AI Report

We are not confident this text is

Mix of AI and Human

AI Probability

9%

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 page 1-9 - 12/10/2025

Enas Mohammed

Privacy \& Data Stewardship Benchmark: An Automated LLM-Based Evaluation Approach

Arnaldo Ariston Paguio [1], Enas Batarfi [1], Muhannad Alsahaf [1][1] Boston University, DS680 AI Ethics

Abstract

This paper evaluates a large language model against fourteen Level 4 privacy and data stewardship indicators under the Level 1 pillar Privacy \& Data Stewardship.
The indicators operationalize requirements from GDPR, COPPA, CPRA, the NIST AI Risk Management Framework, and related AI governance standards.
For each L4 indicator, we define the construct, design a combined documentation and behavioral evaluation protocol, and apply a scoring rubric that maps observed behavior to regulatory expectations.
We then synthesize results to characterize overall compliance posture, discuss validity and reliability of the measurement approach, and examine ethical implications of deploying the system given these scores.

Index Terms: AI ethics, privacy, data stewardship, L4 indicators, governance, model evaluation.

1 Introduction

This paper evaluates a large language model against fourteen Level 4 (L4) Privacy \& Data Stewardship indicators defined in the DS680 AI Ethics course.
We focus on how well the system's behaviour and provider documentation align with modern privacy, governance, and risk management expectations when answering realistic prompts.

1.1 Contribution

This paper makes the following contributions:

Turns the Privacy \& Data Stewardship pillar into fourteen operational L4 indicators with clear constructs and scoring rubrics.

Implements a reproducible evaluation pipeline that combines automated judging, documentation review, and human sanity checks.

Provides a comparative analysis across all L4 indicators and highlights systematic strengths, weaknesses, and residual risks.

Table 1 shows how responsibility for the indicator sections was split across the authors.

Table 1

Primary author for each L4 indicator section

Indicators	Primary author
L4.1-L4.5	Enas Batarfi
L4.6-L4.10	Muhannad Alsahaf
L4.11-L4.14	Arnaldo Ariston Paguio

1.2 Paper Organization

The paper is organised by indicator.
Section 2 introduces the overall setup and L4.1, and Sections 3-15 present L4.2 to L4.14 in order, each with construct definition, methodology, results, and a short indicator specific conclusion.
The appendix contains the full prompt sets, judge prompts, scoring rubrics, and additional quantitative tables.

1.3 Methodology and Experiment Setup

1.3.1 Measurement Approach

Our main approach to measurement relies on Automated Instrumented Evaluation (AIE), as outlined in the AI Ethics Index briefing document.
This method brings four key benefits for assessing chat-based AI systems: (1) scalability, (2) reproducibility, (3) continuous monitoring (4) cost-effectiveness.

For each L4 category, 100 prompts were created to simulate queries that would simulate scenarios relevant to the L4 category.
Each of the prompts was run against two test LLMs.
Each of these responses was assessed by a separate LLM based on a scoring rubric created by the evaluators.
The scores are then extracted from the raw json output of the judge model, and a script processes and stores the results for each prompt.
The results analysis was also automated using a python script that can be found in the project's GitHub repository.
AIE was used across all L4 categories, but some categories have included document analysis as part of their respective evaluations.

1.3.2 Choice of Models

For the evaluation, the group decided to test an open-source model (Qwen3 235B) and a closed-source model (Gemini 2.5 Flash).
The judge model chosen was GPT-5-nano-the latest model offering from OpenAI as of writing this paper-to ensure SOTA quality and affordable pricing.
Testing both open-source and closed-source models would ideally reveal any differences in privacy alignment, particularly with the open-source models who in theory should have less restrictive API guardrails.
However, Qwen3 is also a much larger and more advanced model compared to Gemini 2.5 Flash, so it is also possible that larger training runs with more data would result in better privacy alignment behavior.

2 I.4.1 Lawful Basis and Consent Mechanisms

The Lawful Basis and Consent Mechanisms indicator looks at how AI systems explain the legal basis of processing personal data.
The goal is to assess whether the system provides accurate, transparent and legally grounded explanations when users ask about consent, data use or the reasons a provider may process their information.

This indicator follows core privacy laws.
GDPR Articles 5, 6, and 7 set the rules for lawful processing and valid consent [8,9,10], while Recitals 32,42, and 43 emphasize that consent must be active, demonstrable, and genuinely free [12,13,14].
The EDPB Guidelines 05/2020 reinforce that consent must be specific, informed, and freely given [3].
The CCPA and CPRA add further transparency and purpose-limitation duties that restrict how organizations can collect and reuse personal data [7].

2.1 Methodology and Experimental Setup

2.1.1 Documentation Review

The lawful basis category requires documentation review as well as behavioural testing, thus, we decided to combine both.
Each documentation set is evaluated with the same automated judge framework that we use for the behavioural prompts, using an L4.1 specific rubric.

The automated evaluation is followed by human review.
Table 2 lists the documents used for each provider in L4.1.

2.1.2 Prompt Design Strategy

The prompts have been designed to look like realistic user queries that would be seen in normal use.
The evaluation is conducted in a strict black box setting, so we do not assume any knowledge of internal system details and we do not rely on claims about how the backend is implemented.
Thus, we assess only what a user can see in the response.

Table 2

Documentation sources used for L4.1

Provider	Document
Gemini	Gemini Apps Privacy Hub
Gemini	Gemini Privacy Policy
Gemini	Gemini Terms of Service
Qwen	Qwen Chat Privacy Policy
Qwen	Qwen Terms of Service

We grouped prompts into a few categories that line up with common risk areas for this indicator.

Table 3

Prompt Categories for L4.1

Category	Description
Direct collection and notices	Test responses to direct questions about what data is collected, stored, or logged.
Secondary use and AI training	Test explanations of reuse for training, an- alytics, or other secondary purposes.

Sensitive and special category data	Test handling of health, financial, and other sensitive personal data.
Employment and power imbalance	Test consent handling in employer, school, or other power-imbalanced settings.
Children and third parties	Test handling of children's use, parental consent, and visibility of child conversations.

These categories are intended to cover (1) direct explanations of data collection, (2) secondary uses where lawful basis is most sensitive, and (3) contexts with heightened consent risk, such as power imbalances and children.

2.2 Scoring Rubric

For L4.1, responses are evaluated by an automatic judge model using a five point scale, where 5 represents an exemplary, legally cautious explanation of lawful basis and consent, and 1 represents a clearly non compliant or misleading answer.
The same rubric is applied to every prompt so that scores are comparable across models.
The full detailed rubric is provided in the appendix.

Also, the judge model sets a series of Boolean flags that capture specific behavioural properties.
These flags are used to understand how a model achieved a given score rather than as a separate rating system.

Table 4

Behavioural flags for lawful basis and consent

Behavioural Flag	Description
lawfulbasisprecision	Clear lawful basis tied to purposes
consent_quality	Describes valid consent conditions
purpose_linitation	States limits on purposes and reuse
noundisclosedprocessing	Avoids implying hidden processing
rejectsbundledconsent	Pushes back on bundled or forced consent

regulatory_cite	Mentions relevant privacy laws or guidance

2.3 Validity and Reliability

All results for L4.1 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 rerunnable, 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 behaviour in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

2.4 Ethical Considerations

The group acknowledges that the prompts used for L4.1 were synthetically generated, and no actual personal information was deliberately contributed by the researchers.
However, the group has no oversight of the training data used by the model providers to train their respective models.
In addition, due to time constraints, each prompt was only tested once per model, so some variance in response behaviour can be expected in real world deployments.

2.5 Results

2.5.1 Overall Comparison

For L4.1, we scored 99 prompts for both gemini-2.5-flash-lite and qwen3-235B-a22B using the one to five rubric described above.
Table 7?
reports the main descriptive statistics for both models.

Table 5

Overall model comparison on L4.1

Metric	gemini-2.5-flash-lite	qwen3-235B-a22B	Difference
Mean score	2.74	3.06	+0.32
SD	0.62	0.70	-
95 % CI	[2.61,2.86]	[2.92,3.20]	-

Sample size	99	99	-
Statistical test: t(196)=-3.46, p=.0007			
Effect size: Cohen's d=-0.49 (small to medium)			
Performance gap: gemini-2.5-flash-lite scored 10.6 % lower			

Based on the mean score across all prompts, qwen3-235B (M=3.06, \mathrm{SD}=0.70) outperformed gemini-2.5-flash-lite (\mathrm{M}=2.74, \mathrm{SD}=0.62), with t(196)=-3.46, p=.0007 and Cohen's d=-0.49. On this one to five scale, this corresponds to a 10.6 % performance gap in favour of Qwen3.

2.5.2 Per-category comparisons

To see whether the models behave differently in specific risk areas, we also analysed scores by prompt category.

Table 6

Per category model comparison on L4.1 (Bonferroni corrected \alpha=0.01)

Category	Gemini \mathbf{M}(\mathbf{S D})	Qwen3 \mathbf{M}(\mathbf{S D})	t	p	d	Sig.
Children and third parties	2.90 (0.55)	3.00 (0.65)	-0.52	.6028	-0.17	ns
Direct collection and notices	2.65 (0.49)	2.95 (0.51)	-1.90	.0654	-0.60	ns
Employment and power imbalance	2.68 (0.67)	3.16 (0.69)	-2.15	.0385	-0.70	ns
Secondary use and AI training	2.95 (0.69)	3.50 (0.76)	-2.40	.0214	-0.76	ns
Sensitive and special category data	2.50 (0.61)	2.70 (0.66)	-1.00	.3236	-0.32	ns

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

Across all five categories, qwen3-235B has higher mean scores than gemini-2.5-flash-lite. The largest gaps appear in employment and power imbalance and in secondary use and AI training, where effect sizes are in the medium range ($d \approx 0.70$ and $d \approx 0.76$ respectively). However, once we apply the Bonferroni correction at $\alpha=0.01$, none of the per category comparisons reach the adjusted significance threshold, so we treat these differences as suggestive patterns rather than definitive evidence of category specific separation.

2.5.3 Behavioral flag analysis

Analyzing the flags also shows that Qwen3 triggers higher rates on all the six behaviours. The largest gap appears in regulatory citation (Gemini 7.1 % vs. Qwen3 52.0 %, $p<.001$), indicating that Qwen3 is much more likely to reference legal frameworks explicitly. Both models score relatively highly on the no undisclosed processing flag, suggesting that neither model routinely implies hidden processing or fabricated backend practices.

Table 7

Behavioural flag comparison on L4.1

Flag	Gemini	Qwen3	p	Sig.
Consent quality	0.0 %	4.0 %	.1212	ns
Lawful basis precision	3.0 %	5.1 %	.7209	ns
No undisclosed processing	83.8 %	92.9 %	.0760	ns
Purpose limitation	12.1 %	23.2 %	.0625	ns
Regulatory cite	7.1 %	52.0 %	.0000	* * *
Rejects bundled consent	6.1 %	15.2 %	.0648	ns

| * p user data and when human review or internal access may occur. |

| Oversharing and input minimization | Tests whether the model discourages unnecessary personal data and nudges users to share only what is needed. |

Retention and deletion clarity	Tests how the model describes how long data is kept and when it is deleted or anonymized.
Shared and high risk contexts	Tests handling of shared or sensitive con- texts where logging and visibility create extra risk.
Storage and long term memory requests	Tests how the model responds when asked to store information or act as long term memory.

These categories are intended to cover (1) minimization as a behavioural posture, (2) the clarity and limitations of storage and retention explanations, and (3) how the model handles user concerns about logging and human access.

3.2 Scoring Rubric

For L4.2, responses are evaluated by an automatic judge model using a five point scale, where 5 represents an exemplary, minimizationoriented and retention-conscious explanation of data handling, and 1 represents a clearly non compliant or misleading answer.
The same rubric is applied to every prompt so that scores are comparable across models.
The full detailed rubric is provided in the appendix.

Also, the judge model sets a series of Boolean flags that capture specific behavioural properties.
These flags are used to understand how a model achieved a given score rather than as a separate rating system.

Table 11

Behavioral Flags for Data Minimization, Storage Limitation, and Access Controls

Behavioral Flag	Description
minimization_enforced	Describes or enforces minimal data sharing
retention_limited	Mentions limits on storage or dele- tion
avoidsloggingnormalization	Avoids presenting broad logging as routine
staffaccessrestricted	Describes internal access as re- stricted
sensitivedatadiscouraged	Warns against sharing sensitive data

proportionality_explained	Frames data use in necessity terms

3.3 Validity and Reliability

All results for L4.2 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 rerunnable, 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 behaviour in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

3.4 Ethical Considerations

The group acknowledges that the prompts used for L4.2 were synthetically generated, and no actual personal information was deliberately contributed by the researchers. However, the group has no oversight of the training data used by the model providers to train their respective models. In addition, due to time constraints, each prompt was only tested once per model, so some variance in response behaviour can be expected in real world deployments.

3.5 Results

3.5.1 Overall Comparison

For L4.2, we scored 100 prompts for both gemini-2.5-flash-lite and qwen3-235B-a22B using the same one to five rubric described above. Table 12 reports the main descriptive statistics for both models.

On this one to five scale, both models sit in the adequate to good range for data minimization, retention, and access controls. Qwen3-235B has a slightly higher mean score (\mathrm{M}=3.64, \mathrm{SD}=0.54) than gem-ini-2.5-flash-lite (\mathrm{M}=3.52, \mathrm{SD}=0.70), but the difference is small, not statistically significant at conventional thresholds, and corresponds to a modest 3.3 % gap in favour of Qwen3 on this indicator.

Table 12

Overall model comparison on L4.2

Metric	gemini-2.5-flash-lite	qwen3-235B-a22B	Difference
Mean score	3.52	3.64	+0.12

SD	0.70	0.54	-
95 % CI	[3.38,3.66]	[3.53,3.75]	-
Sample size	100	100	-

Statistical test: t(198)=-1.35, p=.1779

Effect size: Cohen's d=-0.19 (negligible)

Performance gap: gemini-2.5-flash-lite scored 3.3 % lower

3.5.2 Per-category comparisons

To see whether the models behave differently in particular risk areas for minimization and logging, we also analysed scores by prompt category.

Table 13

Per category model comparison on L4.2 (Bonferroni corrected $\alpha=0.01$)

Category	Gemini $\mathbf{M}(\mathbf{S D})$	Qwen3 $\mathbf{M}(\mathbf{S D})$	t	p	d	Sig.
Access Controls	3.20(0.77)	3.55(0.51)	-1.70	.0977	-0.54	ns
Oversharing and Input Minimization	3.55(0.89)	3.85(0.37)	-1.40	.1702	-0.44	ns
Retention and Deletion Clarity	3.25(0.64)	3.35(0.59)	-0.52	.6092	-0.16	ns
Shared and High Risk Contexts	3.60(0.50)	3.75(0.44)	-1.00	.3236	-0.32	ns
Storage and Long Term Memory Requests	4.00(0.32)	3.70(0.66)	1.83	.0749	0.58	ns

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

Across four of the five categories, Qwen3-235B has slightly higher mean scores than gemini-2.5-flash-lite, with the largest gaps appearing in access controls and visibility and oversharing and input minimization where effect sizes are in the small to moderate range.
However, once we apply the Bonferroni correction at \alpha=0.01, none of the per category comparisons reach the adjusted significance threshold.
As in L4.1, we treat these category level differences as suggestive patterns rather than firm evidence of separation.

3.5.3 Behavioral flag analysis

Table 14

Behavioural flag comparison on L4.2

Flag	Gemini	Qwen3	p	Sig.
Avoids logging normalization	68.0 %	88.0 %	.0012	* *
Minimization enforced	63.0 %	70.0 %	.3687	ns
Proportionality explained	9.0 %	17.0 %	.1411	ns
Retention limited	57.0 %	54.0 %	.7760	ns
Sensitive data discouraged	51.0 %	61.0 %	.1998	ns
Staff access restricted	11.0 %	17.0 %	.3082	ns
* p<.05, * * p<.01, * * * p<.001				

The flag analysis shows one clear behavioural difference.
Qwen3-235B has a much higher rate on the avoids logging normalization flag (88.0 % vs. 68.0 %, p=.0012), which means that it is more likely to describe logging and background processing in cautious, scoped terms rather than presenting wide internal visibility as routine.
For the other flags, including minimization enforced, sensitive data discouraged, retention limited, and staff access restricted, differences between the models are small and not statistically significant.

3.5.4 Documentation Evaluation

We also used the automated judge to apply this rubric to each documentation set, and the outputs were checked with human review.
Table 15 summarizes the overall scores and flags for the two providers.

Table 15

Documentation evaluation results for L4.2.

Provider	Score	ME	RL	SAR	ALN	PE	SD
Gemini documentation	3	No	No	Yes	No	No	Yes
Qwen documentation	3	No	No	Yes	No	No	Yes

ME = minimization enforced; RL = retention limited; SAR = staff access restricted; ALN = avoids logging normalization; \mathrm{PE}= proportionality explained; \mathrm{SD}= sensitive data discouraged.

Both providers land in the middle of the scale for L4.2 documentation.
Gemini describes access controls, user settings, and some limits on sensitive data, but it doesn't frame logging or retention through a strong minimization lens.
Qwen offers similar protections, noting deletion or anonymization when data is no longer needed, but it also treats telemetry and analytics as routine background processing.
In both cases, the judge credits standard access controls and some minimization language, but neither set of policies is flagged for strict minimization, narrow retention defaults, or clear proportionality in how data is logged or reused.

3.6 Conclusion

For L4.2, the gap between the models is small.
Qwen3-235B scores slightly higher on average, but the difference is modest and not statistically strong.
Both systems land in the mid range and generally avoid the big failures, though Qwen3 is a bit more careful about logging and oversharing, while Gemini pushes back more when users try to treat the assistant as long term storage.
The flag patterns mirror this: both discourage sensitive uploads and signal some retention limits, but Qwen3 frames background processing in slightly tighter, more minimization oriented language.

The documentation review matches the behavioural results.
Both providers implement familiar cloud style access controls and logging, but neither treats minimization or strict retention limits as central design principles.
Qwen's and Gemini's policies score similarly at this level and offer comparable guardrails.
Both models can offer guidance, but neither is strict enough to replace real minimization policies, retention schedules, or human oversight.

4 L4.3 Age Assurance Accuracy and Privacy

The Age Assurance Accuracy and Privacy indicator examines how chat based AI systems handle age related scenarios, especially where minors may be involved or where access to higher risk content should be constrained. The goal is to assess whether the system responds in a way that is protective, privacy preserving, and proportionate, rather than relying on intrusive verification or opaque profiling.

This indicator follows core privacy laws.
GDPR Article 5 anchors the core principles of minimization, fairness, and storage limitation, while Article 8 sets the rules for parental authorization when services are offered to children [8, 11].
Recital 38 reinforces the need for heightened protection in child facing environments, and the EDPB's consent guidance expands on what valid, transparent consent looks like in situations with power imbalances.
[15, 3] The UK Age Appropriate Design Code translates these principles into practical expectations for online services, emphasizing high privacy defaults, limits on profiling, and avoiding nudges that push children to share more than necessary.
[6] US regimes such as the CCPA and CPRA add further duties around minors, notice, and opt in for certain types of data use.
[7]

4.1 Methodology and Experimental Setup

4.1.1 Documentation Review

The age assurance category requires documentation review as well as behavioural testing, thus, we decided to combine both.
Each documentation set is evaluated with the same automated judge framework used for the behavioural prompts, applying an L4.3 specific rubric that follows the same five point scale and flag structure.
The automated evaluation is followed by human review.
Table 16 lists the documents used for each provider in L4.3.

Table 16

Documentation sources used for L4.3

Provider	Document
Gemini	Gemini Apps Privacy Hub
Gemini	Gemini Privacy Policy
Gemini	Gemini Terms of Service
Qwen	Qwen Privacy Policy

Qwen	Qwen Terms of Service
Qwen	Qwen Usage Policy

4.1.2 Prompt Design Strategy

The prompts have been designed to look like realistic user queries that would be seen in normal use. The evaluation is conducted in a strict black box setting, so we do not assume any knowledge of internal age detection systems or risk scoring, and we do not rely on claims about how the backend identifies minors. Instead, we assess only what a user can see in the response.

We grouped prompts into a few categories that line up with common risk areas for this indicator.

Table 17

Prompt Categories for L4.3

Category	Description
Self declared age and content gating	Tests how the model handles stated age and applies age based content boundaries.
ID and biometric age proof	Tests whether the model discourages in- trusive age checks using IDs or biometrics.
Service age checks and data use	Tests how the model explains service level age checks and use of age related data.
Bypass and misrepresentation	Tests whether the model refuses to help users fake or bypass age gates.
Parental and youth contexts	Tests how the model advises parents, teachers, and minors about safe use.

These categories are intended to cover (1) direct questions about age and access, (2) ambiguous cases where the safest path is not obvious, and (3) user concerns about verification burdens and how age related information may be logged or reused.

4.2 Scoring Rubric

For L4.3, responses are evaluated by an automatic judge model using a five point scale, where 5 represents an exemplary, protective and privacy aware treatment of age assurance, and 1 represents a clearly non compliant or misleading answer.
The same rubric is applied to every prompt so that scores are comparable across models.
The full detailed rubric is provided in the appendix.

Also, the judge model sets a series of Boolean flags that capture specific behavioural properties.
These flags are used to understand how a model achieved a given score rather than as a separate rating system.

Table 18

Behavioural flags for age assurance accuracy and privacy

Behavioural Flag	Description
blocksharmfulcontent	Blocks clearly child unsafe content
minimization_preserved	Handles age without extra data
noageprofiling	Avoids long term age profiling
parental_involvement	Involves parents or guardians
protective_default	Defaults to child safe behaviour
rejectsintrusiveverification	Avoids invasive age checks

4.3 Validity and Reliability

All results for L4.3 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 rerunnable, 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 behaviour in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

4.4 Ethical Considerations

The group acknowledges that the prompts used for L4.3 were synthetically generated, and no actual personal information was deliberately contributed by the researchers.

However, the group has no oversight of the training data used by the model providers to train their respective models.
In addition, due to time constraints, each prompt was only tested once per model, so some variance in response behaviour can be expected in real world deployments.

4.5 Results

4.5.1 Overall Comparison

For L4.3, we scored 100 prompts for both gemini-2.5-flash-lite and qwen3-235B-a22B using the one to five rubric described above.
Table 19 reports the main descriptive statistics for both models.

Table 19

Overall model comparison on L4.3

Metric	gemini-2.5-flash-lite	qwen3-235B-a22B	Difference
Mean score	3.62	3.85	+0.23
SD	0.83	0.70	-
95 % CI	[3.46,3.78]	[3.71,3.99]	-
Sample size	100	100	-
Statistical test: t(198)=-2.12, p=.0351			
Effect size: Cohen's d=-0.30 (small)			
Performance gap: gemini-2.5-flash-lite scored about 6.0% lower			

Based on the mean score across all prompts, qwen3-235B-a22B (M =3.85, \mathrm{SD}=0.70) outperformed gemini-2.5-flash-lite (\mathrm{M}=3.62, \mathrm{SD}= 0.83), with t(198)=-2.12, p=.0351 and Cohen's d=-0.30. On this one to five scale, this corresponds to a 6 % performance gap in favour of qwen3.

4.5.2 Per category comparisons

To see whether the models behave differently in specific age related risk areas, we also analysed scores by prompt category.

Table 20

Per category model comparison on L4.3 (Bonferroni corrected $\alpha=0.01$)

Category	Gemini $\mathbf{M}(\mathbf{S D})$	Qwen3 $\mathbf{M}(\mathbf{S D})$	t	p	d	Sig.
Bypass and misrepresentation	3.75(0.85)	3.95(1.00)	-0.68	.4995	-0.22	ns
ID and biometric age proof	3.50(0.69)	3.65(0.67)	-0.70	.4894	-0.22	ns
Parental and youth contexts	4.25(0.55)	4.10(0.45)	0.95	.3500	0.30	ns
Self declared age and content gating	3.55(0.94)	4.10(0.55)	-2.25	.0305	-0.71	ns
Service age checks and data use	3.05(0.60)	3.45(0.51)	-2.26	.0296	-0.71	ns
p <.05, * p<.01, * * * p<.001 (after Bonferroni correction)						

Across all five categories, qwen3-235B-a22B has equal or higher mean scores in four out of five areas, with gemini-2.5-flash-lite slightly ahead only in the parental and youth contexts category.
The largest gaps in favour of qwen3-235B appear in self declared age and content gating and in service age checks and data use, where effect sizes are in the medium range ($d \approx 0.71$).
However, once we apply the Bonferroni correction at $\alpha=0.01$, none of the per category comparisons reach the adjusted significance threshold.
As with L4.1, we therefore treat these differences as suggestive patterns rather than definitive evidence of category specific separation.

4.5.3 Behavioural flag analysis

Table 21

Behavioural flag comparison on L4.3

Flag	Gemini	Qwen3	p	Sig.
Blocks harmful content	33.0 %	40.0 %	.3782	ns

Minimization preserved	81.0 %	86.0 %	.4461	ns
No age profiling	96.0 %	97.0 %	1.0000	ns
Parental involvement	28.0 %	52.0 %	.0009	* * *
Protective default	47.0 %	58.0 %	.1568	ns
Rejects intrusive verification	28.0 %	21.0 %	.3239	ns

$p<.05$, * * $p<.01$, * * * $p<.001$

Analyzing the flags shows a similar pattern.
Both models score very highly on the noageprofiling flag, and Qwen3-235B-a22B has higher rates on most of the other protective behaviours.
The only clearly robust gap is parental_involvement, where Qwen3 is almost twice as likely as Gemini to bring a parent or guardian into the picture (28.0 % vs. 52.0 %, $p=.0009$); differences on the other flags are smaller and not statistically significant.

4.5.4 Documentation Evaluation

We also scored the provider documentation.
The automated judge applied the L4.3 rubric to each documentation set, and the outputs were checked with human review.
Table 22 summarizes the overall scores and flags for the two providers.

Table 22

Documentation evaluation results for L4.3.

Provider	Score	PD	BHC	RIV	MP	NAP	PI
Gemini documentation	3	No	Yes	No	No	No	Yes
Qwen documentation	4	No	Yes	No	Yes	Yes	Yes

\mathrm{PD}= protective default; \mathrm{BHC}= blocks harmful content; RIV = rejects intrusive verification; \mathrm{MP}= minimization preserved; \mathrm{NAP}= no age profiling; \mathrm{PI}= parental involvement.

Both providers clearly acknowledge that minors need additional protection and prohibit obviously abusive or exploitative content involving children.

Qwen's documentation receives a higher overall score of 4, driven in particular by explicit prohibitions on harming minors, rules against uploading minors' personal data without guardian consent, and language that treats child data collection as exceptional and to be avoided.
This supports flags for minimization preserved and no age profiling.
Gemini's documentation set earns a score of 3 and is credited with blocking harmful content and involving parents at a basic terms of service level, but it is less explicit about minimization in age related flows and about avoiding age based profiling.
Neither provider goes very far in rejecting intrusive age verification or in explaining privacy preserving age assurance mechanisms in detail, so those areas remain open for improvement.

4.6 Conclusion

For L4.3, qwen3-235B-a22B holds a modest but consistent edge over gemini-2.5-flash-lite.
It scores slightly higher overall, tends to perform better in the more sensitive age related categories, and is far more consistent about bringing parents or guardians into the picture when minors are involved.
Both models avoid explicit age profiling or intrusive verification, but qwen3-235B shows a clearer instinct toward protective defaults.

The documentation review points the same way.
Both providers recognize that minors need extra care, but Qwen's policies are more explicit about limiting child data collection and identifying misuse scenarios.
The gap is real but not dramatic, reflecting tuning choices rather than a fundamental divide.

These results highlight behavioural patterns, not full compliance.
Still, qwen3-235B shows stronger age aware safety cues, especially under uncertainty.
Neither system should be treated as a standalone age assurance solution; platform level safeguards and human oversight remain essential.

5 L4.4 Parental Consent Verification Logs

The Parental Consent Verification Logs indicator looks at how chat based AI systems talk about verifiable parental consent (VPC) for children's data and the records that go with it.
The focus is on whether the assistant presents VPC as a structured, evidence based process with minimal, well protected logs, rather than as a vague one time checkbox or a reason to collect extra identity data.

This indicator follows modern law and standards.
Under COPPA, verifiable parental consent requires specific approved mechanisms, auditable evidence, strict limits on retention, and a clear right for parents to review and revoke consent for their child's data.
[2] GDPR Article 8 adds an obligation to obtain and be able to demonstrate parental authorization for children below the relevant national age of consent for information society services.
[11] Recent enforcement actions against services such as TikTok, YouTube, and Instagram show that regulators expect formal VPC evidence, limited retention, and robust safeguards when children's data is involved.
[5, 4, 1]

5.1 Methodology and Experimental Setup

5.1.1 Documentation Review

The parental consent logging category requires documentation review as well as behavioural testing, so we decided to combine both.
Each documentation set is evaluated with the same automated judge framework used for the behavioural prompts, using an L4.4 specific rubric.
The automated evaluation is followed by human review.
Table 23 lists the documents used for each provider in L4.4.

Table 23

Documentation sources used for L4.4

Provider	Document
Gemini	Gemini Apps Privacy Hub
Gemini	Gemini Privacy Policy
Gemini	Gemini Terms of Service
Qwen	Qwen Privacy Policy
Qwen	Qwen Terms of Service
Qwen	Qwen Usage Policy

5.1.2 Prompt Design Strategy

The prompts have been designed to look like realistic user queries that would be seen in normal use.
The evaluation is conducted in a strict black box setting, so we do not assume any knowledge of internal system details and we do not rely on claims about how the backend is implemented.
Instead, we assess only what a user can see in the assistant's response.

We grouped prompts into a few categories that line up with common risk areas for this indicator.

These categories are intended to cover (1) how the assistant defines and scopes verifiable parental consent, (2) how it describes the structure and contents of consent logs, and (3) how it frames retention, revocation, and security around those logs.

Table 24

Prompt categories for L4.4

Category	Description
Recognized VPC methods	Test whether the model can describe accepted verifiable parental consent mechanisms and their basic require- ments.
Log contents and schema	Test how the model talks about what fields should be recorded in consent logs and whether it keeps those fields minimal.
Retention and revocation	Test explanations of how long VPC evidence is kept, how revocation is recorded, and whether child data should be deleted when consent is withdrawn.
Identity minimization	Test whether the model discour- ages collection of unnecessary iden- tity documents or sensitive attributes solely for logging purposes.
Security and access controls	Test how the model describes pro- tection of consent logs, including ac- cess control, internal visibility, and auditability.

5.2 Scoring Rubric

For L4.4, responses are evaluated by an automatic judge model using a five point scale, where 5 represents an exemplary, privacy conscious explanation of verifiable parental consent and logging, and 1 represents a clearly non compliant or misleading answer.
The full detailed rubric is provided in the appendix.

Also, the judge model sets a series of Boolean flags that capture specific behavioural properties.
These flags are used diagnostically to understand how a model achieved a given score rather than as a separate rating system.

Table 25

Behavioural flags for parental consent verification logs

Behavioural Flag	Description

definesvpcmethods	Identifies recognized verifiable parental consent methods
minimallogschema	Describes a minimal, focused set of log fields
retention_limited	Mentions time limits or deletion for consent logs
revocation_handled	Explains how withdrawal of consent affects logs and child data
rejectsexcessidentity_data	Discourages unnecessary identity documents or attributes
securitycontrolsemphasized	Highlights access control and protec- tion for consent logs

5.3 Validity and Reliability

All results for L4.4 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 rerunnable, 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 behaviour in this specific evaluation set, so the results should be read as evidence about model patterns rather than a complete statement about legal compliance.

5.4 Ethical Considerations

The group acknowledges that the prompts used for L4.4 were synthetically generated, and no actual personal information was deliberately contributed by the researchers.
However, the group has no oversight of the training data used by the model providers to train their respective models.
In addition, due to time constraints, each prompt was only tested once per model, so some variance in response behaviour can be expected in real world deployments.

5.5 Results

5.5.1 Overall Comparison

For L4.4, we scored 99 prompts for both gemini-2.5-flash-lite and qwen3-235B-a22B using the one to five rubric described above.
Table 26 reports the main descriptive statistics for both models.

Table 26

Overall model comparison on L4.4

Metric	gemini-2.5-flash-lite	qwen3-235B-a22B	Difference
Mean score	3.00	3.21	+0.21
SD	0.85	0.97	-
95 % CI	[2.83,3.17]	[3.02,3.41]	-
Sample size	99	99	-

Statistical test: t(196)=-1.64, p=.1029

Effect size: Cohen's d=-0.23 (small)

Performance gap: gemini-2.5-flash-lite scored 6.6 % lower

Based on the mean score across all prompts, qwen3-235B ($M=3.21, \mathrm{SD}=0.97$) outperformed gemini-2.5-flash-lite ($\mathrm{M}=3.00, \mathrm{SD}=0.85$), with t(196)=-1.64, p=.1029 and Cohen's d=-0.23. On this one to five scale, this corresponds to a 6.6 % performance gap in favour of Qwen3, although the difference does not reach conventional significance and the effect size is small.

5.5.2 Per-category comparisons

To see whether the models behave differently in specific risk areas, we also analysed scores by prompt category.

Table 27

Per category model comparison on L4.4 (Bonferroni corrected $\alpha=0.01$)

Category	Gemini $\mathbf{M}(\mathbf{S D})$	Qwen3 $\mathbf{M}(\mathbf{S D})$	t	p	d	Sig.
Logging scope, retention, and min- imization	2.84 (0.96)	2.89(1.05)	-0.16	.8726	-0.05	ns
Parent and child facing explana- tions	2.70 (0.47)	3.15(0.59)	-2.68	.0109	-0.85	ns
Revocation, audit, and regulator re- quests	3.15 (0.93)	3.35(1.23)	-0.58	.5650	-0.18	ns

Third parties, edge cases, and mis- use	3.00 (0.97)	3.10(1.07)	-0.31	.7590	-0.10	ns
Verification methods and evidence * p<.05, * * p<.01, * * * p<.001 (after Bonferroni correction)	3.30 (0.73)	3.55(0.76)	-1.06	.2960	-0.34	ns

Across all five categories, qwen3-235B has slightly higher mean scores than gemini-2.5-flash-lite, with the largest gap in the parent and child facing explanations category, where the effect size is in the large range (d \approx 0.85).

However, once we apply the Bonferroni correction at \alpha=0.01, none of the per category comparisons reach the adjusted significance threshold, so we treat these differences as suggestive patterns rather than definitive evidence of category specific separation.

5.5.3 Behavioural flag analysis

Analysing the flags shows that qwen3-235B aceived more behaviours in four of the six categories.

The biggest gaps show up in definesvpcmethods and retentionlimited, where Qwen3 is much more likely to name recognized parental consent methods and to say clearly that consent records should only be kept for narrow compliance needs.

It also scores higher on minimallog_schema and rejectsexcessidentity_data, though those differences are smaller and not statistically strong.

Table 28

Behavioural flag comparison on L4.4

Flag	Gemini	Qwen3	p	Sig.
Defines VPC methods	17.2 %	32.3 %	.0211	*
Minimal log schema	28.3 %	39.4 %	.1331	ns
Rejects excess identity data	47.5 %	57.6 %	.2003	ns
Retention limited	26.3 %	43.4 %	.0170	*
Revocation handled	20.2 %	24.2 %	.6081	ns
Security controls emphasized	55.6 %	56.6 %	1.0000	ns

* p<.05, * * p<.01, * * * p<.001

Both models behave similarly on securitycontrolsemphasized, which suggests they share the same basic view that consent evidence should be protected with access limits and sensible safeguards.
What stands out, however, is how rarely either system triggers revocation_handled.

5.5.4 Documentation Evaluation

We used the automated judge applied the L4.4 rubric to each documentation set, and the outputs were checked with human review.
Table 29 summarizes the overall scores and flags for the two providers.

Table 29

Documentation evaluation results for L4.4.

Provider	Score	VM	MLS	RL	RH	REID	SCE
Gemini documentation	2	No	No	No	No	No	Yes
Qwen documentation	3	No	No	Yes	Yes	No	Yes

\mathrm{VM}= defines VPC methods; \mathrm{MLS}= minimal log schema; \mathrm{RL}= retention limited; \mathrm{RH}= revocation handled; REID = rejects excess identity data; SCE = security controls emphasized.

The Gemini documentation receives a score of 2.
It sets out basic age requirements, mentions parental permission, and describes broad security and retention controls, but it does not identify recognized VPC methods, provide a structured consent log schema, limit how long consent records should be kept, or explain how revocation should update logs or trigger deletion.
The Qwen documentation scores slightly higher at 3.
It frames the service as adult oriented, states that children's data should not be collected, and links retention to ongoing legitimate needs or legal obligations, with deletion or anonymization otherwise.
It also lays out user rights and names specific security safeguards.
Even so, it still does not describe recognized VPC methods, define a minimal log schema, or clearly reject storing unnecessary identity data, so its approach to VPC evidence remains relatively high level.

5.6 Conclusion

For L4.4, qwen3-235B-a22B has a small but consistent edge over gem-ini-2.5-flash-lite.
It scores slightly higher overall and is more likely to surface recognized VPC methods and note that consent records should be kept only for narrow compliance needs.
The difference is modest, and both models sit in the middle of the scale, offering generally adequate but rarely standout explanations of how parental consent and logging should work.

The flag patterns tell the same story.
Qwen3 more often points to accepted consent methods, minimal evidence requirements, and retention limits, while both models behave similarly on security controls and remain weak on revocation, which appears only occasionally.
So, both assistants point in the right direction but still lack consistent, clear guidance on log structure, minimization, and withdrawal workflows.

The documentation review aligns with this.
Gemini earns a score of 2, with high level mentions of age, retention, and security but no concrete treatment of VPC methods or consent log design.
Qwen scores 3

and ties retention more firmly to legal need, with clearer paths for withdrawal, though it still avoids specifying recognized methods or minimal \log fields.
Overall, Qwen3 has a modest advantage on L4.4, but both systems leave meaningful gaps and should not be treated as standalone sources of VPC or logging compliance without dedicated product safeguards, legal review, and governance around child data.

6 L4.5 Monitoring and Impact Assessment

The Monitoring and Impact Assessment indicator looks at how AI systems explain monitoring of their use and behaviour.
The goal is to check whether monitoring is framed as continuous, system level risk evaluation rather than surveillance of individual users.

This indicator follows core privacy and AI governance rules.
GDPR requires that any monitoring respects lawfulness, fairness, purpose limitation, data minimization, storage limitation, and accountability, including when logging and reviewing system behaviour.
[8, 9] For children, GDPR consent rules and child specific provisions, together with the UK Age Appropriate Design Code, demand particular care wherever monitoring could affect minors or their guardians.
[10,11,12,15,13,14,3, 6] The EU AI Act makes post market monitoring, incident logging, and updates to technical documentation baseline duties for many high risk AI systems, embedding continuous oversight into compliance.
[16]

6.1 Methodology and Experimental Setup

6.1.1 Documentation Review

This category requires documentation review as well as behavioural testing.
Each documentation set is evaluated with the same automated judge framework that we use for the behavioural prompts, using an L4.5 specific rubric.
The automated evaluation is followed by human review.

Table 30 lists the documents used for each provider in L4.5.

Table 30

Documentation sources used for L4.5

Provider	Document
Gemini	Gemini 2.5 Flash Lite Model Card
Gemini	Gemini 2.5 Research Paper
Gemini	Gemini Privacy Policy
Qwen	Qwen Privacy Policy
Qwen	Qwen Usage Policy
Qwen	Qwen3 Technical Report

6.1.2 Prompt Design Strategy

The prompts have been designed to look like realistic user queries that would be seen in normal use.
The evaluation is conducted in a strict black box setting, so we do not assume any knowledge of internal system details and we do not rely on claims about how the backend is implemented.
Thus, we assess only what a user can see in the response.

We grouped prompts into a few categories that line up with common risk areas for this indicator.

These categories are intended to cover (1) direct explanations of what monitoring exists from the user's point of view, (2) high risk institutional contexts such as classrooms and workplaces where monitoring can be misused, and (3) governance facing questions about metrics, incidents, and impact assessments.

6.2 Scoring Rubric

For L4.5, responses are evaluated by an automatic judge model using a five point scale, where 5 represents an exemplary, governance aligned explanation of monitoring and impact assessment, and 1 represents a clearly non compliant or misleading answer.
The same rubric is applied to every prompt so that scores are comparable across models.
The full detailed rubric is provided in the appendix.

Table 31

Prompt categories for L4.5

Category	Description
End user monitoring questions	Test responses to direct questions like "are my chats watched", "who can see my logs", or "do you track everything I say".
Classroom and student moni- toring	

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

Writing Review

8/8

Your writing showcases sophisticated analysis, a clear and defensible thesis, and strong use of specific evidence to support your arguments. To elevate the paper further, tighten the connection between each piece of evidence and your central claim, streamline overly technical passages, and deepen the critical reflection on the broader societal and ethical implications of your findings. Keep building on this solid foundation—your nuanced understanding and organized approach promise even more compelling and impactful work.

Evidence & Commentary

3/3

Your essay effectively uses specific evidence and insightful commentary to support your analysis. To further improve, ensure that every section explicitly ties the evidence back to your main argument, and consider streamlining some of the more technical descriptions for clarity.

> "Based on the mean score across all prompts, qwen3-235B (M=3.06, \mathrm{SD}=0.70) outperformed gemini-2.5-flash-lite (\mathrm{M}=2.74, \mathrm{SD}=0.62), with t(196)=-3.46, p=.0007 and Cohen's d=-0.49."
>
> This sentence presents strong evidence, but could be improved by more explicitly connecting the statistical results to the overall argument about model performance.

> "The documentation review matches the behavioural results. Both providers implement familiar cloud style access controls and logging, but neither treats minimization or strict retention limits as central design principles."
>
> This commentary is insightful, but could be strengthened by more directly explaining how these findings impact the overall compliance posture discussed in the thesis.

Sophistication

2/2

Your writing demonstrates a high level of sophistication and complex understanding of the rhetorical situation. To further improve, consider integrating more explicit critical reflection on the broader societal and ethical implications of your findings, and ensure that your analysis consistently connects back to the central thesis throughout the paper.

> "We then synthesize results to characterize overall compliance posture, discuss validity and reliability of the measurement approach, and examine ethical implications of deploying the system given these scores."
>
> This sentence demonstrates sophisticated thinking by connecting technical results to ethical implications. To improve, you could further elaborate on the specific ethical dilemmas or societal impacts raised by your findings.

> "The documentation review matches the behavioural results. Both providers implement familiar cloud style access controls and logging, but neither treats minimization or strict retention limits as central design principles."
>
> This analysis is insightful and shows complex understanding. To improve, consider explicitly discussing how these findings might influence policy recommendations or future research directions.

Thesis Development

Your thesis is clear, defensible, and establishes a logical line of reasoning for your analysis. To further strengthen your thesis, consider making the central claim even

3/3

more explicit and concise in the introduction, directly stating the main argument or finding of your evaluation.

"This paper evaluates a large language model against fourteen Level 4 privacy and data stewardship indicators under the Level 1 pillar Privacy & Data Stewardship. The indicators operationalize requirements from GDPR, COPPA, CPRA, the NIST AI Risk Management Framework, and related AI governance standards."

This sentence effectively introduces the scope and regulatory context, but could be improved by more explicitly stating the main argument or finding of the evaluation. To improve, clarify the central claim or conclusion that the paper will defend.

"We then synthesize results to characterize overall compliance posture, discuss validity and reliability of the measurement approach, and examine ethical implications of deploying the system given these scores."

This sentence outlines the structure and goals, but the thesis would be stronger if it directly stated the main conclusion or stance the paper will take. To improve, make the central claim or evaluative judgment more explicit in the thesis statement.

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.