

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

8. Knowing When Not to Answer: Selective Prediction for Public-Sector Question Answering

Authors: A M Esfar E Alam¹; Md Nisar Uddin Adil²

Affiliations: 1. A M Esfar E Alam — Department of Informatics, University of Oslo, Norway 2. International Social Welfare and Health Policy, Oslo Metropolitan University, Norway

Abstract: When AI answers questions about entitlements, obligations, or compliance, the damaging failure is rarely an obviously wrong answer. It is a confident, fluent, well-cited answer to a question the underlying record cannot support. The citizen has no way to detect it, and the provenance trail looks identical to a correct response.

We argue that for public-facing systems, the ability to withhold an answer should be reported as a first-class result alongside accuracy. We evaluate a question-answering configuration that combines retrieval, session memory, and symbolic validation, and that declines to answer when the combined evidence is insufficient. On 4,021 queries over a public product database it was not developed against — a schema with different field names and substantial genuinely missing data — it reaches 97.21% decision accuracy (Wilson 95% CI 96.66–97.68) at 73.0% answer coverage, with an expected calibration error of 0.021 and abstention precision of 0.998.

Coverage varies sharply by question type: where the underlying fields are sparsely populated, coverage drops to 38% while accuracy is preserved. The system responds to thin evidence by answering less rather than guessing, and its confidence is lowest exactly where it answers least often.

The broader claim is a governance one. Whether 73% coverage is acceptable is not a technical question but a policy question about what a withheld answer costs in a given setting — and that question can only be asked if abstention is measured and published rather than hidden inside an accuracy figure.