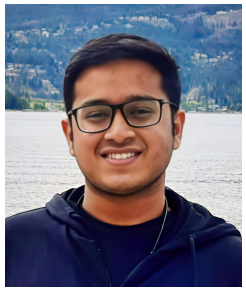
InvisibleInk

High-Utility and Low-Cost Text Generation with Differential Privacy

NeurIPS 2025

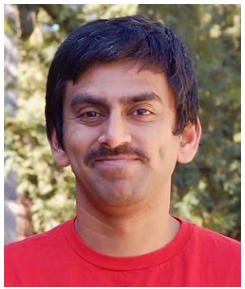


Vishnu Vinod

CeRAI, IIT Madras

Krishna Pillutla

WSAI & CeRAI, IIT Madras



Abhradeep Thakurta

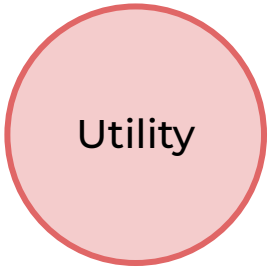
Google DeepMind

InvisibleInk (in one sentence)

High-quality long-form differentially private synthetic text generation with low computational overhead

InvisibleInk (in one sentence)

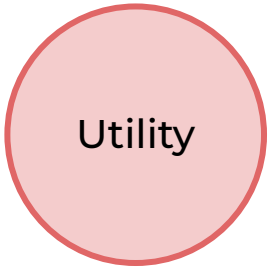
High-quality long-form differentially private synthetic text generation with low computational overhead



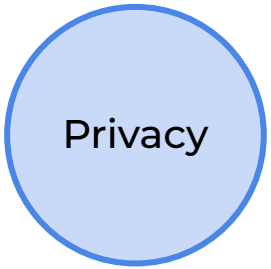
Quality of generated text; measured using MAUVE scores etc.

InvisibleInk (in one sentence)

High-quality long-form differentially private synthetic text generation with low computational overhead



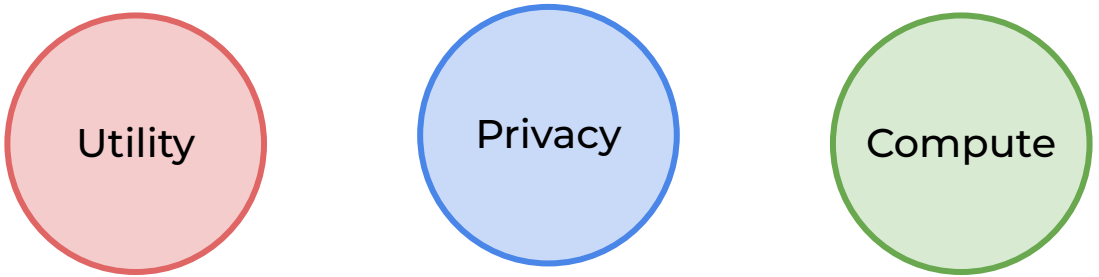
Quality of generated text; measured using MAUVE scores etc.



Parameters for the differential privacy guarantee (ε and δ)

InvisibleInk (in one sentence)

High-quality long-form differentially private synthetic text generation with low computational overhead



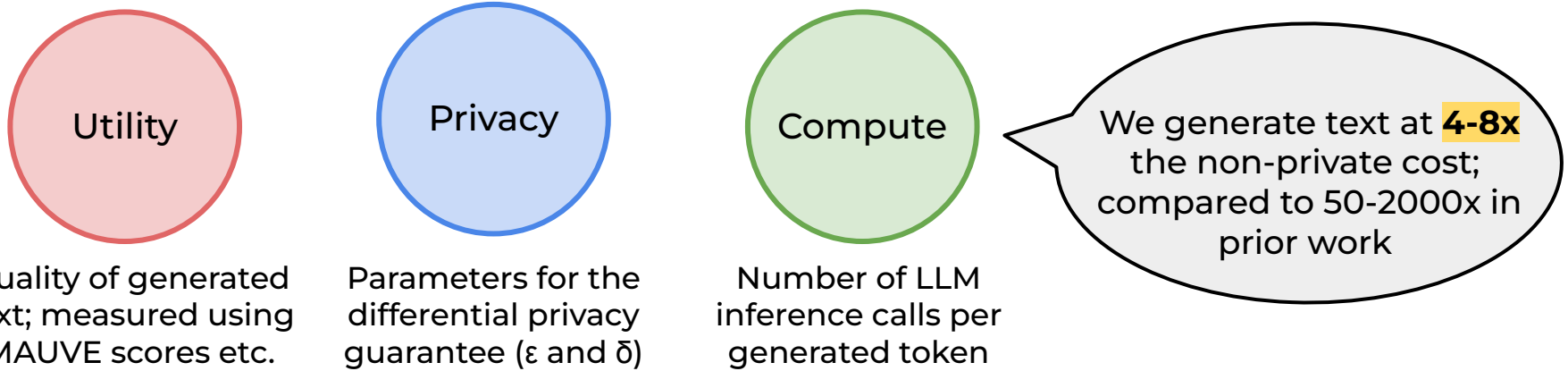
Quality of generated text; measured using MAUVE scores etc.

Parameters for the differential privacy guarantee (ε and δ)

Number of LLM inference calls per generated token

InvisibleInk (in one sentence)

High-quality long-form differentially private synthetic text generation with low computational overhead



Recent work shows LLM inference can leak privacy!

- LLMs can leak in-context information from prompts - *Wu et al.*, ICLR 2024
- LLM-generated synthetic data can leak privacy - *Meeus et al.*, ICML 2025
- RAG over sensitive databases can leak information - *Naseh et al.*, CCS 2025

Recent work shows LLM inference can leak privacy!

- LLMs can leak in-context information from prompts - *Wu et al.*, ICLR 2024
- LLM-generated synthetic data can leak privacy - *Meeus et al.*, ICML 2025
- RAG over sensitive databases can leak information - *Naseh et al.*, CCS 2025

We focus on the task of **synthetic text generation** where an LLM is prompted to generate text similar to *private references* it observes at *inference-time*.

Our Contributions

1. **InvisibleInk**: scalable private synthetic text generation by privatizing LLM decoding

2. **Experiments:** generate high-quality synthetic text at 8 times lower computational cost than baselines

3. **Practical and User-friendly:** heuristics for choosing optimal hyperparameters for private text generation

Our Contributions

1. **InvisibleInk**: scalable private synthetic text generation by privatizing LLM decoding
2. **Experiments:** generate high-quality synthetic text at 8 times lower computational cost than baselines
3. **Practical and User-friendly:** heuristics for choosing optimal hyperparameters for private text generation

LLM Decoding from logits

is cast as

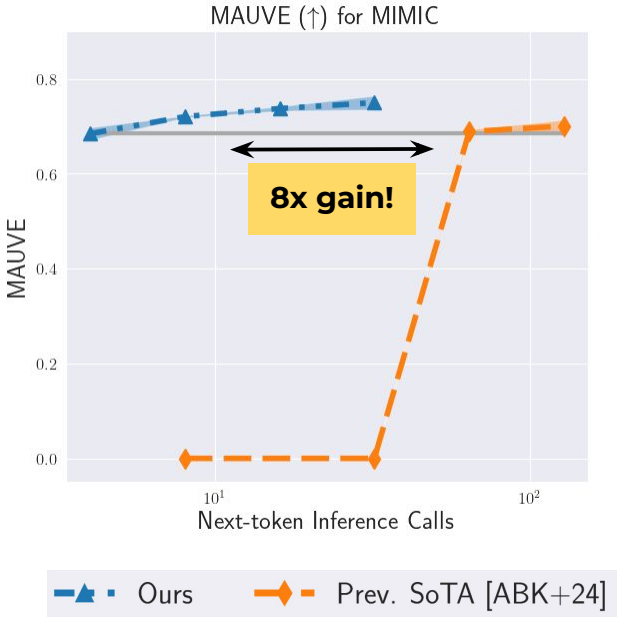
An instance of the canonical Exponential Mechanism for DP

Key Innovations:

- Difference Clipping (DClip)
- Expanded Top-k Sampling (top-k+)

Our Contributions

1. **InvisibleInk**: scalable private synthetic text generation by privatizing LLM decoding
2. **Experiments:** generate high-quality synthetic text at 8 times lower computational cost than baselines
3. **Practical and User-friendly:** heuristics for choosing optimal hyperparameters for private text generation



Our Contributions

1. **InvisibleInk**: scalable private synthetic text generation by privatizing LLM decoding
2. **Experiments:** generate high-quality synthetic text at 8 times lower computational cost than baselines
3. **Practical and User-friendly:** heuristics for choosing optimal hyperparameters for private text generation

Hyperparameter tuning adds to privacy cost - Papernot & Steinke, ICLR 2022

InvisibleInk has user-friendly non-adaptive (pre-hoc) DP-accounting.

Given a target privacy level and compute budget, we can calculate optimal hyperparameters prior to generation.

Background: Public and Private Logits

ϕ_i ⇒ Logits when LLM is prompted with *sensitive references* in-context

"Generate text like: *Agreed, let's meet at ....*"

ϕ_{pub}⇒ Logits when LLM has no sensitive information in-context

"Generate an email."

Contribution 1: Developing InvisibleInk

1a. Is clipping private logits the best we can do?

ϕ_{pub} : General information about language, syntax, grammar and world knowledge

ϕ_i : General information about language, syntax, grammar and world knowledge + Sensitive info about private references

We can isolate the contribution of the sensitive references by clipping the difference ϕ_i- ϕ_{pub} only !

1a. Is clipping private logits the best we can do?

ϕ_{pub} : General information about language, syntax, grammar and world knowledge

ϕ_i : General information about language, syntax, grammar and world knowledge + Sensitive info about private references

We can isolate the contribution of the sensitive references by clipping the difference $\phi_i - \phi_{pub}$ only !

$$\text{DClip}_C(\phi_i, \phi_{pub}) = \phi_{pub} + \text{clip}_C(\phi_i - \phi_{pub})$$

Sensitivity of clipped logits remains C/B

Clip the difference to lie between [-C, + C]

1a. Is clipping private logits the best we can do?

ϕ_{pub} : General information about language, syntax, grammar and world knowledge

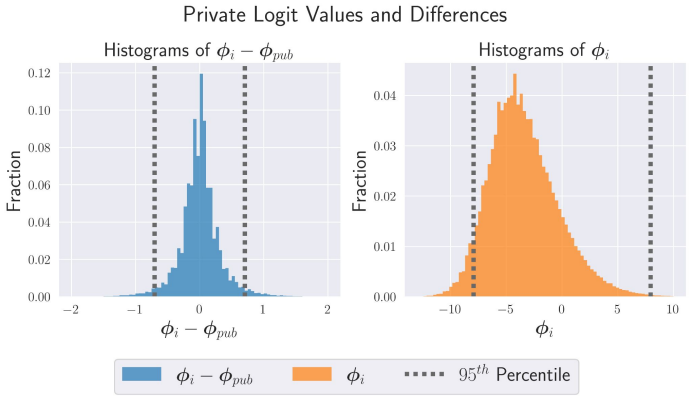
ϕ_i : General information about language, syntax, grammar and world knowledge + Sensitive info about private references

We can isolate the contribution of the sensitive references by clipping the difference ϕ_i- ϕ_{pub} only !

$$\text{DClip}_C(\phi_i, \phi_{pub}) = \phi_{pub} + \text{clip}_C(\phi_i - \phi_{pub})$$

Sensitivity of clipped logits remains C/B

Clip the difference to lie between [-C, + C]



Contributions of sensitive references (difference) are around 10x smaller in magnitude

1a. Is clipping private logits the best we can do?

ϕ_{pub} : General information about language, syntax, grammar and world knowledge

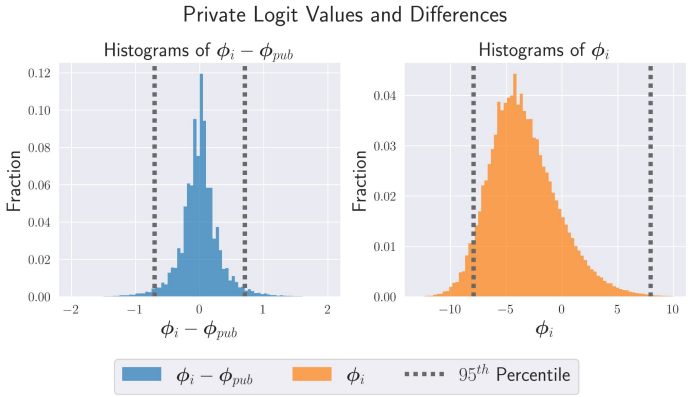
ϕ_i : General information about language, syntax, grammar and world knowledge + Sensitive info about private references

We can isolate the contribution of the sensitive references by clipping the difference ϕ_i - ϕ_{pub} only !

$$\text{DClip}_C(\phi_i, \phi_{pub}) = \phi_{pub} + \text{clip}_C(\phi_i - \phi_{pub})$$

Sensitivity of clipped logits remains C/B

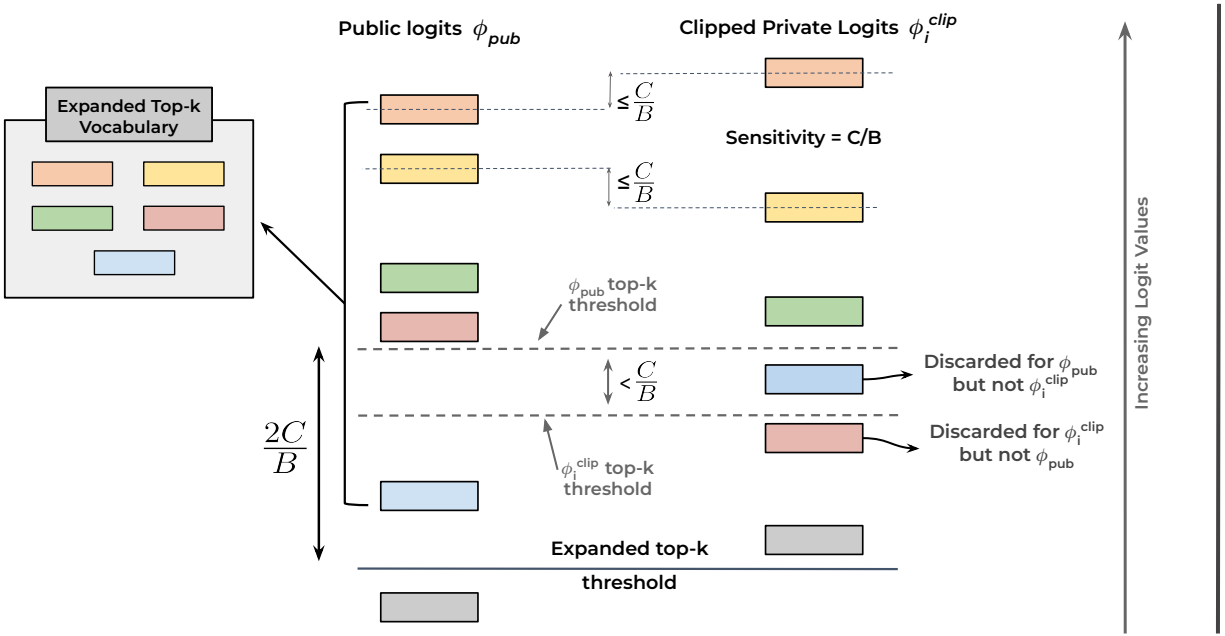
Clip the difference to lie between [-C, + C]



Contributions of sensitive references (difference) are around 10x smaller in magnitude

10x more data efficient ⟸ 10x smaller C ⟹ 10x smaller compute cost

1b. Expanded top-k sampling



- We align DP text generation with best practices in non-private settings using truncated decoding
- Using an expanded top-k set dependent only on public logits does not have additional privacy cost.

1c. Privacy Guarantees (zCDP)

Text sequence with at most T tokens, using batch of B references, clip norm C and sampling temperature τ:

$$\rho_{seq} \leq T \cdot \underbrace{\frac{\left(\mathrm{sens}(f_{\mathrm{DClip}})\right)^2}{2\tau^2}} = T \cdot \frac{C^2}{2B^2\tau^2}$$

Per-token privacy cost of exponential mechanism on clipped-and-averaged logits

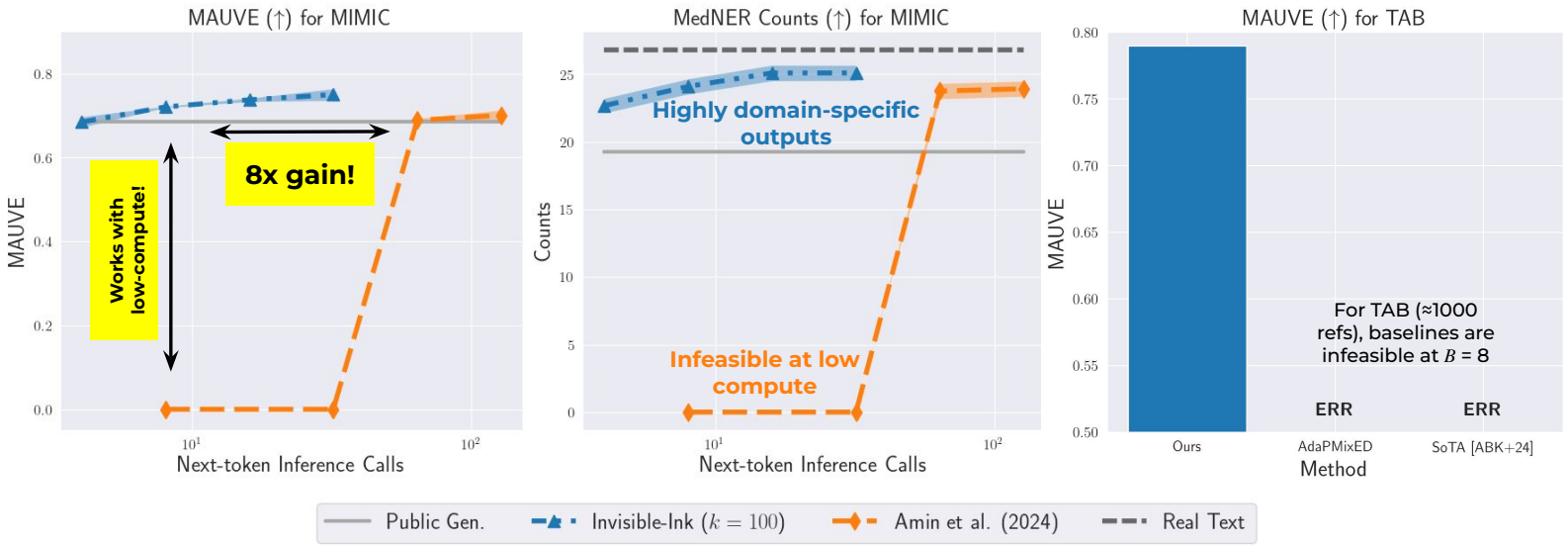
- Compose over T tokens using adaptive sequential composition of zCDP
- Compose over multiple sequences for disjoint batches using parallel composition of zCDP

Contribution 2: Experimental Work

2. Experimental Evaluation

Comparing InvisibleInk and prior SOTA (Amin et al., 2024) over MIMIC IV Notes and Text Anonymization Benchmark (low resource legal dataset).

Both are white-box approaches - assume access to model logits !



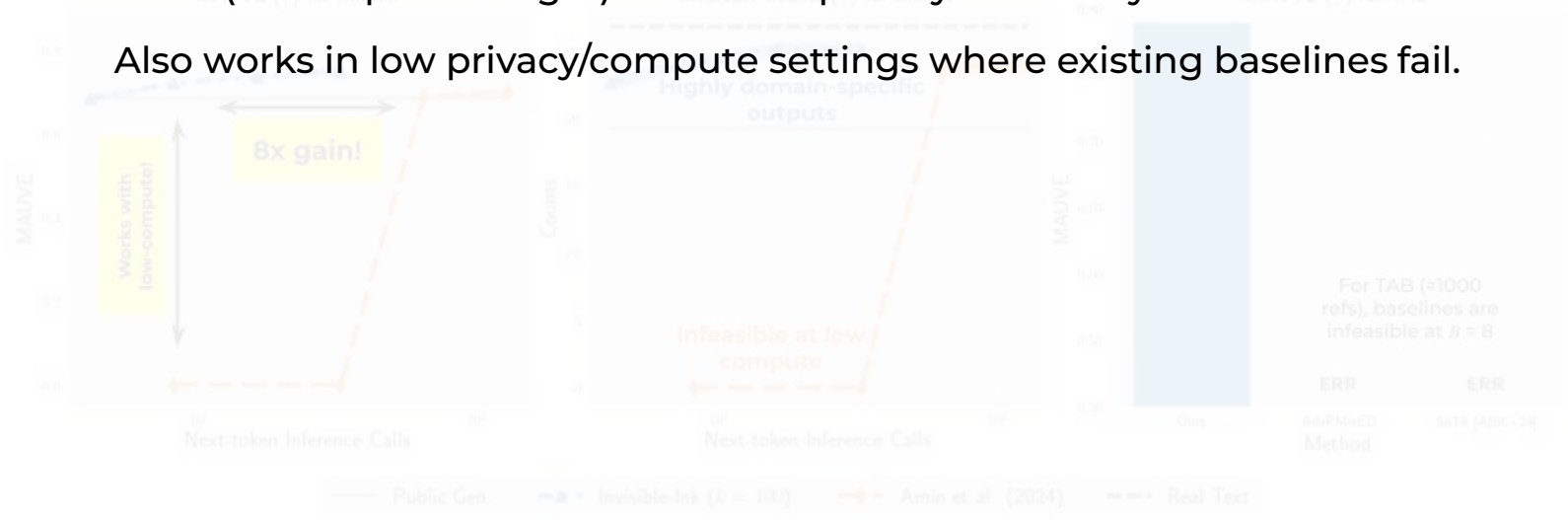
2. Experimental Evaluation

Comparing InvisibleInk and prior SOTA (Amin et al., 2024) over MIMIC IV Notes and Text Anonymization Benchmark (low resource legal dataset).

Both are white-box approaches - assume access to model logits !

Key Takeaways: InvisibleInk improves upon baselines by a factor of 8 or more (in compute budget) for similar privacy and utility levels.

Also works in low privacy/compute settings where existing baselines fail.



Contribution 3: Practical Recommendations

3. Practical Recommendations

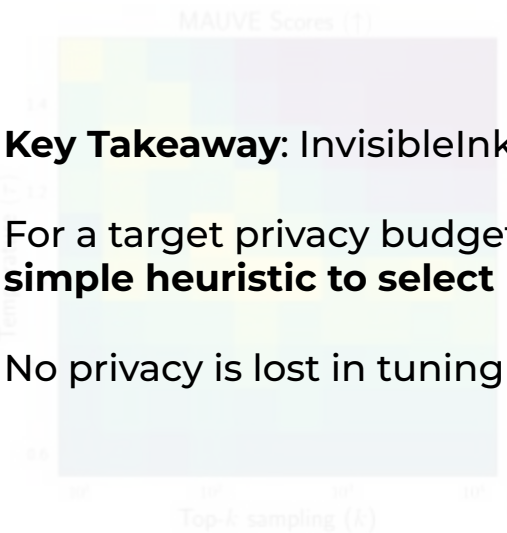
Practically, we know:

- ρ_{seq} : target privacy level
- B : fix a compute budget
- T : maximum new tokens

Then we set $\tau \approx 1$ and k ≈100 for generation and compute the required C:

$$C = \frac{B\tau}{\sqrt{2\rho_{seq}/T}}$$

3. Practical Recommendations



Key Takeaway: InvisibleInk has *user-friendly pre-hoc* privacy accounting.

For a target privacy budget and a maximum compute constraint, we give a **simple heuristic to select hyperparameters for near-optimal generations**!

No privacy is lost in tuning hyperparameters!

Best non-private generation quality is around $\tau \approx 1$ and k $\approx$ 100

Practically, we know:

- ρ_{seq} : target privacy level
- B : budget
- $\tau \approx 1$
- T : maximum new tokens

Then we can compute the required C

$$C = \frac{B\tau}{\sqrt{2\rho_{seq}/T}}$$

And finally set k $\approx$ 100 for generation !

Scalable DP Text Generation using InvisibleInk

Paper PDF (arXiv): https://arxiv.org/pdf/2507.02974

Python software package:
https://github.com/cerai-iitm/invisibleink

Code to reproduce results from the paper:
https://github.com/cerai-iitm/InvisibleInk-Experiments