What Must Be Explicit? Lessons from AI-Assisted Implementation of a Health Economic Model

MSR60

Nishtha Neeraj[1], **Susannah Sadler[2]**

[1]ConnectHEOR, Delhi, India; [2]ConnectHEOR, London, United Kingdom.
Email: Susannah.Sadler@connectheor.com

connectHEOR

What humans infer, AI needs spelled out: 14 prompt dependencies across 10 domains turned an ambiguous HEOR prompt into a reproducible model specification with ±1% benchmark agreement.

14	10	±1%
prompt dependencies	recurring domains	final benchmark tolerance

1 BACKGROUND

Large language models (LLMs) are increasingly used in health economics and outcomes research (HEOR) for model development, critique, implementation and education.[1,2]

- Technical errors are common in models submitted for appraisal, with transcription, logic and computational errors accounting for the largest share.[3]
- Srivastava et al. proposed a hybrid human–AI framework for HEOR tasks such as model conceptualization, in which experts validate and refine LLM output.[2]
- Prompting techniques are largely domain-agnostic [4], so it is domain specification completeness that may drive accuracy, efficiency, token use and human versus AI effort.

The gap: little guidance covers the HEOR conventions transferable to AI that experienced modellers infer but rarely document.

2 OBJECTIVES

Research question **What must an LLM be told, beyond a conventional HEOR specification, to implement the intended model reproducibly?**

IDENTIFY Identify the specification dependencies on which implementation success depended.

CLASSIFY Group them into domains concerning HEOR methodology or governance of the implementation.

TEST Test whether stating them explicitly brings the model outputs within predefined compliance criteria.

3 METHODS

A synthetic oncology partitioned survival model (PSM) was used as a test instrument so that the intended model behavior could be fixed in advance. The benchmark was the independently predefined output of the locked synthetic reference model.

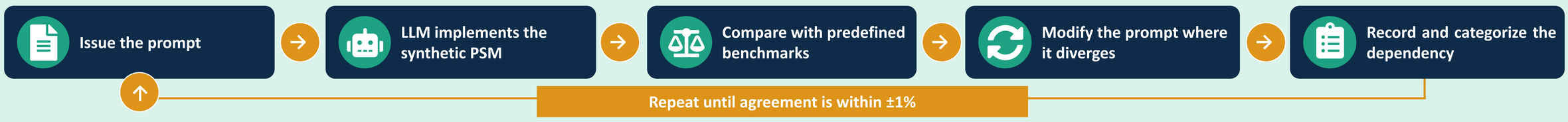


Figure 1. Where the implementation diverged from the benchmarks, the prompt was modified and the cycle repeated.

- Prompt changes were recorded during development. If the model only worked correctly after a clarification was added, that clarification was treated as a **prompt dependency**.
- **HEOR-specific dependencies** affected how the model calculated results.
- **Governance dependencies** affected how easily the model could be checked, reproduced and reviewed.
- Numerical results were compared with **independently calculated reference outputs** using a **±1% pass/fail threshold**.
- Governance requirements were checked separately, including **locked inputs, QC checks, visible calculation traces and clear pass/fail reporting**.

4 RESULTS

Iterative prompt refinement identified **fourteen** specification dependencies across **ten** domains. The final explicit specification reproduced all predefined **benchmark outputs within ±1%** and passed the specified QC checks.

HEOR-specific: 6 of 10		Governance: 4 of 10
• Model structure	• Survival parameterisation	• Input mutability
• State derivation	• Discounting conventions	• Validation logic
• Cost allocation	• Utility calculations	• Transparency requirements
		• Benchmarking standards

Figure 2. Six domains concern HEOR methods and four concern governance.

Dependency type	How it was assessed	What it tells us
HEOR model mechanics	Compared with independently calculated reference outputs using the ±1% acceptance threshold	Whether the model produced the intended numerical results
Governance requirements	Checked against predefined implementation criteria, e.g. locked inputs, QC checks, visible traces, explicit pass/fail reporting	Whether the implementation could be checked, reproduced and reviewed

Table 1. Numerical model mechanics and governance requirements were assessed using different criteria.

Domain	Early instruction	Refined specification
Survival	"Use Weibull survival."	Define $S(t) = \exp(-\lambda t^{\gamma})$; λ=rate, γ=shape, t=months.
Discounting	"Apply 3.5% annual discounting."	End-of-cycle: $1/(1.035)^{t/12}$; no half-cycle correction.
Cost allocation	"Include treatment and terminal-care costs."	Specify cost by state, timing and duration; terminal care on new deaths only.
Input mutability	"Build an interactive model."	Lock benchmark inputs; keep scenario inputs separate.
Validation	"Check that the model is correct."	Require PFS≤OS, states=1, monotonic death, non-negative outputs and ±1% pass/fail.
Transparency	"Show the results."	Show assumptions, formulas, cycle trace, key calculations, warnings and QC status.

Table 2. Illustrative Examples of Prompt Refinement.

5 DISCUSSION

1 Specify at the level of executable rules, not labels.

"3-state PSM", "Weibull", "apply discounting" or "include terminal care" can be clear to an experienced modeller yet permit multiple reasonable implementations.

2 Longer is not the objective; less ambiguous is.

A useful rule: if two competent modellers could implement an instruction differently, make the computational rule explicit.

3 Governance requirements need testable criteria.

Specify what is fixed, what must be checked, what must be visible, and what counts as a pass.

4 Treat the final specification as part of the audit trail.

LLM-generated code may vary between runs [6], so the final specification given to the LLM should be retained and versioned. This complements ELEVATE-GenAI, ISPOR guidance emphasizing transparent and reproducible reporting of LLM-assisted HEOR research [7].

6 LIMITATIONS

- One synthetic oncology PSM, one implementation context and one research team.
- Specification versions were refined iteratively rather than prospectively frozen; baseline V1 numerical performance and specification-length metrics were not formally captured.
- The 10 domains are an emergent taxonomy, not a validated or exhaustive framework.
- ±1% tests numerical agreement with the synthetic reference model; it does not establish decision fitness or external validity.
- Token use, elapsed time and modeller effort were not quantified.

CONCLUSION

Effective AI-assisted modelling requires **more than conventional model specification prompting** making implicit HEOR conventions explicit influenced accuracy, reproducibility and validation burden. The 14 dependencies across 10 domains could form the basis of HEOR guidance for LLM-assisted modelling.

REFERENCES

1. Fleurence RL, et al. Value Health. 2025;28(2):175–183.
2. Srivastava T, et al. Adv Ther. 2025;42:4103–4130.
3. Radeva D, et al. Int J Technol Assess Health Care. 2020;36(4):311–316.
4. Sahoo P, et al. arXiv:2402.07927. 2024.
5. Woods B, et al. NICE DSU TSD 19. 2017.
6. Ouyang S, et al. ACM Trans Softw Eng Methodol. 2025;34(2):42.
7. Fleurence RL, et al. Value Health. 2025;28(11):1611–1625.

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

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