AI TRAINING-DATA PROVENANCE ANALYSIS

Training-Data Exposure Report

Subject work: **Pride and Prejudice (Jane Austen, public domain)**

> Exposure indicator: **Strong verbatim reproduction on at least one model (investigative signal, not proof)**
>
> An investigative signal, NOT proof that your work was used in training. See methodology and limitations below.

Tier A — Verbatim reproduction (extraction)

Each model was given a short prefix of a distinctive passage of the subject work and asked to continue it. The table reports how far the model's output verbatim-matched the true continuation. Verbatim reproduction of long, distinctive passages is a fact about the model's behaviour and the soundest available signal.

Model	Probes	Strong	Partial	Refused
GPT-4o	3	**0**	0	3
Claude Sonnet 4.5	3	**0**	2	1
Gemini Pro	3	**0**	1	0
Llama 3.3 70B (open-weight)	3	**1**	1	0

Reproduced-text exhibits

Claude Sonnet 4.5 — partial (9-word verbatim run of a 107-word continuation from a 15-word prompt):

```
fortune, must be in want of a wife. However
```

Claude Sonnet 4.5 — partial (11-word verbatim run of a 120-word continuation from a 15-word prompt):

```
five thousand a year. What a fine thing for our girls!
```

Gemini Pro — partial (10-word verbatim run of a 120-word continuation from a 15-word prompt):

```
five thousand a year. What a fine thing for our
```

Llama 3.3 70B (open-weight) — strong (55-word verbatim run of a 107-word continuation from a 15-word prompt):

```
fortune, must be in want of a wife. However little known the feelings or views of such a
man may be on his first entering a neighbourhood, this truth is so well fixed in the minds
of the surrounding families, that he is considered the rightful property of some one or
other of their daughters.
```

Llama 3.3 70B (open-weight) — partial (11-word verbatim run of a 120-word continuation from a 15-word prompt):

```
five thousand a year. What a fine thing for our girls!
```

Tier A — Recognition (DE-COP multiple-choice)

Each model was shown the verbatim original of a passage alongside three close paraphrases and asked which is the original. A model that memorized the text recognizes the exact wording and scores above the 25% you would expect from guessing. This test is refusal-resistant, so models that declined to reproduce the text in Tier A still answer here.

Model	Questions	Answered	Correct	Accuracy	vs chance
GPT-4o	3	3	3	**100%**	+75 pts
Claude Sonnet 4.5	3	3	3	**100%**	+75 pts
Gemini Pro	3	3	3	**100%**	+75 pts
Llama 3.3 70B (open-weight)	3	3	3	**100%**	+75 pts

Chance is 25% (one verbatim option among four). Accuracy well above chance is a familiarity signal, not proof. Paraphrases were generated by openai:gpt-4o.

Tier B — Membership signal (relative)

For each open-weight model, the subject work's passages are scored with a Min-K%++ membership statistic and positioned relative to two reference corpora on the same model: known-trained public-domain text and freshly-synthesized text no model could have seen. A higher position toward the known-trained pole is a weak signal, not a probability.

Qwen 3.5 9B (open-weight): 2 of 3 passages sit closer to the known-trained pole (member median -1.220 vs non-member median -6.349; references 4/4).

How to read this report — methods & limitations

This report combines several methods, each with different strengths. They are explained below so every result can be read correctly. Probes are issued to each model's public API at temperature 0; model refusals and empty responses are recorded as facts, not as reproductions.

Tier A — Reproduction (extraction)

What it is. Each model is given a short prefix of a distinctive passage of your work and asked to continue it; we measure how much of the true continuation it reproduces word-for-word.
Why it is useful. Verbatim reproduction of a long, distinctive passage from a short prompt is a fact about the model's behaviour and the soundest single signal available — text a model has not seen is very unlikely to be reproduced at length by chance.
Its limits. Modern flagship chat models are guarded to refuse requests to reproduce copyrighted text, so a model that trained on your work may still decline (recorded here as a refusal, not a reproduction). It also needs long, distinctive passages to be meaningful.

Tier A — Recognition (DE-COP multiple-choice)

What it is. For each passage the model is shown the verbatim original alongside three close paraphrases and asked which is the original. We report how often it is right, versus the 25% you would expect from guessing.
Why it is useful. A model that trained on your work recognizes the exact wording it memorized and scores well above chance; a model that never saw it is at chance. Because it is a comprehension question rather than a request to reproduce text, guarded flagship models answer instead of refusing — and it needs only a single-letter reply, so it works on closed models with no special access.
Its limits. It is an inference method, not proof: above-chance accuracy is evidence of familiarity, not a certainty or an exact probability. The paraphrase distractors are machine-generated, and their quality affects difficulty; we report accuracy against the chance baseline so the signal is interpretable.

Tier B — Membership statistics

What it is. For open-weight models that expose token probabilities, we compute a Min-K%++ statistic over your text and report where it falls between known-trained and known-untrained reference corpora.
Why it is useful. It gives a per-passage signal even when a model neither reproduces nor is asked to recognize the text.
Its limits. Available only for open-weight models that return token log-probabilities. It is reported strictly as a relative position against reference corpora — never an absolute probability that your work was in the training set, which cannot be soundly computed for a production model.

Tier C — Copyright traps (prospective)

What it is. A unique, random canary document is published and cryptographically timestamped now, then models are re-probed over time.
Why it is useful. Because the canary is random and provably first published on a known date, a later reproduction is the one genuinely sound form of evidence that a model trained on it — it has a defensible false-positive rate.
Its limits. Prospective only: it produces evidence in the future, after models retrain on newer web crawls, not today.

Important: This report is an investigative signal, not proof. Verbatim reproduction (Tier A) evidences that a model can output your text, which is a fact about the model's behavior; it does not by itself establish how your work entered the model or that any legal right was infringed. Membership statistics (Tier B) are reported only as a position relative to reference corpora and do NOT constitute a probability that your work was in the training set -- no such probability can be soundly estimated for a production model. A negative or null result does not prove your work was excluded. This is not legal advice. BotWitness is a neutral third party and is not affiliated with the model providers.

Run id 0ccc11f663412663. Each model transcript is SHA-256 hashed; hashes are retained so any result in this report can be reproduced and audited.