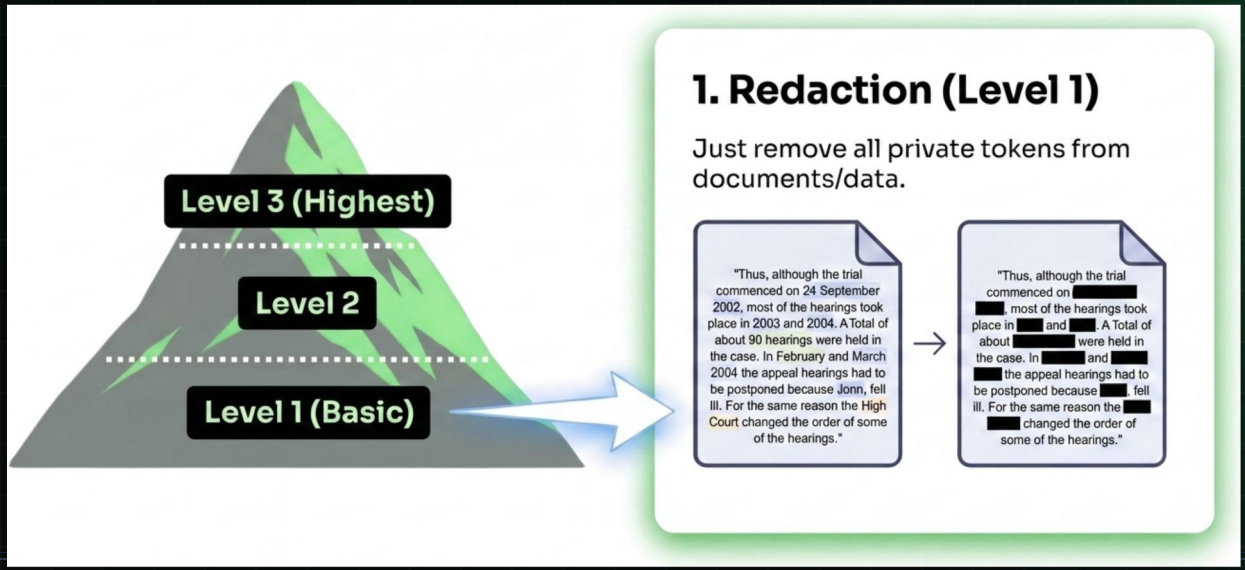
DP-Fusion

DP-Fusion generates a synthetic version of local data with formal differential privacy guarantees



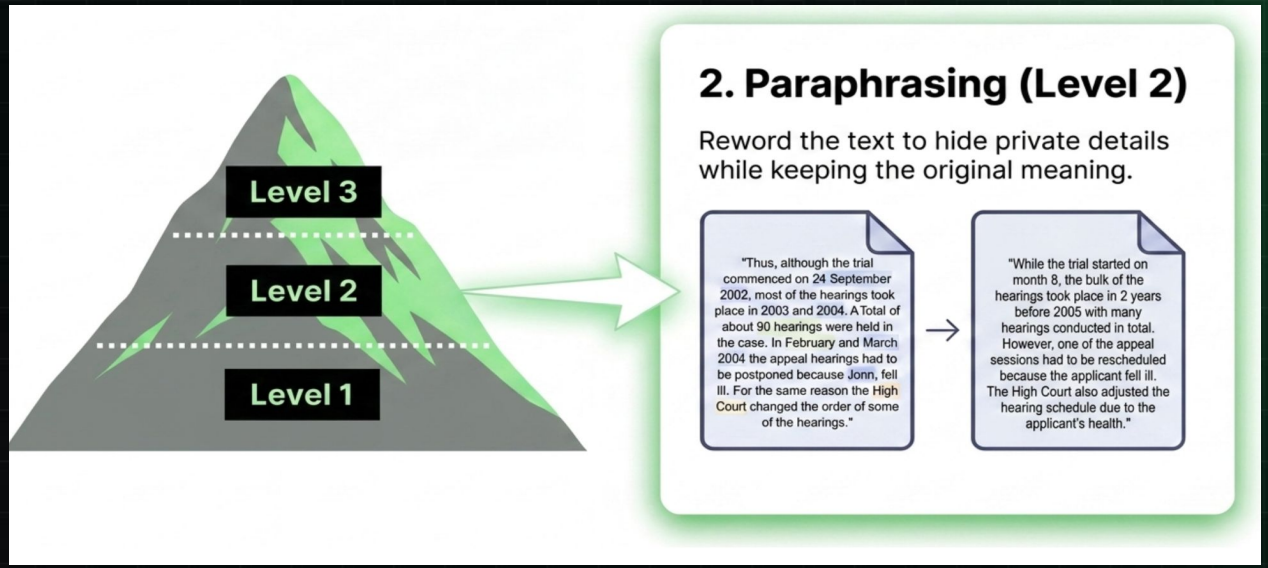
The 3 Levels of Text Privacy

Level - 1 - Redaction



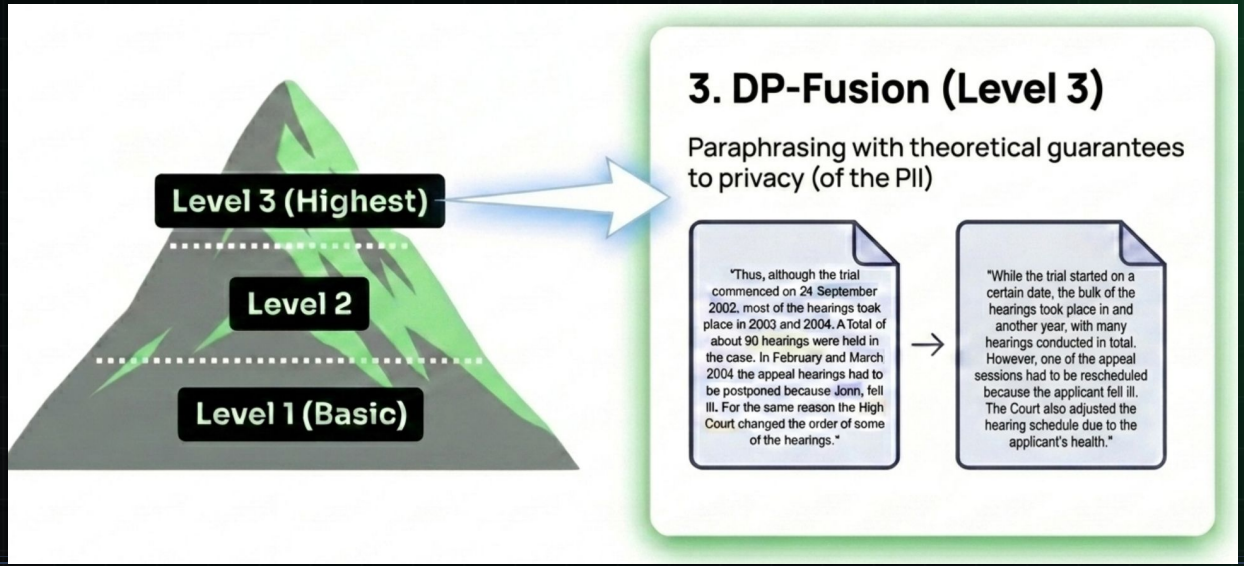
The 3 Levels of Text Privacy

Level – 2 – Redaction



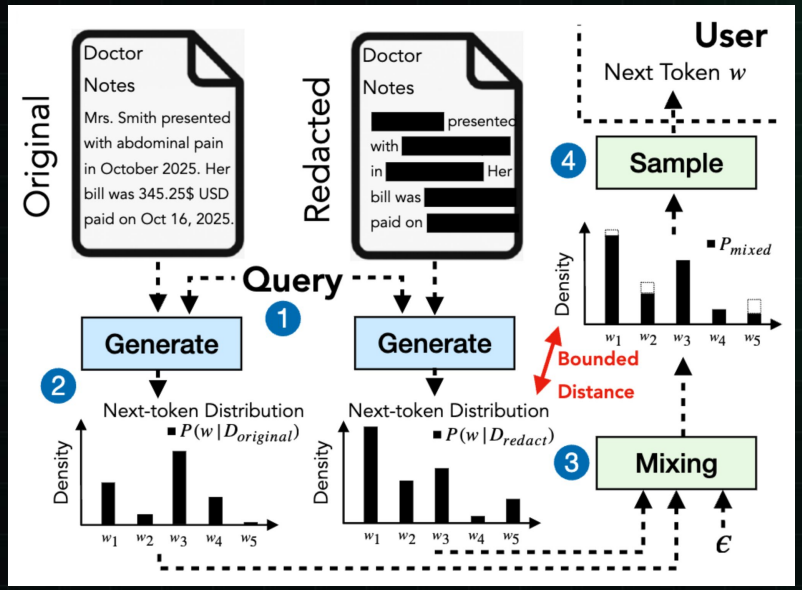
The 3 Levels of Text Privacy

Level - 3 - DP-Fusion



The DP-Fusion DPI Mechanism

DP-FUSION bounds the influence of private tokens on the output of an LLM



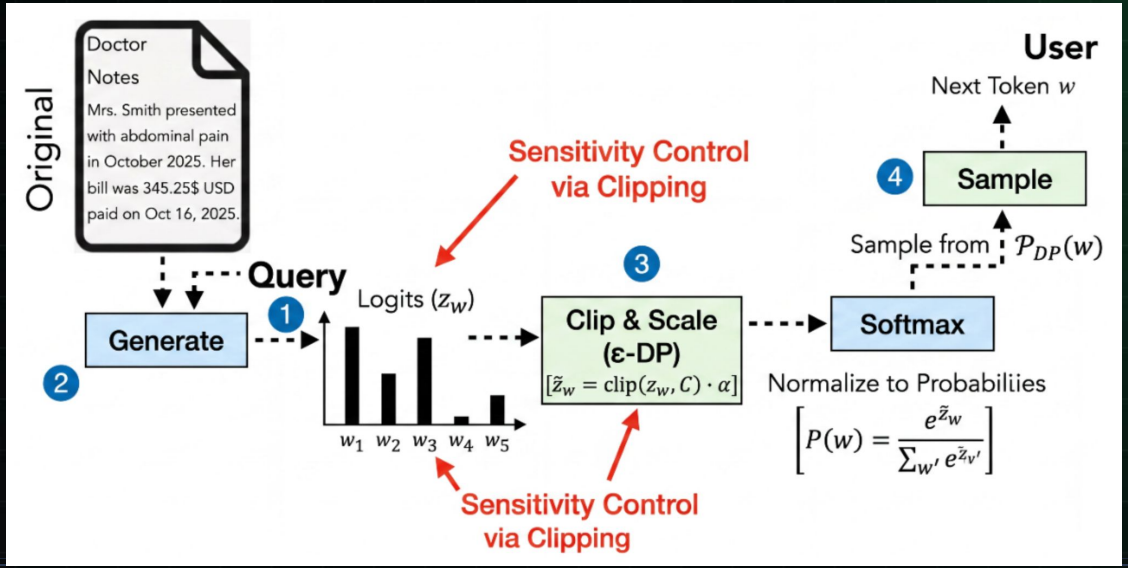
The DP-Fusion DPI Mechanism

Theoretical Guarantees

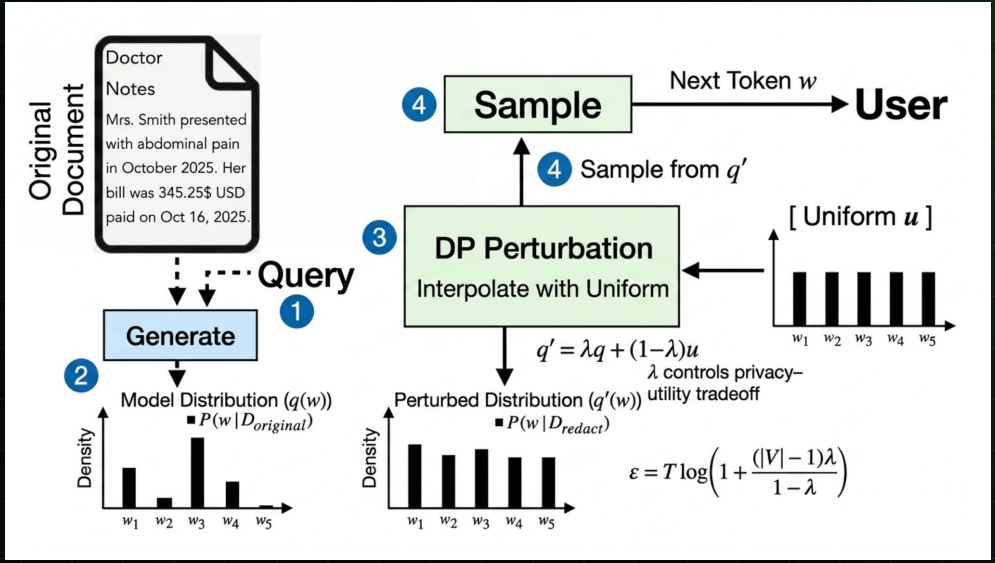
Theorem 4 (Per-group (ε_i, δ)-DP for T tokens). *Assume DP-Fusion M fulfils Definition 4 at order $\alpha > 1$ with budgets $\beta_1, \ldots, \beta_m$. Let $\delta \in (0, 1)$ and generate T output tokens autoregressively with M. Then for every group i the entire T-token transcript is (ε_i, δ)-DP with respect to the add/remove adjacency of Definition 3, where*

$$\varepsilon_i = T \cdot \frac{1}{\alpha - 1} \log\left(\frac{m-1}{m} + \frac{1}{m} e^{(\alpha-1)4\beta_i}\right) + \frac{\log(1/\delta)}{\alpha - 1} \qquad \textit{(Full proof in Flemings et al. (2024)).} \qquad (5)$$

Baselines – DP – Prompt

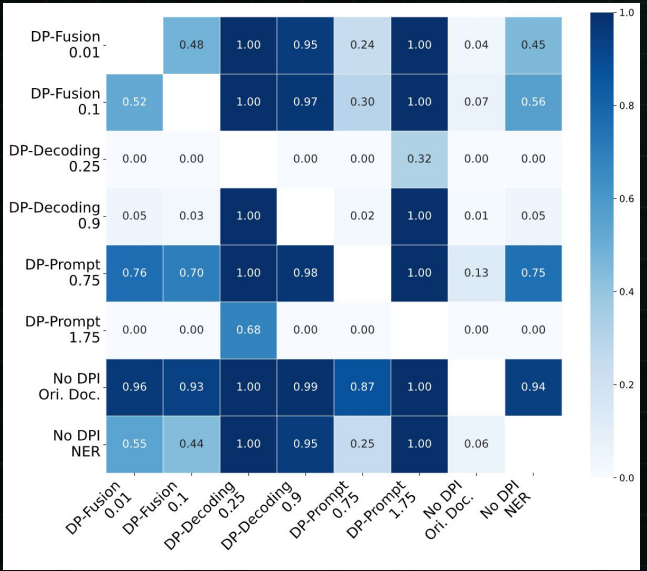


Baselines – DP – Decoding



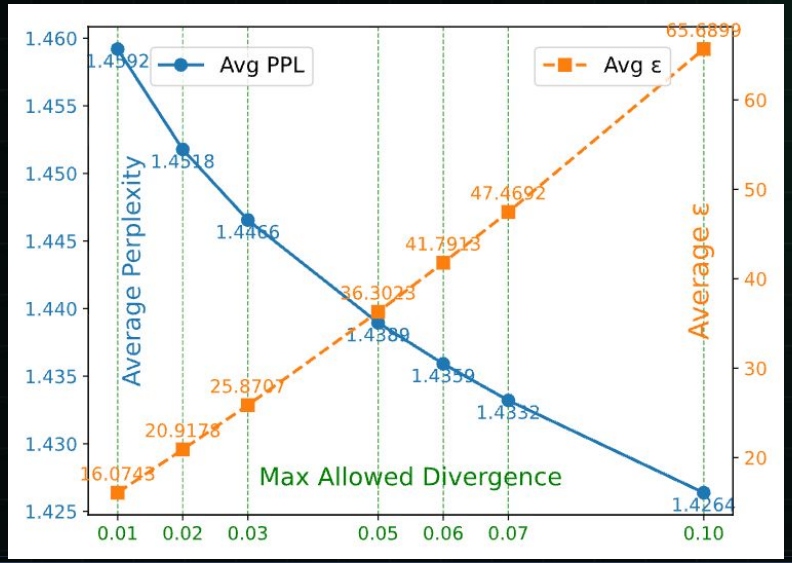
Performance: Win – Rate

Win-Rate (row beats column) of paraphrases



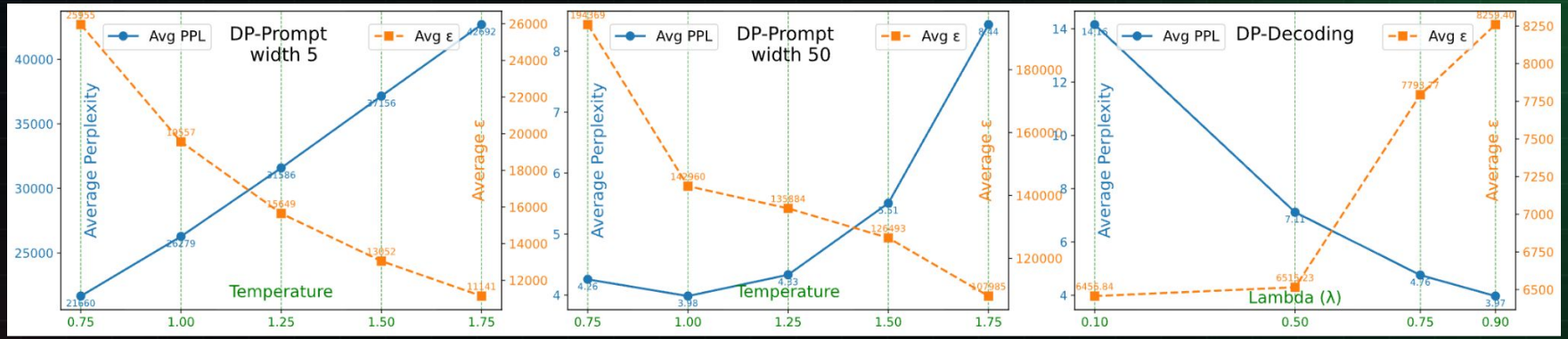
Performance: Privacy vs Utility

DP-Fusion perplexity vs epsilon



Performance: Privacy vs Utility

Privacy-vs-Utility: DP-Prompt and DP-Decoding across their respective parameter settings



Performance: Privacy vs Utility

Perplexity (utility) and ASR (privacy) are reported with Candidates = 5, random guessing gives 20% ASR.

Method	ppl	LOSS	MIN5%	MIN10%	MIN20%	MIN40%
No DPI - Original Document	1.03	0.6267	0.4633	0.5300	0.6033	0.6267
No DPI - NER	1.46	0.2767	0.2767	0.2734	0.29	0.2767
DP-Decoding $\lambda = 0.1$	14.15	0.1567	0.2033	0.1767	0.1600	0.1733
DP-Decoding $\lambda = 0.9$	3.96	0.6600	0.1067	0.1233	0.3567	0.5800
DP-Prompt (w=5,T=0.75)	>100	0.2667	0.2633	0.2533	0.2567	0.2367
DP-Prompt (w=5,T=1.75)	>100	0.1733	0.1933	0.1933	0.1500	0.1467
DP-Prompt (w=50,T=0.75)	4.26	0.5667	0.4300	0.4433	0.4667	0.5200
DP-Prompt (w=50,T=1.75)	8.44	0.2867	0.1633	0.1967	0.1967	0.1833
DP-FUSION (Ours), $\alpha\beta_i$=0.01	1.459	0.2600	0.2700	0.2733	0.2667	0.2633
DP-FUSION (Ours), $\alpha\beta_i$=0.10	1.426	0.2933	0.2933	0.2900	0.2900	0.2867