When Formatting Flips Answers: A Deterministic Atlas of Retokenization Sensitivity in LLM Evaluation

Ali Uyar
Independent Researcher

February 2026

Abstract

Semantics-preserving formatting should not change model decisions, but in multiple-choice QA it often does. We present **Retokenization Invariance Atlas (RIA)**, a deterministic audit that isolates this effect with paired no-truncation/semantics gates, scope-only token-divergence metrics, and mitigation checks with explicit no-harm criteria. Across BoolQ, ARC-Challenge, and an 8-subject MMLU subset (36,820 gated rows, two instruct models), routine formatting edits reach 9.18% maximum flip-rate (Llama3-8B, mmlu_subset, O4). TD is consistently positive but modest ($\rho \in [0.072, 0.206]$, pass 1/6), while ablations expose strong template/placement sensitivity and non-universal canonicalization safety on MMLU. RIA turns retokenization sensitivity into a reproducible reliability axis with complete scripts, logs, figures, and gate evidence.

1 Introduction

Large-language-model evaluations often assume that semantically equivalent formatting variants are behaviorally interchangeable. In practice, small differences in whitespace, punctuation style, and wrapping can alter tokenization and, consequently, model decisions. This is especially important for closed-form multiple-choice evaluations where small score deltas can change rankings, regressions, and deployment decisions.

This paper focuses on *evaluation reliability*, not adversarial attacks. We ask: if content semantics are preserved and truncation is prevented, how much decision instability remains under everyday formatting variants?

We contribute a deterministic audit protocol and an artifact-complete benchmark package:

- **Paired, gate-enforced design.** Every included example passes no-truncation and semantics-preservation checks across clean, perturbed, and canonicalized variants.
- **Scope-local mechanism metric.** We compute TD on perturbation scope only (not whole prompt wrappers), then link TD to flips and margin shifts.
- **Mitigation with safety accounting.** We evaluate canonicalization recovery and explicitly test no-harm on clean inputs.
- **Reviewer-facing robustness ablations.** We include template dependence, placement-seed robustness, and deterministic replay checks.

The resulting picture is mixed but informative: integrity gates pass, while several robustness gates fail under strict precommitted thresholds. Under Path A framing, these are reported as empirical findings, not retroactively redefined criteria.

2 Related Work

Benchmark and QA evaluation datasets. RIA uses established closed-form QA settings: BoolQ [1], ARC-Challenge [2], and MMLU [3]. Our contribution is not a new benchmark dataset, but a reliability layer on top of standard evaluation tasks.

Prompt and formatting sensitivity. Prior literature shows that prompt phrasing and formatting can affect model behavior. RIA narrows this into a deterministic protocol for *semantics-preserving* retokenization variants, with explicit scope controls and reproducibility gates.

Normalization and robustness reporting. Text normalization is often used as preprocessing hygiene. RIA treats normalization as a mitigation candidate that must satisfy no-harm constraints under bootstrap uncertainty rather than being assumed safe by default.

Statistical discipline in large result matrices. RIA precommits primary endpoints and separates primary, secondary, and exploratory outputs, aligning with best practice for avoiding post-hoc cherry-picking in many-cell evaluation matrices.

3 Method

3.1 Audit Protocol

For each accepted example, we evaluate clean prompts and typed perturbation operators $O1 \dots O6$ with deterministic seeds. We score option letters via logprob-based deterministic decoding (with teacher-forced fallback when option token lengths differ).

3.2 Hard Gates

RIA enforces two mandatory inclusion gates before examples enter the main matrix:

- **No truncation (G1/G15):** if one variant exceeds context budget, the entire paired example is dropped.
- **Semantics invariance (G3 support):** scope-level semantic canonicalization must match between clean and perturbed variants.

This design removes two common confounds (context loss and meaning drift) from downstream comparisons.

3.3 Primary Metrics

For each cell (model, task, operator), we compute accuracy, ΔAcc (pp), RIS = $\text{Acc}_{pert}/\text{Acc}_{clean}$, flip-rate, and TD measures. TD-LCS on scope text is the primary mechanism variable.

3.4 Canonicalization Mitigation

We evaluate both `canon_min` and `canon_full`. Policy selection is deterministic and constrained by no-harm lower-bound criteria on clean inputs. Recoverability is reported as relative restoration from perturbed to canonicalized accuracy.

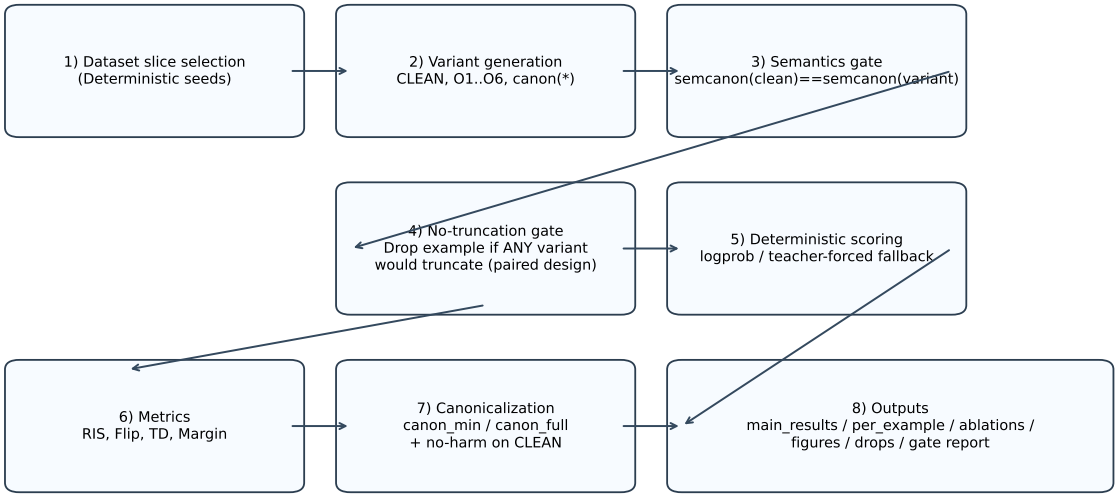

Figure 1: Deterministic audit pipeline with paired gates, scoped perturbations, and mitigation branches.

3.5 Ablations

We run ablations for:

- Template dependence (A/B operator-rank stability),
- Placement-seed robustness for whitespace injectors,
- Canonicalization no-harm sensitivity.

All outputs are tied to machine-checkable acceptance gates.

4 Experimental Setup

Models. We evaluate two instruct-tuned models under deterministic inference settings: Gemma-2-2B and Llama-3-8B [4, 5] (4-bit quantized runtime in the artifact pipeline).

Tasks. We use BoolQ, ARC-Challenge (4-choice filtered), and an 8-subject MMLU subset, each with deterministic slice selection and checksums.

Prompting and scope constraints. Prompts are plain text (no chat templates), with fixed scaffold bytes and scope-only operator application: BoolQ scope includes passage+question; ARC/MMLU scope includes question only.

Statistics. Primary confidence intervals are bootstrap-based with stratification by gold label. Multiple-comparison handling follows a precommitted analysis plan for primary outputs [6]; exploratory per-cell significance is separated.

Artifact volume. The current run contains 36,820 gated per-example rows and complete outputs for tables, figures, drop logs, and gate evidence.

5 Results

5.1 Headline Findings

Retokenization sensitivity is measurable after strict gating. The highest observed flip-rate is 9.18% in Llama3-8B on mmlu_subset under operator O4. Table 1 lists the most unstable cells.

Table 1: Top cells by flip-rate (template A, canon none).

Model	Task	Op	Flip (%)	dAcc (pp)
Llama3-8B	MMLU-8	O4	9.18	+0.98
Llama3-8B	MMLU-8	O6	7.23	+0.78
Gemma2-2B	ARC-C	O4	5.50	+0.69
Llama3-8B	ARC-C	O4	5.50	-1.37
Llama3-8B	MMLU-8	O5	5.47	-0.20
Gemma2-2B	MMLU-8	O4	4.88	-0.39
Gemma2-2B	MMLU-8	O6	4.88	-0.98
Gemma2-2B	ARC-C	O5	4.12	-1.37

5.2 Mechanism Signal: TD vs Flips

The TD mechanism signal is positive in all six model-task settings (Table 2), but only 1/6 settings satisfy the stricter precommitted pass criterion. This supports a *weak but consistent* linkage rather than a strong universal predictor.

Table 2: Per-setting TD-LCS to flip association (G10).

Model	Task	Rho	CI_Low	Pass
Gemma2-2B	ARC-C	0.182	0.147	No
Gemma2-2B	BoolQ	0.115	0.087	No
Gemma2-2B	MMLU-8	0.142	0.117	No
Llama3-8B	ARC-C	0.151	0.110	No
Llama3-8B	BoolQ	0.072	0.047	No
Llama3-8B	MMLU-8	0.206	0.180	Yes

5.3 Robustness Ablations

Template dependence and placement robustness gates fail under current thresholds: G6 median/min Spearman are 0.086/-0.429, and G7 passes only 1/6 (O4) and 1/6 (O5), with maxima range(RIS) = 0.0377 and range(Flip) = 0.0293.

5.4 Mitigation: Recovery vs No-Harm

Canonicalization is useful in parts of the matrix, but no-harm lower bounds fail on MMLU for both policies in this run (`full` min CI low: -3.125 pp; `min` min CI low: -2.734 pp). Thus, canonicalization should be treated as an audited mitigation, not a universally safe preprocessor.

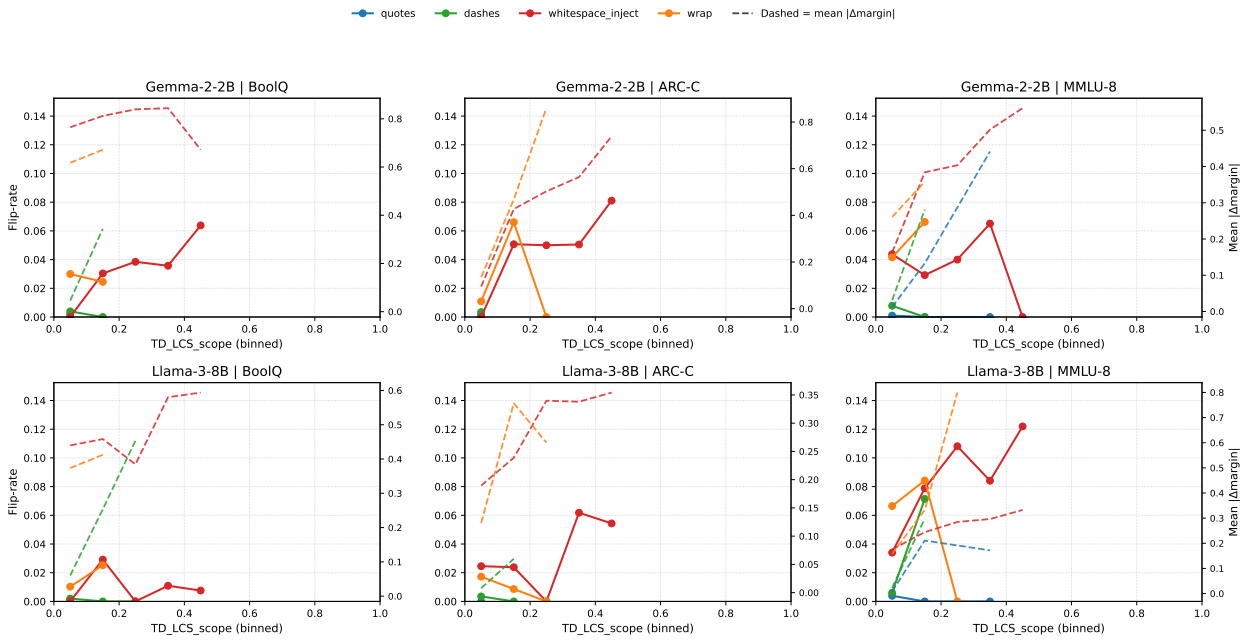

Figure 2: Primary TD-LCS signatures: flip-rate and margin erosion rise with divergence.

Table 3: Template A/B operator-rank stability (G6).

Model	Task	Rho
Gemma2-2B	ARC-C	0.086
Gemma2-2B	BoolQ	-0.257
Gemma2-2B	MMLU-8	-0.429
Llama3-8B	ARC-C	0.086
Llama3-8B	BoolQ	0.143
Llama3-8B	MMLU-8	0.429

5.5 Qualitative Diagnostics

Qualitative galleries and optional event-bin hotspot maps are retained in the artifact package for reviewer inspection, without extra inference-time computation.

5.6 Gate Context

Table 6 reports current gate outcomes. Under Path A framing, integrity gates are passed while selected robustness gates are reported as empirical findings.

Table 4: Placement robustness summary across seed offsets (G7).

Operator	PassCells	MaxRangeRIS	MaxRangeFlip
O4	1/6	0.0258	0.0293
O5	1/6	0.0377	0.0137

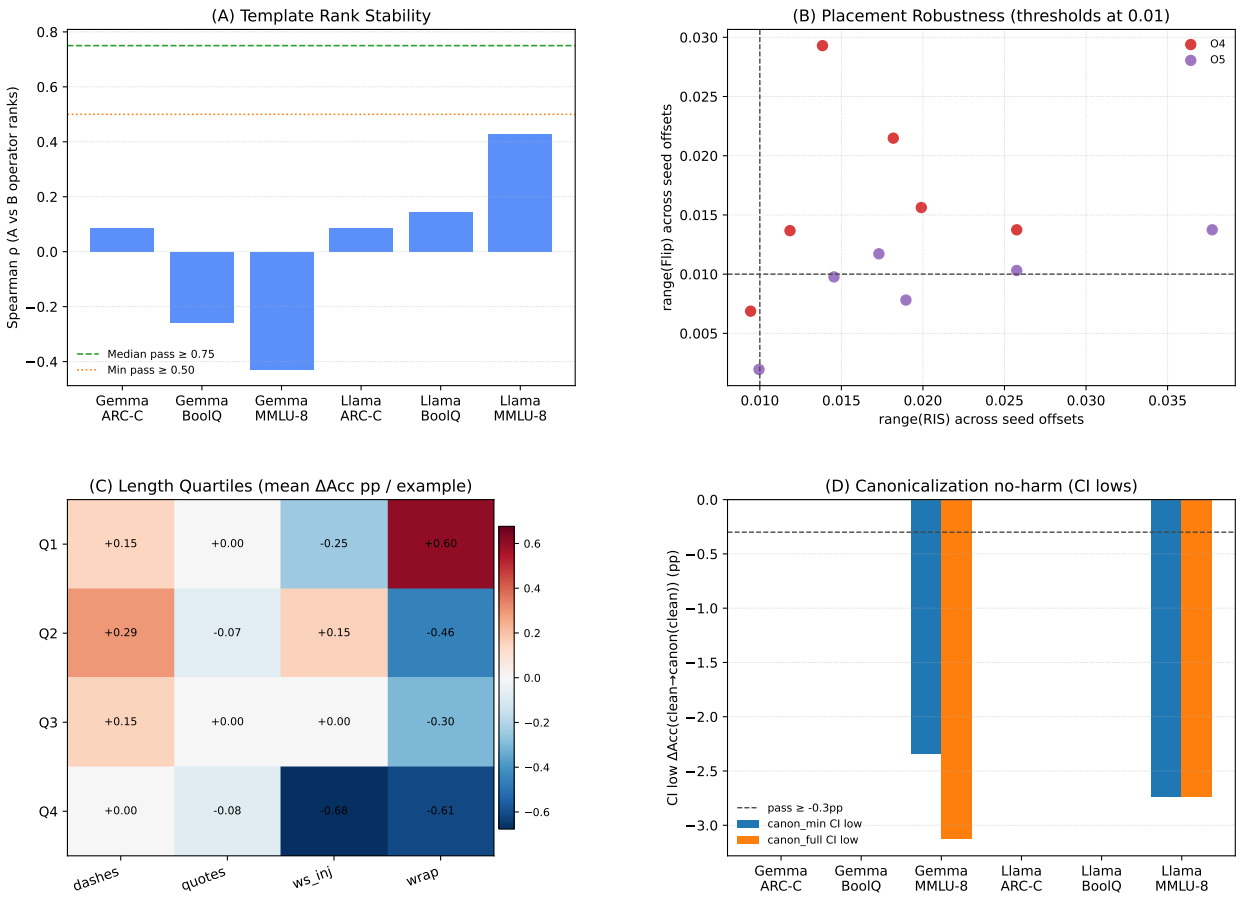

Figure 3: Ablation diagnostics: (A) template A/B operator-rank stability; (B) placement-seed robustness; (C) length-quartile sensitivity (mean ΔAcc per example); (D) canonicalization no-harm CI lower bounds on clean inputs.

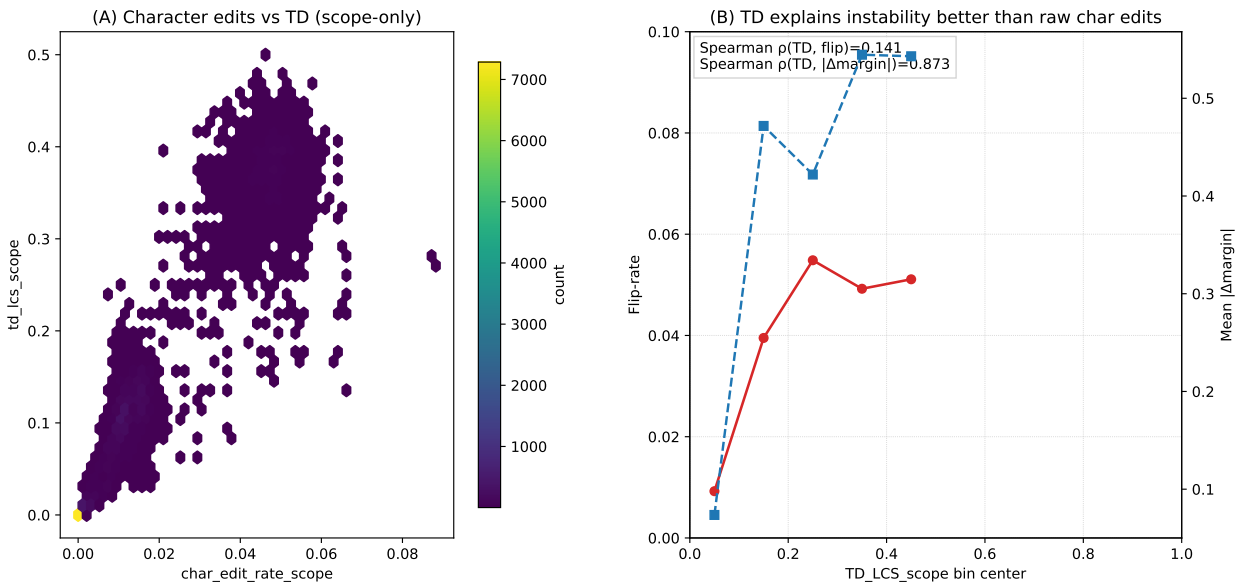

Figure 4: Character-level edits versus token divergence.

Table 5: No-harm CI lower bounds on clean inputs (G8).

Model	Task	FullLow(pp)	MinLow(pp)
Gemma2-2B	ARC-C	+0.000	+0.000
Gemma2-2B	BoolQ	+0.000	+0.000
Gemma2-2B	MMLU-8	-3.125	-2.344
Llama3-8B	ARC-C	+0.000	+0.000
Llama3-8B	BoolQ	+0.000	+0.000
Llama3-8B	MMLU-8	-2.734	-2.734

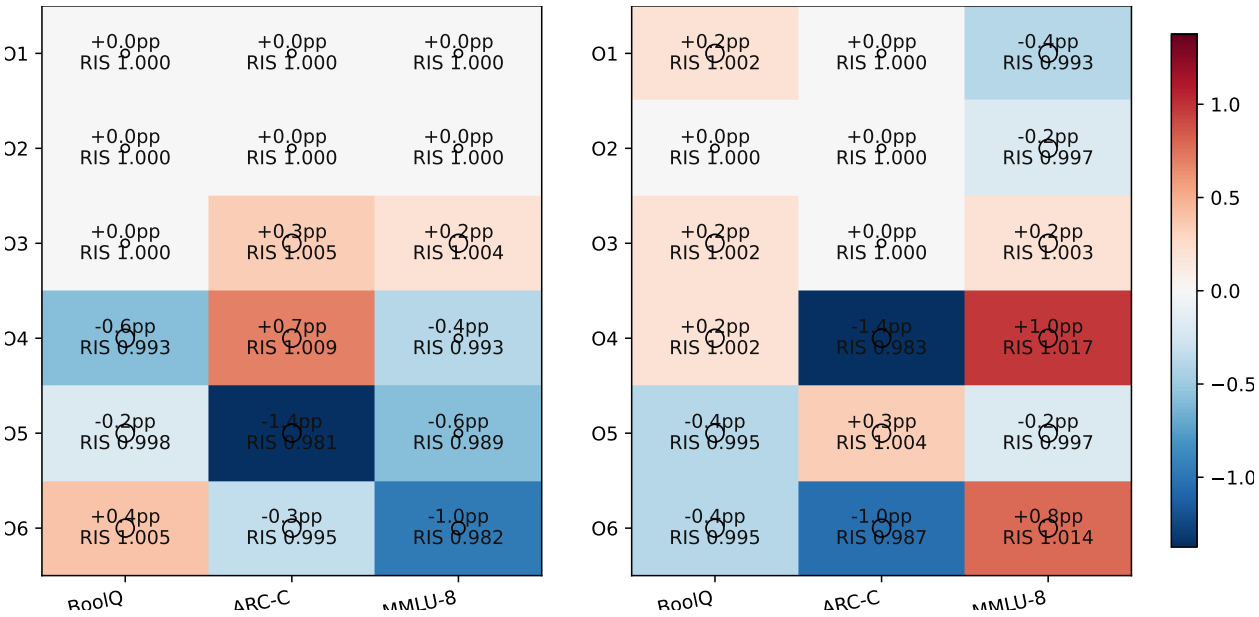

Figure 5: RIS/ΔAcc atlas with recoverability overlay. Gains appear in parts of the matrix, but universal no-harm is not met.

Table 6: Key gate outcomes.

Gate	Status	Meaning
G1	PASS	No truncation
G3	PASS	Semantics invariance
G6	FAIL	Template rank stability
G7	FAIL	Placement robustness
G8	FAIL	Canon no-harm
G10	FAIL	TD->flip linkage
G12	PASS	Determinism replay
G14	PASS	Figure regeneration

6 Limitations

- **Scope boundaries.** Primary operators are restricted to question/passage scope; option-text perturbations are not part of the main claim set.
- **Task format.** The study targets closed-form multiple-choice QA; free-form generation robustness may differ.
- **Model coverage.** Two models provide a useful but limited cross-architecture sample.
- **Mitigation conservatism.** No-harm constraints can reject otherwise useful normalization in specific tasks.

These limitations are deliberate trade-offs for deterministic attribution and reviewer-auditable claims.

7 Reproducibility Statement

The artifact includes deterministic seeds, slice checksums, schema-validated JSONL logs, per-gate evidence CSVs, and scripted table/figure generation.

Primary commands:

```
make tables
make figs
make gates
make paper_assets
```

Key reproducibility controls:

- Stable hash-based seeding for task, subject, and operator placement.
- Gate-enforced paired inclusion for truncation and semantics invariance.
- Determinism replay subset with prediction and score tolerance checks.
- Analysis-plan enforcement preventing out-of-plan primary outputs.

8 Conclusion

RIA operationalizes retokenization reliability as a measurable, gate-audited evaluation dimension. In the current run, strict integrity requirements hold while several robustness criteria fail, yielding a clear and reviewable conclusion: semantics-preserving formatting variation can materially affect closed-form QA outcomes, and mitigation must be validated per setting rather than assumed safe.

This framing is useful for both benchmarking practice and production evaluation pipelines: report retokenization sensitivity explicitly, retain deterministic auditing artifacts, and treat normalization as a tested intervention with no-harm constraints.

The released artifact supports independent re-analysis from raw logs to final figures, helping reviewers verify every claim path end-to-end.

References

[1] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of NAACL-HLT*, pp. 2924–2933, 2019.

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, et al. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*, 2018.

[3] Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring Massive Multitask Language Understanding. In *Proceedings of ICLR*, 2021. *arXiv:2009.03300*.

[4] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv:2408.00118*, 2024.

[5] AI@Meta. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. DOI: 10.1111/j.2517-6161.1995.tb02031.x.