From “ChatGPT, list hazards” to reliable hazard identification

The use of large language models (LLMs) in safety-critical systems engineering promises major productivity gains. It also raises questions about quality assurance, reviewability, and integration into the safety case, especially when AI supports safety-related engineering tasks rather than the operational system itself. In hazard analysis and risk assessment (HARA), building a sufficiently comprehensive set of hazard events typically requires expert workshops and several review cycles. As systems and contexts grow more complex, this becomes a bottleneck. Simple “one-shot” prompting (for example, “list hazards”) does not fit safety engineering. Outputs are often inconsistently structured, hard to trace to inputs, and difficult to review systematically. However, these are basic requirements for any safety process.

We present an engineer-led alternative. LLMs are embedded into a structured, human-in-the-loop workflow with explicit checkpoints and standardized outputs, implemented as a Hazard Identification Tool (HIT). The workflow restricts the LLM’s role, applies systematic prompt engineering, and augments prompts with domain safety knowledge via retrieval-augmented generation (RAG), supported by Hypothetical Document Embeddings (HyDE). It first turns a short system context into a reviewed system description and system-level requirements. Only then does it derive concise, reviewable hazard candidates. In parallel, our HARA Assistant tool AS-PHALT applies the same principles beyond hazard identification. It structures interaction by analysis phase and supports pre-assessment of severity, exposure, probability of occurrence, and avoidability to derive Safety Integrity Levels (SIL).

We evaluate the approaches on a limited collaborative manufacturing use case. The results indicate that the structured, RAG-supported workflow improves the quality and structure of identified hazards compared to an off-the-shelf LLM baseline and reduces expert effort relative to traditional workshops. However, it still falls short of expert-only completeness, confirming the need for human oversight and iterative review. We outline practical implications for industrial adoption, including open questions like confidence measures for AI-based tools, their qualification and certification, and their safe integration into safety engineering processes.