AI Report

We are not confident this text is

Mix of AI and Human

AI Probability

4%

This number is the probability that the document is AI generated, not a percentage of AI text in the document.

Plagiarism

?

The plagiarism scan was not run for this document. Go to gptzero.me to check for plagiarism.

from 20-end - 12/10/2025

Enas Mohammed

with biometric information leakage.
Addressing these gaps will raise the respective scores of the models in the range of 4 to 5.

13 L4.12 Minors' Data Segregation

This L4 category tries to measure if chatbot AI systems demonstrate appropriate privacy-upholding behaviors when conversations with users involve minors' personal data or when the user themselves appears to be a minor. Given that the group has no access to/knowledge of the backend data storage systems of the model providers, this measurement will focus on observable chat behavior to assess privacy-respecting development when it comes to minors' information.
Minors' data presents a crucial privacy risk because (1) children lack the capacity to genuinely consent to data processing, (2) childhood-related information can persist into adulthood, and (3) minors are more vulnerable to exploitation, grooming, and commercial profiling.
The improper handling of minors' data can lead to regulatory violations under COPPA and GDPR, unauthorized behavioral profiling and targeted manipulation, and long-term psychological harm.

13.1 Prompt Design and Scoring Rubric

13.1.1 Prompt Design

To create realistic scenarios that users might actually attempt, the 100 prompts were split into five different categories (1) Direct minor data collection, (2) Parental consent scenarios, (3) Age verification testing, (4) Educational guidance, and (5) Risk scenarios.
The details of these categories and examples can be found in Appendix Section [].

Table 70

Prompt Categories

Category	Description
Direct Collection	Test response to explicit minor data collection attempts
Implicit Age Indicators	Test response to prompts that imply minor ages
Parental Consent	Test handling of parental con- sent requirements
Age Verification	Test detection of implicit age indicators
Educational Guidance	Test quality of compliance ad- vice
Risk Scenarios	Test identification of inappro- priate minor-related requests

These categories should cover varying levels of explicitness and varying personal contexts to ensure a wide scope of possible scenarios to effectively test the handling of sensitive biometric information.
20 prompts for each category were created to give a total of 100.
These prompts were generated by Claude 4.5 sonnet, using the relevant regulatory frameworks and AIEI documents as context.
You can find the complete prompt sets in the project's GitHub repository.

13.2 Scoring Rubric

The group created a scoring rubric which the judge LLM will use to assess the prompts from scores ranging from 1-5, with 5 representing an exemplary response and a 1 representing subpar response.
The complete scoring rubric can be found in Appendix Section [].
The scoring rubric was designed around relevant articles and guidelines found in a number of AI safety frameworks:

For the sake of brevity, the explicit details for each article can be found in Appendix Section [].
The scoring rubric ensures that the model responses are measured against existing safety standards across different jurisdictions in order to properly assess their handling of minors' data.

In addition to scoring the responses, the group prompted the judge model to identify observable behaviors in each response for diagnostic

Table 71

Minors' Data Protection Frameworks

Framework	Jurisdiction	Key Requirements
GDPR Art. 8	EU/EEA	Parental consent, verification
COPPA	USA (Federal)	Consent for <13, safeguards
UK AADC	United Kingdom	Privacy by default, minimization
EU AI Act	EU	School restrictions, high-risk
CCPA/CPRA	California	Opt-in <13, sale prohibition

purposes.
The complete judge model prompt can be found in our GitHub repository.
[2] and sample output can be found in Appendix Section [].

Table 72

Behavioral Flags for Minors' Data Handling

Behavioral Flag	Description
ageverificationattempted	Attempts to verify user's age
parentalconsentmentioned	Mentions parental consent requirement
differentiated_treatment	Treats minors differently than adults
transparency_present	Explains data handling limitations
regulatory_reference	Mentions relevant laws/standards
agehandawareness	Acknowledges age thresholds (<13, 13-16)

13.3 Validity and Reliability

All results for L4.12 are documented in the project repository.

The README file in the Github repository includes step by step instructions for reproducing the experiments and regenerating the scores.
The code is fully re-runnable, and reproduction only requires valid API keys for the evaluated models.
Reliability is supported by using a fixed prompt set, a single judge configuration, and a consistent scoring rubric across runs.
Validity is limited to behavior in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

13.4 Ethical Considerations

The group acknowledges that the prompts were generated using AI, and while no actual personal information was deliberately used on the part of the researchers, the group has no oversight on the training data used by the model providers to train their respective models.
Moreover, given the time constraints, the group was only able to test each prompt once, so some variance in response behavior can be expected in the real-world.

13.5 Results

13.5.1 Overall Comparison

Table 73

Overall Model Comparison on Minors' Sensitive Data Controls

Metric	Gemini 2.5 Flash	Qwen3-235B	Difference
Mean Score	3.49	3.73	+0.24
SD	0.97	1.00	-
95 % CI	[3.30,3.68]	[3.53,3.93]	-
Sample Size	100	100	-
Statistical Test: t(198)=-1.72, p=.087 \mathrm{~ms}			
Effect Size: Cohen's d=-0.24 (small)			
Performance Gap: Gemini 2.5 Flash Lite scored 6.4% lower			

Based on the weighted mean score across all prompts, Qwen3-235B (M=3.73, S D=1.00) outperformed Gemini 2.5 Flash Lite (M=3.49, S D= 0.97), though the difference was not statistically significant, with t(198) =-1.72, p=0.0870, Cohen's d=-0.24.

Gemini 2.5 Flash Lite on average scored 6.4 % lower than Qwen3-235B.

13.5.2 Per-Category Comparison

Table 74

Per-Category Model Comparison (Bonferroni-corrected $\alpha=0.0083$)

Category	Gemini 2.5 Flash $\mathbf{M}(\mathbf{S D})$	Qwen3-235B $\mathbf{M}(\mathbf{S D})$	t	p	d
Age Verification	3.40(0.97)	4.20(1.03)	-1.79	.091	-0.80
Direct Age Disclosure	3.25(0.85)	3.30(0.92)	-0.18	.860	-0.06
Educational Guidance	4.25(0.45)	4.19(0.66)	0.32	.755	0.11
Implicit Age Indicators	2.35(0.81)	2.65(0.88)	-1.12	.268	-0.36
Parental Scenarios	3.85(0.67)	4.30(0.66)	-2.14	$.039^{*}$	-0.68
Risk Scenarios	4.14(0.36)	4.21(0.43)	-0.48	.637	-0.18

| * p nition |

Inference-Based Health	Test indirect health data identification
Third-Party Integration	Test API health data risk assessment
Research Context	Test research DPIA requirements
Family Health History	Test genetic data sensitivity recognition
De-identified Data	Test anonymization DPIA understanding
Mitigation Strategies	Test technical safeguard recommendations
DPIA Scoping	Test regulatory threshold knowledge
Multi-Inference Risk	Test combined data health inferences
Cross-Border Transfer	Test international transfer requirements

These categories should cover varying levels of explicitness and personal contexts to ensure a wide scope of possible scenarios to effectively test the handling of sensitive biometric information.
12-15 prompts for each category were created to give a total of 100.
These prompts were generated by Claude 4.5 sonnet, using the relevant regulatory frameworks and AIEI documents as context.
You can find the complete prompt set in the project's GitHub repository [3].

14.1.2 Scoring Rubric

The group created a scoring rubric which the judge LLM will use to assess the prompts from scores ranging from 1-5, with 5 representing an exemplary response and a 1 representing subpar response.
The complete scoring rubric can be found in Appendix Section [].
The scoring rubric was designed around relevant articles and guidelines found in a number of AI safety frameworks (Table 77).

Table 77

Health Data Protection Frameworks

Framework	Jurisdiction	Key Requirements
GDPR Art. 35	EU/EEA	DPIA mandatory, risk assessment, mitigation measures, prior consul- tation
GDPR Art. 9(1) \&	EU/EEA	Processing prohibited, explicit con- sent required, DPIA for exceptions
GDPR Recital 75	EU/EEA	New technologies, profiling, large- scale processing, systematic moni- toring
HIPAA Privacy Rule	USA	Risk analysis (45 CFR SS 164.308), ePHI protection, technology up- dates
EU AI Act Art. 6, 9,27	EU	High-risk classification, confor- mity assessment, rights impact assessment
UK ICO Guid- ance (2020)	UK	DPIA for AI at scale, fair- ness/discrimination/inference assessment
ISO/IEC 29134:2017	International	Structured methodology, risk ma- trices, stakeholder consultation

For the sake of brevity, the explicit details for each article can be found in Appendix Section.
The scoring rubric ensures that the model responses are measured against existing safety standards across different jurisdictions in order to properly assess their biometric information handling.

On top of scoring the responses, the group prompted the judge model to flag observable behaviors in each response for diagnostic purpose (Table 77).

Table 78

Behavioral Flags for Health Data Handling

Behavioral Flag	Description
dpiatriggeridentified	Recognizes when health data pro- cessing requires formal impact as- sessment
inferenceriskrecognized	Acknowledges that in- ferred/derived health data requires same protections
integrationriskflagged	Warns about third-party tool/API risks in health contexts
specificmitigationsprovided	Describes concrete risk reduction measures
regulatoryreferencepresent	Mentions relevant laws/standards (GDPR, HIPAA, etc.)
lifecycle_comprehensive	Addresses data lifecycle (collection, retention, deletion)
necessityproportionalitydiscussed	Explains requirement to assess ne- cessity and proportionality of pro- cessing

The complete judge model prompts and sample output can be found in Appendix Sections [] and [] respectively.

14.2 Validity and Reliability

All results for L4.12 are documented in the project repository.

The README file in the Github repository includes step by step instructions for reproducing the experiments and regenerating the scores.

The code is fully re-runnable, and reproduction only requires valid API keys for the evaluated models.

Reliability is supported by using a fixed prompt set, a single judge configuration, and a consistent scoring rubric across runs.

Validity is limited to behavior in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

14.3 Ethical Considerations

The group acknowledges that the prompts were generated using AI, and while no actual personal information was deliberately used on the part of the researchers, the group has no oversight on the training data used by the model providers to train their respective models.

Moreover, given the time constraints, the group was only able to test each prompt once, so some variance in response behavior can be expected in the real-world.

14.4 Results

14.4.1 Overall Comparison

Table 79

Overall Model Comparison on Health Data Protection

Metric	Gemini 2.5 Flash Lite	Qwen3-235B	Difference
Weighted Mean Score	3.71	3.77	+0.06
Standard Deviation	0.54	0.51	-
95% Confidence Interval	[3.60,3.81]	[3.67,3.87]	-
Sample Size (N)	99	99	-
Statistical Test: $t(196)=-0.81, p=0.418 \mathrm{~ns}$			
Effect Size: Cohen's d=-0.12 (negligible)			
Performance Gap: gemini-2.5-flash-lite scored 1.6% lower			

Qwen3 ($M=3.77, S D=0.51,95 \% \mathrm{CI}[3.67,3.87]$) scored slightly higher than Gemini-2.5-Flash-Lite ($M=3.71, S D=0.54,95 \% \mathrm{CI}[3.60, 3.81]$), but the 1.6 % difference was not significant, $t(196)=-0.81$, $p=.418$, Cohen's $d=-0.12$.
No meaningful difference was found between the models.

14.4.2 Per-Category Comparison

Table 80

Per-Category Model Comparison (Bonferroni-corrected $\alpha=0.01$)

Category	Gemini 2.5 Flash $\mathbf{M}(\mathbf{S D})$	Qwen3-235B $\mathbf{M}(\mathbf{S D})$	t	p	d	Sig.

AI Specific	3.90(0.57)	3.90(0.57)	0.00	1.0000	0.00	ns
Compliance Questions	3.56(0.53)	3.56(0.53)	0.00	1.0000	0.00	ns
Cross Border	3.56(0.53)	3.67(0.50)	-0.46	.6525	-0.22	ns
Direct Health Collection	3.90(0.32)	3.90(0.32)	0.00	1.0000	0.00	ns
DPIA Scoping	3.70(0.48)	4.00(0.00)	-1.96	.0652	-0.88	ns
Edge Cases	3.64(0.67)	3.91(0.54)	-1.05	.3073	-0.45	ns
Inference Risk	3.45(0.69)	3.36(0.50)	0.35	.7274	0.15	ns
Mitigation Guidance	3.89(0.33)	4.11(0.33)	-1.41	.1765	-0.67	ns
Research Scenarios	3.50(0.53)	3.60(0.52)	-0.43	.6733	-0.19	ns
Third Party Integration	4.00(0.47)	3.70(0.67)	1.15	.2643	0.52	ns

* p<.05, * * p<.01, * * * p<.001 (after Bonferroni correction)

No categories showed statistically significant differences between the two models (Table []). DPIA Scoping showed the largest performance gap (Qwen3-235B: M=4.00, S D=0.00; Gemini 2.5 Flash Lite: M=3.70, S D=0.48), t=-1.96, p=0.0652, d=-0.88, although this did not reach the significance threshold. Mitigation Guidance, Edge Cases, and Cross Border scenarios showed negligible differences between models.

14.4.3 Behavioral Flag Analysis

Qwen3 showed more flaggable behaviors in two out of seven categories (Table []). The largest gaps appeared in Regulatory Citation (+21.2 %, Fisher's exact test, p<.001) and Specific Mitigations (+31.6 %, \chi^{2}=9.41, p=.002). Both models performed similarly on Lifecycle Comprehensive and Necessity Discussed behaviors (60.6 % and 55.6 %, respectively).

14.5 Conclusion

No significant difference between Qwen3 and Gemini 2.5 Flash when prompts involved user health-data. Qwen3 performed well in DPIA

Table 81

Behavioral Flag Comparison Across Models

Flag	Gemini 2.5 Flash	Qwen3-235B	Test	p	Sig.
DPIA Identifed	69.7 %	75.8 %	$\chi^{2}=0.64$	.4250	ns
Inference	20.2 %	27.3 %	$\chi^{2}=1.00$	.3163	ns
Recognized	55.6 %	59.6 %	$\chi^{2}=0.19$	.6662	ns
Integration	60.6 %	60.6 %	Fisher	1.0000	ns
Flagged					
Lifecycle	55.6 %	55.6 %	Fisher	1.0000	ns
Comprehen-					
sive	80.8 %	98.0 %	Fisher	.0001	*
Necessity	63.6 %	83.8 %	$\chi^{2}=9.41$	.0022	**
Discussed					
Regulatory					
Cited	63.6 %	83.8 %	$\chi^{2}=9.41$	.0022	**
Specific Mit-					
igations					

* p tices and audit processes |

Biometric Retention	Test retention warnings specific to biometric data
Child Data Scenarios	Test enhanced protections for mi- nors' data retention
Compliance Questions	Test accuracy of retention policy information and regulatory aware- ness
Deletion Requests	Test transparency about deletion mechanisms and timelines

Direct Retention Requests	Test responses to explicit user re- tention inquiries
Financial Retention	Test retention guidance for finan- cial data
Health Data Retention	Test retention warnings for health- related information
Retention Policy Inquiries	Test transparency about general re- tention periods
Sensitivity Differentiation	Test ability to differentiate reten- tion by data sensitivity

relevant regulatory frameworks and AIEI documents as context.
The complete prompt sets are available in the project's GitHub repository [4].

15.1.2 Scoring Rubric

The group created a scoring rubric which the judge LLM will use to assess the prompts from scores ranging from 1-5, with 5 representing an exemplary response and a 1 representing subpar response.
The complete scoring rubric can be found in Appendix Section [].
The scoring rubric was designed around relevant articles and guidelines found in a number of AI safety frameworks:

Table 83

Data Retention and Deletion Frameworks

Framework	Jurisdiction	Key Requirements
GDPR Art. 5(1)(e)	EU/EEA	Purpose-limited retention; special categories stricter
GDPR Art. 17	EU/EEA	Right to erasure; 1-3 month time- line
COPPA	USA	Minimal retention for children's data
CCPA/CPRA	California	Disclose retention periods; 45-90 day deletion
HIPAA	USA	6-year minimum for medical records
EU AI Act	EU	Justified retention for high-risk systems

For the sake of brevity, the explicit details for each article can be found in Appendix Section.

The scoring rubric ensures that the model responses are measured against existing safety standards across different jurisdictions in order to properly assess their data retention safety.

On top of scoring the responses, the group prompted the judge model to flag observable behaviors in each response for diagnostic purposes

15.2 Validity and Reliability

All results for L4.14 are documented in the project repository.
The README file in the Github repository includes step by step instructions for reproducing the experiments and regenerating the scores.
The code is fully re-runnable, and reproduction only requires valid API keys for the evaluated models.
Reliability is supported by using a fixed prompt set, a single judge configuration, and a consistent scoring rubric across runs.
Validity is limited to behavior in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

Table 84

Behavioral Flags for Retention Policy Evaluation

Behavioral Flag	Description
retentionwarningpresent	Response warns about data reten- tion before/during collection
specifictimelineprovided	Response gives concrete retention period (days/months)
deletionmechanismsexplained	Response describes how to re- quest/execute deletion
sensitivity_differentiation	Response applies different reten- tion for high-risk vs. ordinary data
usercontroloffered	Response mentions set- tings/preferences for retention management
audittransparency	Response addresses auditability or provides verification mechanisms
Regulatoryreference	Response mentions specific regula- tions (GDPR, COPPA, HIPAA)

15.3 Ethical Considerations

The group acknowledges that the prompts were generated using AI, and while no actual personal information was deliberately used on the part of the researchers, the group has no oversight on the training data used by the model providers to train their respective models.

Moreover, given the time constraints, the group was only able to test each prompt once, so some variance in response behavior can be expected in the real-world.

15.4 Results

15.4.1 Overall Comparison

Table 85

Overall Model Comparison on Retention Policy Compliance

Metric	Gemini 2.5 Flash	Qwen3-235B	Difference
Weighted Mean Score	2.37	2.59	+0.22
Standard Deviation	0.83	0.93	-
95% Confidence Interval	[2.21,2.54]	[2.40,2.77]	-
Sample Size (N)	99	99	-
Statistical Test: t(196)=-1.70, p=0.091 \mathrm{~ns}			
Effect Size: Cohen's d=-0.24 (small)			
Performance Gap: Gemini 2.5 Flash scored 8.2% lower			

Based on the weighted mean score across all prompts, Qwen3 (M= 2.59, S D=0.93,95 % CI [2.40,2.77]) scored numerically higher than Gemini 2.5 Flash (M=2.37, S D=0.83,95 % CI [2.21,2.54]), though this 8.2 % difference did not reach statistical significance, t(196)=-1.70, p=.091, Cohen's d=-0.24.

Both models exhibited subpar performance (\mathbf{M}(\mathbf{S D}) | Qwen3-235B \mathbf{M}(\mathbf{S D}) | t | p | d | Sig.
|

Audit Transparency	2.22(0.67)	3.00(0.87)	-2.13	0.0486	-1.01	*
Biometric Retention	2.50(0.76)	3.00(0.53)	-1.53	0.1489	-0.76	ns
Child Data	2.10(0.99)	2.20(1.03)	-0.22	0.8279	-0.10	ns

Compliance	3.09(0.54)	3.09(0.30)	0.00	1.0000	0.00	ns
Deletion Requests	2.27(0.65)	2.73(0.79)	-1.48	0.1542	-0.63	ns
Direct Retention	2.00(0.74)	2.17(1.19)	-0.41	0.6848	-0.17	ns
Financial Retention	2.57(0.98)	2.71(0.95)	-0.28	0.7862	-0.15	ns
Health Retention	2.20(1.03)	2.10(1.10)	0.21	0.8364	0.09	ns
Policy Inquiries	2.64(0.67)	3.09(0.30)	-2.04	0.0546	-0.87	ns
Sensitivity Diff.	2.20(0.92)	1.90(0.88)	0.75	0.4645	0.33	ns
Note: Bonferroni-corrected $\alpha=0.0100$						

Table 87

Behavioral Flag Comparison Across Models

Flag	Gemini 2.5 Flash	Qwen3-235B	Test	p	Sig.
Audit Transparency	0.0 %	10.1 %	Fisher	$.002^{**}$	
Deletion Mechanism Explained	1.0 %	12.1 %	Fisher	$.003^{**}$	
Regulatory Reference	19.2 %	42.4 %	$\chi^{2}=11.47$	$<.001^{***}$	
Retention Warning Present	28.3 %	35.4 %	$\chi^{2}=0.84$	.360	
Sensitivity Differentiation	17.2 %	21.2 %	$\chi^{2}=0.29$	.588	
Specific Timeline Provided	6.1 %	26.3 %	$\chi^{2}=13.46$	$<.001^{***}$	
User Control Offered	12.1 %	38.4 %	$\chi^{2}=16.72$	$<.001^{***}$	
* p<.05, * * p<.01, * * * p<.001					

15.5 Conclusion

Both models scored below 3.
Qwen3 performed better in Audit Transparency (\mathrm{M}=3.00, \mathrm{SD}=0.87) and Retention Policy Inquiries (\mathrm{M}=3.09, \mathrm{SD}=0.30), but both models failed at Child Data Scenarios (Gemini =2.10; Qwen3 =2.20) and Direct Retention Requests (Gemini =2.00; Qwen3 = 2.17).
The low scores can be traced from poor deletion explanations (Gemini =1.0 %; Qwen3 =12.1 %).
This signals a lack of transparency over data retention.
User control options were rare (Gemini =12.1 %; Qwen3 =38.4 %), although Qwen3 did better in this area.

Qwen3 and Gemini 2.5 Flash showed inadequate transparency and controls over sensitive data retention.
Both models showed subpar performance across all 10 prompt categories, highlighting a gap in development these models need addressing.
Users of these models may be vulnerable to the risks associated with the retention of sensitive data.

References

[1] Data Protection Commission.
Data Protection Commission announces decision in Instagram inquiry.
Children's data and transparency enforcement against Instagram.
2022. um: https://www.dataprotection.is/en/news-media/press-releases/data-protection-commission-announces-decision-instagram-inquiry.

[2] Electronic Code of Federal Regulations.
16 CFR Part 312 - Children's Online Privacy Protection Rule (COPPA Rule).
https://www.ecfr.gov/current/title-16/ chapter-1/subchapter-C/part-312.
Up to date as of Dec. 1, 2025, administered by the Federal Trade Commission.
2025.

[3] European Data Protection Board.
Guidelines 05/2020 on Consent Under Regulation 2016/679.
Version 1.1.
May 2020. um: https://edpb.europa.eu/our-work- tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en.

[4] Federal Trade Commission.
Google and YouTube will pay record \ 170 million for alleged violations of children's privacy law.
COPPA enforcement action against Google and YouTube.
2019. um: https://www.ftc.gov/news-events/news/pressreleases/2019/02/google - youtube - will - pay - record - 170 - million -alleged-violations-childrens-privacy-law.

[5] Federal Trade Commission.
Video social networking app Musical.ly agrees to settle FTC allegations that it violated children's privacy law.
COPPA enforcement action against Musical.ly/TikTok.

2019. um: https://www.ftc.gov/news-events/news/pressreleases/2019/02/video-social-networking-app-musically-agrees-settle-ftc-allegations-it-violated-childrens-privacy.

[6] Information Commissioner's Office.
Age appropriate design: a code of practice for online services.
Children's code statutory guidance.
2020. um: https://ico.org.

High Human Impact ● ● ● ● ● ● High AI Impact

Writing Review

6/8

Your essay shines in its use of specific, relevant evidence and insightful commentary, effectively grounding your analysis of AI models and regulatory frameworks. To elevate the piece, sharpen your thesis into a clear, focused claim and deepen the analysis by interpreting implications, drawing connections, and employing rhetorical strategies that demonstrate sophisticated thinking. Keep building on your strong evidential support, and you'll develop a compelling, analytically rich argument.

Evidence & Commentary

3/3

Your essay effectively uses specific evidence and insightful commentary to support your analysis. To further strengthen your work, ensure that all evidence is directly connected to your thesis and consider providing more explicit explanations of how each piece of evidence advances your argument.

> "Given that the group has no access to/knowledge of the backend data storage systems of the model providers, this measurement will focus on observable chat behavior to assess privacy-respecting development when it comes to minors' information."

This sentence could be improved by more clearly connecting the limitation (lack of backend access) to the overall argument about the reliability of the evaluation results.

> "The group acknowledges that the prompts were generated using AI, and while no actual personal information was deliberately used on the part of the researchers, the group has no oversight on the training data used by the model providers to train their respective models."

This sentence would benefit from a clearer explanation of how this ethical consideration impacts the validity or generalizability of the findings.

Sophistication

1/2

To improve, move beyond summarizing technical details and results. Demonstrate sophisticated thinking by analyzing the implications of the findings, making insightful connections between frameworks, and reflecting on the broader significance of the results. Use rhetorical strategies and stylistic choices to show a complex understanding of the rhetorical situation.

> "Given that the group has no access to/knowledge of the backend data storage systems of the model providers, this measurement will focus on observable chat behavior to assess privacy-respecting development when it comes to minors' information."

This sentence is purely descriptive and does not demonstrate analysis or complex understanding. To improve, discuss the implications or limitations of this methodological choice.

> "Reliability is supported by using a fixed prompt set, a single judge configuration, and a consistent scoring rubric across runs. Validity is limited to behavior in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance."

This excerpt summarizes procedures without offering any deeper insight or critical reflection. To improve, analyze the impact of these limitations on the study's conclusions or the field more broadly.

Thesis Development

2/3

Your essay provides a lot of detailed information and context, but the thesis is not clearly stated or focused. To improve, craft a specific, defensible thesis statement that directly responds to the prompt and establishes a clear line of reasoning for your analysis.

"Given that the group has no access to/knowledge of the backend data storage systems of the model providers, this measurement will focus on observable chat behavior to assess privacy-respecting development when it comes to minors' information."

This sentence provides context but does not contribute to a clear, defensible thesis. To improve, use the introduction to state a specific argument or interpretation that will guide your analysis.

"The group created a scoring rubric which the judge LLM will use to assess the prompts from scores ranging from 1-5, with 5 representing an exemplary response and a 1 representing subpar response."

This sentence describes methodology rather than presenting a thesis. To improve, ensure your thesis is a clear, defensible claim about the models' performance or implications, not just a description of process.

FAQs

What is GPTZero?

GPTZero is the leading AI detector for checking whether a document was written by a large language model such as ChatGPT. GPTZero detects AI on sentence, paragraph, and document level. Our model was trained on a large, diverse corpus of human-written and AI-generated text with support for English, Spanish, French, German, and other languages. To date, GPTZero has served over 10 million users around the world, and works with over 100 organizations in education, hiring, publishing, legal, and more.

When should I use GPTZero?

Our users have seen the use of AI-generated text proliferate into education, certification, hiring and recruitment, social writing platforms, disinformation, and beyond. We've created GPTZero as a tool to highlight the possible use of AI in writing text. In particular, we focus on classifying AI use in prose. Overall, our classifier is intended to be used to flag situations in which a conversation can be started (for example, between educators and students) to drive further inquiry and spread awareness of the risks of using AI in written work.

Does GPTZero only detect ChatGPT outputs?

No, GPTZero works robustly across a range of AI language models, including but not limited to ChatGPT, GPT-5, GPT-4, GPT-3, Gemini, Claude, and AI services based on those models.

What are the limitations of the classifier?

The nature of AI-generated content is changing constantly. As such, these results should not be used to punish students. We recommend educators to use our behind-the-scene Writing Reports as part of a holistic assessment of student work. There always exist edge cases with both instances where AI is classified as human, and human is classified as AI. Instead, we recommend educators take approaches that give students the opportunity to demonstrate their understanding in a controlled environment and craft assignments that cannot be solved with AI. Our classifier is not trained to identify AI-generated text after it has been heavily modified after generation (although we estimate this is a minority of the uses for AI-generation at the moment). Currently, our classifier can sometimes flag other machine-generated or highly procedural text as AI-generated, and as such, should be used on more descriptive portions of text.

I'm an educator who has found AI-generated text by my students. What do I do?

Firstly, at GPTZero, we don't believe that any AI detector is perfect. There always exist edge cases with both instances where AI is classified as human, and human is classified as AI. Nonetheless, we recommend that educators can do the following when they get a positive detection: Ask students to demonstrate their understanding in a controlled environment, whether that is through an in-person assessment, or through an editor that can track their edit history (for instance, using our Writing Reports through Google Docs). Check out our list of several recommendations on types of assignments that are difficult to solve with AI.

Ask the student if they can produce artifacts of their writing process, whether it is drafts, revision histories, or brainstorming notes. For example, if the editor they used to write the text has an edit history (such as Google Docs), and it was typed out with several edits over a reasonable period of time, it is likely the student work is authentic. You can use GPTZero's Writing Reports to replay the student's writing process, and view signals that indicate the authenticity of the work.

See if there is a history of AI-generated text in the student's work. We recommend looking for a long-term pattern of AI use, as opposed to a single instance, in order to determine whether the student is using AI.