Powered by industry and academia

DELPHI

Building trust in AI: how to avoid confident nonsense

Quality management demands data-driven decisions, yet processing vast datasets drains critical human resources. Generative AI promises to solve this bottleneck, but at what cost? The fundamental question isn't whether AI can process data faster—it's whether it can match the nuanced judgment that experien ced professionals bring to quality decisions. Finding the right balance between efficiency and human ex pertise remains a key challenge.

Reliability Metrics

Find a measure of confidence that reflects correctness

Establish reliable confidence metrics for GenAI responses to make informed decisions about when you can trust AI outputs. Two approached that may prove particularly useful are:

- Utilize token-level confidence scores as available heuristics, but keep in mind that language models tend to be overconfident in their predictions. To find cases where confidence actually correlates with correctness, you may need to look only at the very top probability scores.
- Deploy multiple model calls to identify straightforward cases through consensus detection. Cases exhibiting high inter-model agreement across outcomes can be classified as trivial and processed automatically, while instances with significant response variance can indicate greater complexity and necessitate human expert evaluation.

Transparent Reasoning

Answers must come with evidence, such as rationale (“because of ”) and source attribution (“based on”)

Require the model to include both traceable sources for its claims and explicit reasoning for its conclusions. This dual approach enables independent verification of the underlying information and the thought process that led to the response.

Be aware of its sensitivities

Know when to rely and when to doubt

Build up your team's experience using AI so they learn to recognize its common quirks - overconfidence in uncertain scenarios, excessive caution that blocks useful outputs, generic responses that miss the mark, systematic biases relevant to your use case, context drift in longer conversations, and verbose responses that hurt the user experience.

Partners:

PHILIPS

Contact

anca.lichiardopol@tno.nl

www.esi.nl