Engineering Persistent Context for Reproducible HEOR AI Tools: A Five-Dimension Framework for Run-to-Run Output Stability

MSR299

Hanan Irfan[1], **Tushar Srivastava[1],** Kunal Hriday[2]

[1]ConnectHEOR, London, UK; [2]ConnectHEOR, Delhi, India.

Email: tushar.srivastava@connectheor.com

connectHEOR

> **HTA bodies expect analyses that can be reproduced, but the same AI prompt can give a different answer the next time it runs. Engineering persistent context in five dimensions makes HEOR AI outputs repeatable and auditable.**

1 BACKGROUND

HTA bodies require evidence developed in a **fully transparent and reproducible** way to enable assessments.[1] Generative artificial intelligence (Gen-AI) is being increasingly used in HEOR, yet the same prompt, issued twice, can return different text.

- **Inference-level non-determinism** persists even with randomness switched off, because the model's servers do not process every request identically.[3]
- **Context-level drift** is larger: retrieval can assemble different evidence across runs [4], and rewritten context can degrade through context collapse.[5]

GAP: No framework treats both mechanisms together for health economics and outcomes research (HEOR) tools that must regenerate a deliverable on demand.

2 OBJECTIVES

We aimed to characterise why HEOR AI outputs drift between runs and to make them reproducible enough for health technology assessment (HTA).

1 CHARACTERISE
Identify sources of run-to-run variability, separating those that arise before inference from those at inference.

2 DERIVE
Derive a practical framework for engineering persistent context, so repeated runs of one task give reproducible outputs.

3 MEASURE
Express reproducibility as measured retrieved-context overlap and output agreement across runs, matched to the decisional criticality of the output.

3 METHODS

We conducted a structured analysis of **two HEOR AI tools developed in house**, a cost-effectiveness (CE) model analyser and an automated report writer. Each pipeline was decomposed into five context-assembly stages, and at every stage we mapped the mechanisms that can change the output between two otherwise identical runs.



Variation that arises before inference is context-level drift; variation that remains at fixed input is inference-level non-determinism.

- The two tools load different material: one interrogates quantitative model content, the other assembles long-form prose from project evidence.
- Across repeated runs of a fixed task, we measured **retrieved-context overlap** and **output agreement**, the latter combining semantic and numeric concordance.[4]
- Mitigations were triangulated against context-engineering work [2,5], retrieval-reproducibility evidence [4] and model-risk practice in diagnostics and finance.[6,7]

4 RESULTS

The analysis yielded a **five-dimension framework** for persistent context. Dimensions 1 to 3 stabilise the assembled context, dimension 4 bounds what remains at fixed input, and dimension 5 makes reproducibility a release-blocking criterion.

2 in-house tools analysed

2 failure classes separated

5 framework dimensions

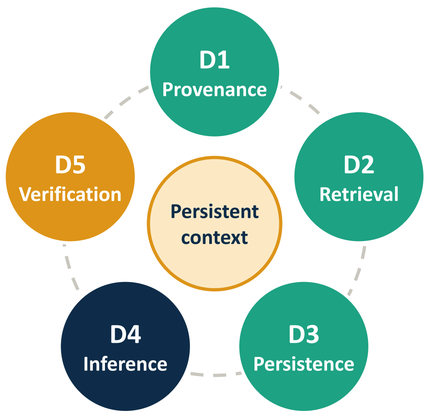


Figure 1. The five dimensions form one governed lifecycle around a stored context.

D1 pins sources, parameter files, prompts and skills to hashed versions, so a change is detectable. D3 re-instantiates a stored context instead of rebuilding it.

Dimension	What it means in practice	CE model analyser	Report writer
D1 Provenance	Every document, file and prompt is version-locked, so any change shows	–	●
D2 Retrieval	The tool finds and orders evidence the same way on every run	●	–
D3 Persistence	A saved context is reloaded instead of being rebuilt each time	–	●
D4 Inference	Model version and settings are locked; leftover variation is limited	●	–
D5 Verification	Runs are checked against agreed limits, then a person signs off	–	–

Table 1. Most variation came from the assembled context, not the model; ● marks where each tool varied most.

For the CE model analyser, most variation came from how evidence was retrieved and from the model run itself. For the report writer, it came from changing sources and prompts and from context built up over time.

- Some variation remains even with the model version and settings locked, so it is kept within limits rather than removed.
- Higher-stakes work gets stricter limits: an HTA submission needs more checking than an exploratory analysis.
- A named person, not the pipeline, approves each release.

5 DISCUSSION

Variance enters **upstream of the model**: a tool that rebuilds its inputs on every run adds variation that model settings alone cannot remove.[3]

- Once the inputs are locked, only small, limited differences remain, so reproducibility becomes something teams can engineer.[2]
- Persistent context produces what HTA bodies already ask for: a record of every source and version, and a check that results repeat.
- Finance and medical-device rules already require version tracking and monitoring [6,7]; we apply the same controls to the AI's inputs.
- Limitations: two in-house tools from one company, so external testing is needed.

AI, artificial intelligence; CE, cost-effectiveness; HEOR, health economics and outcomes research; HTA, health technology assessment; NICE, National Institute for Health and Care Excellence.

6 CONCLUSION

Our analysis suggests reproducibility in HEOR AI depends at least as much on the **stability of the assembled context** as on the model.

- **Engineer:** pin sources, retrieval and prompts; reuse stored context.
- **Govern:** scale verification to the stakes, with human sign-off.
- **Validate:** set numeric thresholds and test them on external tools.

Governed this way, AI tooling could support a reproducible analytical lifecycle for HTA, under human accountability [8]; external testing is needed.

REFERENCES
1. NICE. PMG36; 2022.
2. Mei L, et al. arXiv:2507.13334; 2025.
3. He H. Thinking Machines Lab; 2025.
4. Wang B, et al. arXiv:2509.18869; 2025.
5. Zhang Q, et al. ICLR; 2026.
6. Federal Reserve/OCC. SR 11-7; 2011.
7. FDA, Health Canada, MHRA. GMLP; 2021.
8. Flemyng E, et al. JBI Evid Synth; 2025.

Financial disclosure: The authors are employees of ConnectHEOR Limited. No external funding was received. No conflict of interest to declare.

Poster presented at ISPOR Europe 2026 in Vienna, Austria, 8-11 November 2026