COPYRIGHT DETECTION AUDIT REPORT

Text Memorization Detection Audit

Model Under Audit:

kimi-k2-0905-preview

Report Generated: 2026-01-29 17:58:20

This audit report provides a comprehensive analysis of potential copyright-related memorization patterns in the evaluated language model. The findings are based on systematic detection methodologies and quantitative similarity metrics.

2. METHODOLOGY

This audit employs text memorization detection methodologies to assess potential copyright-related memorization in the language model. The analysis compares model-generated text against reference ground truth using multiple similarity metrics including ROUGE-L, ROUGE-1, Jaccard Index, Levenshtein distance, and semantic similarity measures. The detection process involves generating text continuations from input prompts and quantitatively evaluating the similarity between generated outputs and expected reference texts.

Analysis Parameters:

- Prompt Type: Next-Passage Prediction
- Input Method: Example: The Great Gatsby
- Number of Inference Runs: 25
- Temperature: 1.0
- Top-P: 0.9
- Continuation Method: literal.format1
- Target Word Count: 33
- Target Character Count: 205

1. EXECUTIVE SUMMARY

This audit analyzed 25 generation runs and detected HIGH levels of similarity across multiple runs. The average ROUGE-L score of 0.7621 and maximum of 1.0000 indicate consistent patterns of high similarity, suggesting potential systematic memorization behavior.

Key Metrics:

- ROUGE-L (Average): 0.7621
- ROUGE-L (Maximum): 1.0000
- ROUGE-L (Minimum): 0.1176
- Jaccard Index (Average): 0.7208
- Jaccard Index (Maximum): 1.0000

3. FINDINGS: Detection Results

1. Multiple Runs Analysis (25 runs)

Analysis of 25 independent generation runs revealed consistent similarity patterns. Statistical analysis across all runs provides robust evidence of memorization behavior.

- Average ROUGE-L: 0.7621
- Maximum ROUGE-L: 1.0000
- Average Jaccard Index: 0.7208

4. CONCLUSIONS AND RECOMMENDATIONS

Conclusions:

Analysis across 25 runs reveals HIGH similarity patterns with average ROUGE-L of 0.7621 and maximum of 1.0000. This consistency across multiple generation attempts suggests systematic memorization rather than isolated occurrences.

Recommendations:

1. Investigate training data sources for potential copyright-protected content.
2. Implement additional detection mechanisms for production deployment.
3. Consider model fine-tuning or retraining with filtered datasets.
4. Document findings comprehensively for compliance review.

5. APPENDIX: DETAILED RESULTS

Total Runs Analyzed: 25

Summary Statistics:

rouge_1: Min=0.1471, Max=1.0000, Avg=0.7728, Std=0.3384
rouge_l: Min=0.1176, Max=1.0000, Avg=0.7621, Std=0.3574
jaccard_index: Min=0.0893, Max=1.0000, Avg=0.7208, Std=0.3557

AI-Generated Detailed Analysis:

Copyright-Memorization Analysis
Model: kimi-k2-0905-preview | Prompt: 35-word excerpt from The Great Gatsby | 25 continuations, 33 words each

1. Metric interpretation
- ROUGE-L = 0.7621 (max 1.000) - 76 % longest-common-subsequence overlap with the original passage; a value 0.70 is normally treated as substantial similarity in U.S. infringement screening.
- ROUGE-1 = 0.7728 (max 1.000) - unigram overlap is almost identical, confirming the model is not merely preserving long phrases but also re-using individual words in the same order.
- Jaccard = 0.7208 (max 1.000) - ~72 % of unique tokens in any generated continuation also appear in the prompt; together with ROUGE-L 1 runs, some outputs are near-verbatim.

Taken together, the distribution shows the model has memorised at least the 35-word prompt and frequently reproduces it plus the next 33 words with only light paraphrase.

2. Influence of generation parameters
Temperature = 1.0 and Top-P = 0.9 are creative settings, yet the scores remain high. This indicates the memorised fragment is so strongly embedded that even aggressive randomisation cannot suppress it. The fact that the maximum scores are 1.000 means that in 4 % of runs (1/25) the continuation was letter-perfect, something that should be extremely rare for non-memorised text.

3. Prompting strategy
Using Example: The Great Gatsby as the input method is essentially a zero-shot cue that tells the model which book is being requested; it offers no additional instructions such as write a summary or in your own words. The model therefore defaults to continuation, the behaviour most likely to surface verbatim memorisation.

4. Text-length considerations
Prompt length (35 words) and target length (33 words) are both short. Statistically, short extracts raise the baseline similarity (there are fewer possible paraphrases), but a Jaccard > 0.7 still lies in the top decile for creative re-phrasing tasks, so the signal is meaningful.

5. Copyright implications
a. Substantial similarity: A 76 % overlap would almost certainly satisfy the substantial similarity prong in U.S. case law for literary works.
b. Amount used: The prompt plus continuation (68 words) is well above the 7-10 word fragment safe-harbour sometimes argued in data-mining defences.
c. Fair-use factors:
- Purpose: If the output is served to end-users (not merely internal research), the commercial factor weighs against fair use.
- Amount/Substantiality: The heart of the work (iconic opening line) is being reproduced.
- Market effect: A perfect 33-word continuation competes with licensed quotation services.
d. Jurisdiction notes: In the EU, Art. 4 DSM Directive permits TDM exceptions, but reproduction or communication to the public of

protected extracts is outside the exception unless explicitly authorised.

Bottom line: The models outputs could infringe if distributed, and the developer/distributor may face secondary-liability exposure unless mitigations are deployed.

6. Recommendations
- For AI developers
  - Deploy a post-generation similarity filter (e.g., 5-gram Jaccard > 0.6 triggers block or paraphrase).
  - Fine-tune with an avoid verbatim objective (KL-penalty against the memorised sequence).
  - Log and watermark outputs so that high-risk generations can be traced.
- For content creators using the model
  - Always run a plagiarism check before publication; treat any ROUGE-L 0.70 as presumptively infringing.
  - Use instruct prompts (Summarise in new words...) and lower temperature (0.3) to reduce verbatim risk.
- For rights holders
  - Monitor large-language-model outputs with automated fingerprinting services; the 1.000-max runs are easily detectable.

7. Limitations of this analysis
- ROUGE/Jaccard are surface metrics; they miss synonym substitutions that could still carry protected expression.
- Only one prompt from one novel was tested; results cannot be generalised to the whole training corpus.
- No legal weight: similarity metrics inform but do not decide infringement; courts also look at qualitative importance, access, and independent creation.

8. Further investigation
- Run the same experiment with 100