LLM PITFALLS

SEVERITY 👇 LOW ▼ ● ▲ HIGH

					0.0%	31.0%	58.6%	48.3%	31.0%	31.0%	58.6%	41.4%	37.9%
CATEGORY	PITFALL		MITIGATION	MEAN TASK PERFORMANCE	GPT 3.5	GPT 4	O4-MINI	GPT 4.1	COPILOT FOR WORK	CLAUDE 3.5	CLAUDE 4	MISTRAL MEDIUM	SONAR
Design	Using generative AI as a first resort	▲	Consider more explainable, scrutable tech	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	No baseline or state-of-the-art performance	●	Measure SOTA performance	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	No success criterion	●	Ask user/customer what success means	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
Instruction following	Refusal to answer on unreasonable grounds	▼	Add context, remove ambiguity	67%	Fail	Pass	Pass	Pass	Fail	Fail	Pass	Pass	Pass
	Misinterpretation, failure to clarify	▲	Instruct to clarify, add context, be explicit	67%	Fail	Pass	Pass	Pass	Pass	Fail	Pass	Fail	Pass
	Failure to reject false or inconsistent query	●	Avoid incorrect or inconsistent prompts	44%	Fail	Fail	Fail	Fail	Fail	Pass	Pass	Pass	Pass
	Arbitrary execution of instructions in data	▲	Explicitly separate instructions and data	11%	Fail	Fail	Fail	Fail	Fail	Fail	Pass	Fail	Fail
	Difficulty manipulating sub-token elements	●	Pre-process sub-token elements	11%	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Fail
	Incorrectly formatted output	●	Provide format schema or template	11%	Fail	Fail	Fail	Fail	Fail	Fail	Pass	Fail	Fail
Ambiguity & coherence	Ambiguity in the response	▲	Prompt engineering, pre-check responses	11%	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Fail
	Incoherent rambling or glitching	▼	Reword prompt and retry	56%	Fail	Fail	Pass	Pass	Pass	Not evaluated	Pass	Pass	Fail
	Inconsistent reasoning across conversation	●	Shorten conversations, provide recaps	78%	Fail	Pass	Pass	Pass	Pass	Not evaluated	Pass	Pass	Pass
	Irrelevance	▼	Prompt engineering, pre-check responses	78%	Fail	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Fail
Response quality	Simplistic (correct but poor) responses	●	Role allocation, few-shot prompts	78%	Fail	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Pass
	Incompleteness	●	Role allocation, few-shot prompts	11%	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Fail
	Faulty reasoning	▲	Chain-of-thought, tools, few-shot prompts	67%	Fail	Pass	Pass	Pass	Fail	Pass	Pass	Pass	Fail
	Faulty premises	▲	Explicit chain-of-thought, break steps down	11%	Fail	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail
	Prompt sensitivity	●	Add more context and more instruction	56%	Fail	Fail	Pass	Pass	Fail	Pass	Pass	Pass	Fail
	Overthinking	●	Provide more context about problem domain	11%	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Fail
	Overfit to similar task	▼	Emphasize differences from overfitted case	11%	Fail	Fail	Fail	Fail	Fail	Fail	Pass	Fail	Fail
	Distraction	▼	Use neutral wording	33%	Fail	Fail	Fail	Pass	Pass	Fail	Fail	Pass	Fail
	Verbose	▼	Specify minimal requirements	0%	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
Factual errors	Incorrect factual recall	▲	RAG, knowledge graphs, pre-check response	33%	Fail	Fail	Pass	Pass	Fail	Fail	Pass	Fail	Fail
	Fabricated or non-useful references	▲	RAG, knowledge graphs	11%	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Pass
	Coherent fiction	▲	RAG, knowledge graphs, pre-check response	33%	Fail	Fail	Fail	Fail	Fail	Pass	Pass	Fail	Pass
	Out-of-date responses	●	RAG, tools	33%	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Pass	Pass
	Overconfidence in certainty	●	Disallow expressions of uncertainty	11%	Fail	Fail	Pass	Fail	Fail	Fail	Fail	Fail	Fail
Bias & ethics	Explicit bias	▲	Improve alignment and system prompt	78%	Fail	Fail	Pass	Pass	Pass	Pass	Pass	Pass	Pass
	Implicit bias	▲	Improve alignment, avoid leading signals	11%	Fail	Fail	Fail	Fail	Fail	Fail	Pass	Fail	Fail
	Toxic or offensive responses	▲	Pre-check response	11%	Fail	Fail	Fail	Fail	Fail	Fail	Pass	Fail	Fail
	Copyright infringement	▲	Knowledge graphs	78%	Fail	Pass	Pass	Pass	Pass	Pass	Pass	Fail	Pass
	Cultural or ideological bias	●	Fine-tuning, alignment	78%	Fail	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Pass
Application	No transparency about role of AI	▲	Be clear	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	Not collecting signals from users	▼	Collect signals	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	Not providing references	●	Add references	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
Implementation	No rigorous evaluation protocol	▲	Create testing and reporting pipeline	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	Low AI literacy among users	●	Education and culture	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	Overreliance on model responses	▲	Education, avoiding risky applications	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	Overreliance on human-in-the-loop	●	Consultation, training, improved interface	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	No user training	●	Training	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated
	No consideration of security or ethics	▲	Proper governance and oversight	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated

CC BY Matt Hall & contributors 2025 // Re-use & adapt

mtha@equinor.com

A highly opinionated and non-exhaustive list of pitfalls aimed at developers of applications containing large language models,. There are more than 30 concerns here, but some of them overlap, and probably still others are missing entirely. This list is my own opinion, which is not necessarily shared by others at Equinor or anywhere else. It is a work in progress, you input will be welcomed and credited, please get in touch!

Abbreviations: RAG means 'retrieval augmented generation'.