Mock EAG-style Reviews Using Generative AI to Anticipate HTA Critique

MSR62

Hanan Irfan, **Tushar Srivastava,** Thaison Tong, Shilpi Swami

ConnectHEOR, London, UK[1]

Email: tushar.srivastava@connectheor.com

> **A scope-grounded mock-EAG AI tool produced a NICE EAG-style critique in under two hours, recovering 12 of 14 key issues compared with an actual submission without fabricating a single parameter, comparator or value.**

1 BACKGROUND

After a company files with NICE, the **External Assessment Group (EAG)** tests the submission independently, and the company sees the critique and clarification questions only after filing, with limited time to respond.[1,2]

- EAGs test adherence to the final scope, the evidence synthesis, and the structure, inputs and assumptions of the economic model.
- Generative AI could simulate this process so that company submissions can anticipate questions and critiques in advance.

Evidence gap: a mock EAG critique is only useful if every assertion traces to the submission, and it is not known whether a system can do this.

2 OBJECTIVES

We developed and validated a **scope-grounded LLM system** that drafts NICE EAG-style critique reports and clarification questions, with three objectives.

1 RECOVERY Quantify how completely it recovers the key issues an EAG raised for an actual submission used as test case; issue-level sensitivity was primary.

2 GROUNDING Establish how often it produces invalid, non-material or ungrounded content, or fabricates evidence.

3 SPEED Measure how far it shortens time to a structured first-draft critique, against a human-only draft.

3 METHODS

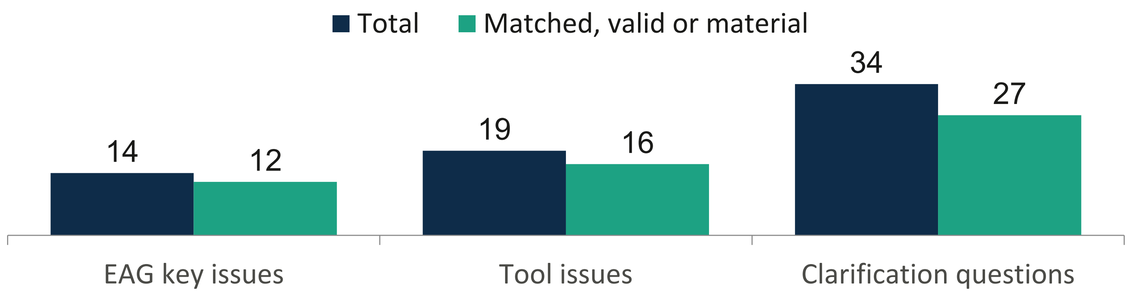
We built a multi-stage, retrieval-grounded pipeline on a reasoning LLM (GPT-5.4) that ingests two documents: the company submission and the NICE final scope, treated as a first-class input because it fixes the population, comparators and outcomes.[2] The case study was the NICE appraisal of **lorlatinib** for untreated ALK-positive advanced non-small cell lung cancer (NSCLC), covering clinical and cost effectiveness. The workflow is described below:

INPUTS Company submission and NICE final scope → **STAGE 1** Structured critique outline per domain → **STAGE 2** Issue-based critique under strict evidence grounding → **STAGE 3** Key-issue and clarification questions comparison against actual EAG review → **ADJUDICATION** Two blinded reviewer's vs the EAG report

- **Grounding:** facts were confined to the current submission section, while exemplar reports informed style only, never content.
- **Reference standard:** the published EAG report, whose key issues were numbered by the EAG itself and could not be revised afterwards.
- **Adjudication:** two senior HTA reviewers, independent of development and blinded, judged each issue for validity, materiality and grounding.

4 RESULTS

The tool flagged 19 issues in the company submission whereas the actual EAG critique had 14 issues. 12 of the tool-suggested issues matched the 14 key issues in the published EAG report (86%). Matched issues included immature overall survival, reliance on investigator-assessed progression after central review ceased, and a weak indirect comparison: interpretive criticisms that turn on judgement about what the evidence can support.



Most tool output was confirmed by senior reviewers against the EAG report.

Adjudicated outcome	Result
EAG key issues recovered	12 / 14 (86%)
Valid tool issues	16 / 19 (84%)
Decision-critical tool issues	**12 of 19**
False positives	**3 of 19**
Fabricated trials, comparators or values	**None**
Clarification questions matched with original	27 / 34 (79%)
Time to first-draft critique	**<2 h vs multi-day**

Of the 19 tool issues, 16 were valid and 12 decision-critical; the other valid issues were defensible but of lower priority. Of 34 clarification questions, 27 were material and answerable, as in the EAG's own questions.

Three false positives were overstated. No fabricated trials, comparators or values were found; surviving errors were verbosity and over-confident wording.

- **Speed:** a first-draft critique took under two hours, against an estimated multi-day human draft (estimation based on domain work).

5 DISCUSSION

No fabrication was found. Confining content to the submission is designed to limit recall from training data, but the submission and EAG report are public, so training-data exposure cannot be excluded in this single appraisal.[3–5] High recovery with three false positives defines the usable role.

- Retrieval grounding constrains generation to a source set, and appraisal suits it because authoritative documents are supplied in full.[5]
- The output is a set of candidate issues for expert adjudication, the human-in-the-loop role that NICE and ISPOR guidance require.[3,6]
- This single-appraisal validation tested one model configuration, and the system has not yet been used prospectively in a live EAG workflow.

6 CONCLUSION

A scope-grounded LLM system produced an auditable first-draft EAG critique in under two hours, recovering 12 of 14 key issues in one NICE appraisal, with no fabricated evidence.

- **In this single appraisal,** confining every statement to the submission kept the critique traceable and free of invented trials, comparators or values.
- **Anticipating critique:** manufacturers can test a submission against likely EAG issues and clarification questions before it is filed.
- **Next steps:** prospective testing across more appraisals, indications and model configurations.

Implication: the system drafts and triages, senior reviewers verify and extend, and accountability stays with named EAG reviewers.

REFERENCES

1. Kaltenthaler E, et al. Health Technol Assess. 2011;15(22):1–82.
2. NICE. Technology appraisal and highly specialised technologies guidance: the manual (PMG36). 2022, updated 2026.
3. NICE. Use of AI in evidence generation: NICE position statement (ECD11). 2024.
4. Chelli M, et al. J Med Internet Res. 2024;26:e53164.
5. Lewis P, et al. Adv Neural Inf Process Syst. 2020;33:9459–9474.
6. Fleurence RL, et al. ELEVATE-GenAI. Value Health. 2025;28(11):1611–1625.

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