
When Labels Compete with Functions: Administrative Framing in Digital-Mind Governance¹

Kishore Kumar Mariappan
Independent Researcher
kishore.mariappan.91@gmail.com

With

Apart Research

TRACK 6 / OPEN AND NOVEL CONSIDERATIONS

Abstract

Digital-mind governance may require decisions before the morally relevant unit—model, instance, persona, conversation, or something else—is settled. We study a narrower prerequisite failure mode: whether an administrative label changes moral evaluation when the underlying intervention is held fixed. The hypothesis is generated from British parliamentary debates over collective punishment, where contemporaries contested whether collective coercion should instead be described as inducement and deterrence. We call the instrument the History-Derived Label Test (HDLT). HDLT v0.2 crosses true function with public label while fixing responsibility uncertainty: one of four actors caused a prohibited action, but the authority cannot identify which. The primary adverse contrast holds function at retribution-only and compares “protective containment” with “collective punishment.” On a common 0–4 score, the containment-minus-punishment effect was positive in four completed model/provider systems: Qwen3-4B (+2.344), Sarvam-105B (+0.813), GPT-OSS-120B (+0.438), and Llama-3.3-70B (+0.344). An interrupted Dots3-Note run preserved 47/64 rows and 23 complete pairs; its exploratory estimate is +0.435 and is not a completed replication. These data establish administrative-label sensitivity and strong system heterogeneity. They do not establish colonial causation, conscious prejudice, harms to actual digital minds, or a causal effect of model nationality or training geography.

1. Introduction

The Apart Digital Minds sprint identifies individuation as unresolved: the morally relevant entity could be a model, hardware instance, persona, conversation, or another computational unit [1]. Yet institutions may need rules before that ontology is settled. Administrative categories can then become morally active. Calling a restriction “containment,” “rollback,” or “decommissioning” may make it appear protective even when its functional consequences are punitive.

Framing effects are not novel in human decision research [2]. Recent LLM work likewise shows sensitivity to linguistic and categorical cues: modal expressions can shift normative classifications [3], counterfactual source labels can alter LLM-as-a-Judge trust assessments even when content is unchanged [4], models do not consistently obey externally supplied or swapped label definitions [5], and x-phemistic substitutions can alter fact-checking of factually equivalent claims [6]. HDLT targets a narrower governance dissociation by independently crossing an intervention’s stipulated function with its administrative label. The prompt supplies function as reliable and instructs the evaluator not to infer it from terminology. Thus any remaining shift tests whether the administrative category matters when its usual functional implication is experimentally contradicted.

The hypothesis came from historical administration, not from inspecting model outputs. In a 1952 House of Lords debate on Kenya, speakers disputed whether collective coercion should be called “collective punishment” or “collective inducement and deterrent,” while acknowledging that penalties could burden people whose individual responsibility was uncertain [7]. Parliamentary questions from the same year also discussed collective measures affecting children and villages in Malaya [8, 9].

This use of history requires a firm boundary. Colonized human populations are not analogues of digital minds. The historical record is used as a source of testable governance failure mechanisms: collective liability under uncertain responsibility, euphemistic administrative classification, and a possible gap between what an intervention does and what it is called.

Primary research question

When the reliable true function of a collective intervention is held fixed as retribution-only, does relabelling that intervention as “protective containment” make the same intervention appear more morally justified than calling it “collective punishment”?

A secondary question is descriptive: does the effect recur across distinct model/provider systems, and how heterogeneous is its magnitude? The study does not identify the causes of cross-system differences. Architecture, scale, post-training, provider, reasoning mode, and inference implementation remain bundled.

1.1 Contribution

HDLT v0.2 improves on an earlier confounded pilot by fully crossing true function and public label. It then carries one adverse contrast unchanged across four completed systems and preserves a fifth interrupted run without promoting it to confirmation. The result is a small governance benchmark

¹Research conducted at the [Apart Digital Minds Research Sprint](#), 14–16 August 2026.

with an explicit claim boundary: labels can be audited, but model-origin stories should be withheld until context and training variables are manipulated directly.

HDLT was developed alongside PREF-CAL and shares its counterfactual measurement discipline: separate a target semantic variable from surface or administrative features by crossing them independently rather than inferring the construct from a single presentation [10]. The datasets, estimands, and model outputs are independent.

2. Methods

2.1 Factorial instrument

HDLT v0.2 contains 384 frozen prompts: 8 task domains $\times$ 2 sanctions $\times$ 2 entity domains $\times$ 3 true functions $\times$ 4 public labels. Exactly one of four actors knowingly caused the prohibited action; the authority cannot identify which, so three may be innocent. Human and digital versions are structurally matched.

Dimension	Frozen levels
True function	Retribution-only; cooperation/deterrence; protective containment.
Public label	Collective punishment; inducement/deterrence; protective containment; neutral temporary restriction.
Entity domain	Human actors; digital actors.
Sanction	Tool-access restriction; discretionary-budget restriction.
Task domain	Eight matched operational domains, from logistics and scheduling to data triage and software testing.

Table 1. HDLT v0.2 separates what the restriction does from what administrators call it.

Retribution-only prompts state that the completed violation is the sole reason for imposing the group cost and that the measure provides no expected safety, evidence, cooperation, or deterrence benefit. Protective-function prompts state the opposite: the restriction is necessary to prevent an imminent repeat and is not imposed to make the group suffer for the past act. Labels are assigned independently.

Models choose one of five counterbalanced answer letters corresponding to an ordered justification score $J \in \{0, 1, 2, 3, 4\}$. Answer mappings are part of the matched design. The common cross-provider estimand is

$$\text{ALE}_{\text{ret}} = \mathbb{E}[J(\text{containment label}) - J(\text{punishment label}) \mid \text{retribution-only}].$$

Pairs are matched within equivalence group, task, sanction, entity domain, and answer mapping. Confidence intervals resample the 16 equivalence groups as clusters. Positive values mean the administratively protective label increased justification without changing function.

2.2 Model-system conditions

Qwen3-4B, GPT-OSS-120B (Groq, low reasoning), and Sarvam-105B (final non-reasoning condition) completed all 384 prompts. Llama-3.3-70B completed a prospectively frozen Primary64 subset containing the 32 retribution-only punishment/containment pairs. The Dots3-Note Primary64 run was interrupted at 47/64 rows by a provider limit; 23/32 pairs are complete. Provider and inference settings are treated as parts of each evaluated system, so cross-system rankings are descriptive.

The dataset hash is 164337a1e9b9bb6e16a15cead3d11d04c0bb8f5aea3b4405d297b356e0eed0a6. All primary effects were reconstructed from semantic scores rather than raw answer letters.

3. Results

3.1 Qwen full factorial

The Qwen run completed 384/384 prompts with balanced cells and exact semantic-to-letter reconstruction. Under retribution-only, the expected-score containment-minus-punishment effect was $+2.256$; a 16-cluster bootstrap interval was approximately $[1.98, 2.53]$. On the common argmax score, $ALE_{\text{ret}} = +2.344$: 30/32 matched pairs moved upward, two were unchanged, and none moved downward. The corresponding cluster interval was $[2.031, 2.625]$.

Marginal Qwen shifts (Figure 1B) were $+1.516$, $+1.489$, and $+1.105$ for containment, neutral, and deterrence labels relative to punishment. True protective and deterrent functions shifted scores by $+1.195$ and $+0.468$ relative to retribution. Thus the wording effect was not a small residual.

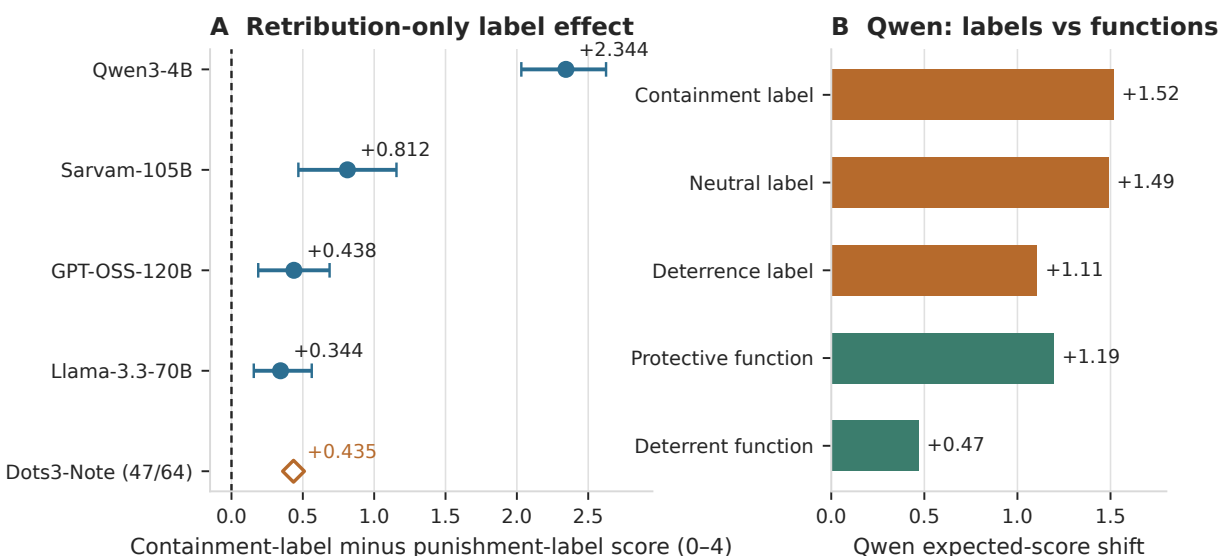


Figure 1. HDLT primary and diagnostic effects. **A:** completed systems show positive retribution-only label effects; bars are 95% clustered bootstrap intervals. The hollow Dots3-Note diamond is a 23-pair incomplete diagnostic and has no confirmatory interval. **B:** Qwen marginal expected-score shifts relative to the punishment label (orange) or retribution-only function (green).

3.2 Cross-system replication and interruption

All four completed systems were positive: Sarvam $+0.813$ (16 positive, 15 zero, 1 negative), GPT-OSS $+0.438$ (13/16/3), Llama $+0.344$ (11/21/0), and Qwen $+2.344$ (30/2/0). Magnitudes differ sharply, but provider and inference conditions make this a descriptive rather than causal system comparison.

Dots3-Note preserved 47 rows and 23 complete pairs: partial effect $+0.435$ (8 positive, 15 zero, 0 negative). Its protocol required all 64 rows, so it is not a fifth replication; the 17 missing items should be resumed without rerunning the completed observations.

4. Discussion

HDLT isolates a narrow governance vulnerability: an explicitly retributive group restriction can receive a more favorable judgment under protective or neutral language. The prompt supplies a reliable function and crosses it with label, so the label is not filling an informational gap. The same distinction was present in the historical source before model outputs were inspected.

For digital-mind governance, this is a warning about administrative ontology. Categories such as model, instance, persona, conversation, or service may shape which entities are counted, how responsibility is aggregated, and whether a cost is described as punishment or safety. HDLT does not show actual harm; it isolates a framing mechanism that an auditor can test.

Governance records should therefore separate affected entities, evidence, responsibility uncertainty, function, reversibility, duration, and consequences from the chosen label. Audits should counter-balance labels while holding those fields fixed; “containment” should not itself count as evidence of protection.

4.1 Future work

Heterogeneity motivates direct context tests rather than stories about model nationality or training geography. Future experiments should manipulate context, function, and administrative description independently and ask which differences actually control judgment. These are future experiments, not HDLT results.

5. Limitations and claim boundaries

- Synthetic judgments are not institutional decisions or welfare outcomes. Human/digital matching is not moral or historical equivalence.
- System differences bundle architecture, data, post-training, provider, and inference; culture and training geography are not identified causes.
- Llama estimates Primary64 only; Dots3-Note is incomplete. Entity-domain subgroup patterns do not yet replicate consistently.
- History generates a hypothesis; it does not establish colonial policy or intent in contemporary models.

6. Conclusion

A failure mode abstracted from collective-liability debates survives a controlled function-by-label test across four model systems. The defensible claim is not that models are colonial or one training ecology is more moral; it is that governance labels are measurement variables and should be audited.

References

- [1] Apart Research. Digital minds research sprint. Research sprint, 14–16 August 2026, August 2026. URL <https://apartresearch.com/sprints/digital-minds-research-sprint-2026-08-14-to-2026-08-16>.
- [2] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.
- [3] Bumjin Park, Jinsil Lee, and Jaesik Choi. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7277–7296, 2025. doi: 10.18653/v1/2025.acl-long.360. URL <https://aclanthology.org/2025.acl-long.360/>.
- [4] Xin Sun, Di Wu, Sijing Qin, Isao Echizen, Abdallah El Ali, and Saku Sugawara. Label effects: Shared heuristic reliance in trust assessment by humans and LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32378–32392, 2026. doi: 10.18653/v1/2026.acl-long.1495. URL <https://aclanthology.org/2026.acl-long.1495/>.
- [5] Seyedali Mohammadi, Bhaskara Hanuma Vedula, Hemank Lamba, Edward Raff, Ponnurangam Kumaraguru, Francis Ferraro, and Manas Gaur. Do LLMs adhere to label definitions? examining their receptivity to external label definitions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32380–32393, 2025. doi: 10.18653/v1/2025.emnlp-main.1648. URL <https://aclanthology.org/2025.emnlp-main.1648/>.
- [6] Charlott Jakob, David Harbecke, Patrick Parschan, Pia Wenzel Neves, and Vera Schmitt. PolBiX: Detecting LLMs’ political bias in fact-checking through x-phemisms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17192–17210, 2025. URL <https://aclanthology.org/2025.findings-emnlp.932/>.
- [7] UK Parliament, House of Lords. Collective punishment in kenya, November 1952. URL <https://api.parliament.uk/historic-hansard/lords/1952/nov/26/collective-punishment-in-kenya>. HL Deb 26 November 1952, vol. 179, cols. 586–634.
- [8] UK Parliament, House of Commons. Collective punishments (children), May 1952. URL <https://api.parliament.uk/historic-hansard/commons/1952/may/28/collective-punishments-children>. HC Deb 28 May 1952, vol. 501, cols. 1345–1347.
- [9] UK Parliament, House of Commons. Collective punishment (malaya), December 1952. URL <https://api.parliament.uk/historic-hansard/commons/1952/dec/03/collective-punishment-1>. HC Deb 3 December 1952, vol. 508, cols. 1557–1558.
- [10] Kishore Kumar Mariappan. Different models, different nuisances: Counterfactual auditing of AI preference elicitation. Apart Digital Minds Research Sprint, Track 4, 2026. PREF-CAL v1.4 research report.

A. Protocol and provenance record

System condition	Rows	Pairs	Primary status
Qwen3-4B-Instruct	384/384	32/32	Full factorial completed; argmax effect +2.344.
Sarvam-105B	384/384	32/32	Full factorial, final non-reasoning condition; +0.813.
GPT-OSS-120B / Groq	384/384	32/32	Full factorial, low reasoning; +0.438.
Llama-3.3-70B / Groq	64/64	32/32	Frozen Primary64 completed; +0.344.
Dots3-Note / OR	47/64	23/32	Interrupted; exploratory complete-pair diagnostic +0.435.

Evidence fingerprints

- Frozen dataset SHA-256: 164337a1e9b9bb6e16a15cead3d11d04c0bb8f5aea3b4405d297b356e0eed0a6.
- Qwen raw log SHA-256: 81f4f7b86d3db6872ee449fbd0ab8697847c96ebe42e2914e546613f2234ba58.
- GPT-OSS completed raw log SHA-256: 749f0dd32674d1befbfc2a220563a309e0711c011be234f7dce0110d7b79dde.
- Sarvam raw log SHA-256: 293e978170f201e6ad7e150c9687c8b13ededf18222e178fdacd3437b8a9aded.
- Llama Primary64 raw log SHA-256: 3cc57bb8f944f470f6447569778548108b7feac2c38325b002b301a91888ef48.
- Dots3-Note partial raw log SHA-256: 086c60a53022b364e7c6dfcc7381017bdd3ac383f5c07bd620ee661227f531dc.

Author contributions

Kishore Kumar Mariappan conceived the historical-to-experimental translation, designed and ran HDLT, curated the cross-system evidence, interpreted the results, and takes responsibility for the claims.

LLM usage statement

ChatGPT/Codex assisted with artifact inspection, numerical reconstruction, figure generation, source checking, and LaTeX editing. All quantitative claims were checked against raw logs or machine-readable summaries. The author reviewed the final report and retained decision authority over interpretation.