



Design	CATEGORY	PITFALL		MITIGATION
No success criterion Ask user/customer what success means	Design	Using generative AI as a first resort		Consider more explainable, scrutable tech
Refusal to answer on unreasonable grounds Misinterpretation, failure to clarify Avoid incorrect or inconsistent prompts Arbitrary execution of instructions in data Difficulty manipulating sub-token elements Incorrectly formatted output Ambiguity & Incorrectly formatted output Ambiguity in the response Incoherence Inconsistent reasoning across conversation Irrelevance Simplistic (correct but poor) responses Incompleteness Faulty reasoning Prompt engineering, pre-check responses Faulty reasoning Prompt engineering, pre-check responses Incompleteness Faulty reasoning Prompt engineering, pre-check responses Faulty reasoning Prompt engineering, pre-check responses Response quality Overthinking Distraction Verbose Incorrect factual recall Fabricated or non-useful references Coherent fiction ARAG, knowledge graphs, pre-check response Factual errors Coherent fiction Out-of-date responses Deveronfidence in certainty Explicit bias Implicit bias Implicit bias Implicit bias No transparency about role of Al Application Not providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on human-in-the-loop No user training No user training Verbouse training Verbouse frequences ARAG, cols Coherent fiction ARAG doels ARAG cols Coherent fiction ARAG cols ARAG cols Coherent fiction A		No baseline or state-of-the-art performance	•	Measure SOTA performance
Instruction Failure to clarify Failure to reject false or inconsistent query Avoid incorrect or inconsistent prompts Arbitrary execution of instructions in data Difficulty manipulating sub-token elements Pre-process sub-token elements Prompt engineering, pre-check responses Prompt engineering, pre-check responses Robert Prompt engineering, pre-check responses Robert Prompt engineering, pre-check responses Prompt engineering, pre-check responses Prompt engineering, pre-check response Robert Prompt engineering, pre-check response Robert Prompt engineering, pre-check response Robert Prompt engineering, pre-check response Prompt engineering, pre		No success criterion	•	Ask user/customer what success means
Instruction following		Refusal to answer on unreasonable grounds		Add context, remove ambiguity
Arbitrary execution of instructions in data Difficulty manipulating sub-token elements Pre-process sub-token elements Provide format schema or template Provide more context and more instruction Provide more context and more instruction Provide more context and more instruction Provide more context about problem domain Distraction Provide more context about problem domain Provide more context about probl		Misinterpretation, failure to clarify		Instruct to clarify, add context, be explicit
Difficulty manipulating sub-token elements Incorrectly formatted output Ambiguity & coherence Incorrectly formatted output Ambiguity in the response Incoherent rambling or glitching Inconsistent reasoning across conversation Irrelevance Simplistic (correct but poor) responses Incompleteness Incompleteness Incompleteness Incompleteness Incompleteness Incompleteness Faulty reasoning Prompt engineering, pre-check responses Response Faulty reasoning Faulty premises Application Verbose Factual errors Coherent factual recall Fabricated or non-useful references Overconfidence in certainty Incorrect factual responses Implicit bias Implicit bias Implicit bias Incorrect or femsive responses Copyright infringement Cultural or ideological bias Not transparency about role of Al Application Veribose Informatical Applications Overreliance on model responses Application Coverediance on culture responses Application Not collecting signals from users Overrectiance on model responses Acquired for incorrect account on provide recaps Acquired for incorrect factoring incorrect on a collect incorrect factoring pipeline Consultation, varieting, improved interface Implementation Overreliance on human-in-the-loop Nouser training Training Training Provide format schema or template Prompt engineering, pre-check responses Reword prompt engineering, pre-check responses Reword prompt engineering, pre-check responses Reword prompt engineering, pre-check responses Application, solidion, pre-check responses Application App		Failure to reject false or inconsistent query	•	Avoid incorrect or inconsistent prompts
Incorrectly formatted output Ambiguity & Ambiguity in the response Ambiguity & Coherence Incoherent rambling or glitching Incoesistent reasoning across conversation Irrelevance Simplistic (correct but poor) responses Incompleteness Incomplete		Arbitrary execution of instructions in data		Explicitly separate instructions and data
Ambiguity & coherence Incoherent rambling or glitching Incoherent rambling or glitching Incoherent rambling or glitching Inconsistent reasoning across conversation Irrelevance Vertex poor) responses Role allocation, few-shot prompts Incompleteness Relative reasoning Response Relative reasoning Response Relative reasoning Response Responses Relative reasoning Response Responses Response Relative reasoning Response Responses Response Responses Response Respons		Difficulty manipulating sub-token elements	•	Pre-process sub-token elements
Incoherent rambling or glitching Veword prompt and retry		Incorrectly formatted output	•	Provide format schema or template
Inconsistent reasoning across conversation Irrelevance Simplistic (correct but poor) responses Role allocation, few-shot prompts		Ambiguity in the response		Prompt engineering, pre-check responses
Irrelevance Simplistic (correct but poor) responses Incompleteness Incompleteness Incompleteness Incompleteness Faulty reasoning Faulty premises Faulty premises Quality Prompt sensitivity Add more context about problem domain Distraction Verbose Verbose Verbose Verbose Factual errors Factual errors Bias & ethics Copyright infringement Cultural or ideological bias Application Application No transparency about role of Al Application Irrelevance Simplistic (correct but poor) responses A Role allocation, few-shot prompts A Chain-of-thought, tools, few-shot prompts A Explicit chain-of-thought, break steps down Add more context and more instruction Overthinking Provide more context about problem domain Verbose Ve		Incoherent rambling or glitching		Reword prompt and retry
Simplistic (correct but poor) responses Role allocation, few-shot prompts		Inconsistent reasoning across conversation	•	Shorten conversations, provide recaps
Incompleteness Faulty reasoning Achain-of-thought, tools, few-shot prompts Faulty premises quality Faulty premises Prompt sensitivity Overthinking Distraction Verbose Verbose Verbose Verbose Factual errors Factual errors Factual errors Factual errors Taxic or offensive responses Copyright infringement Cultural or ideological bias Application Application Not ransparency about role of Al Application Application Application Application Application A Role allocation, few-shot prompts Achain-of-thought, broak steps down Add more context and more instruction Achad more context about problem domain Verbose		Irrelevance		Prompt engineering, pre-check responses
Faulty reasoning		Simplistic (correct but poor) responses	•	Role allocation, few-shot prompts
Faulty premises Add more context and more instruction		Incompleteness	•	Role allocation, few-shot prompts
Prompt sensitivity Overthinking Distraction Verbose Incorrect factual recall Factual errors Coherent fiction Overconfidence in certainty Explicit bias Implicit bias Implicit or offensive responses Copyright infringement Cultural or ideological bias Not providing references Not providing references Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on human-in-the-loop No user training Provide more context and more instruction Provide more context about problem domain Provide more context and more instruction Provide more context about problem domain instruction Provide more context about problem domain Provide more context about problem domain instruction Provide more context about problem domain Provide more context about problem domain instruction Provide more con		Faulty reasoning		Chain-of-thought, tools, few-shot prompts
Overthinking Distraction Verbose Incorrect factual recall Factual errors Coherent fiction Overconfidence in certainty Explicit bias Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias Not providing references No transparency about role of Al Application Not collecting signals from users No rigorous evaluation protocol Low Al literacy among users Overreliance on human-in-the-loop No user training Verbose V Specify minimal requirements A RAG, knowledge graphs, pre-check response A RAG, knowledge graphs, pre-check response RAG, tools Out-of-date responses Overreliance on human-in-the-loop No user training Provide more context about problem domain Vuese neutral wording A RAG, knowledge graphs, pre-check response RAG, tools Disallow expressions of uncertainty Improve alignment and system prompt Improve alignment and system prompt Improve alignment, avoid leading signals Pre-check response A Knowledge graphs Fine-tuning, alignment A Be clear Collect signals Create testing and reporting pipeline Education and culture Overreliance on model responses A Education, avoiding risky applications Consultation, training, improved interface Training		Faulty premises		Explicit chain-of-thought, break steps down
Distraction Verbose Verbose Incorrect factual recall Fabricated or non-useful references Coherent fiction Out-of-date responses Overconfidence in certainty Explicit bias Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias No transparency about role of Al Application Not collecting signals from users No providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on human-in-the-loop No user training Incorrect factual recall RAG, knowledge graphs, pre-check response RAG, tools Disallow expressions of uncertainty Disallow expressions of uncertainty Improve alignment and system prompt Improve alignment, avoid leading signals Fre-check response Knowledge graphs Fine-tuning, alignment Be clear Collect signals Add references Add references Add references Create testing and reporting pipeline Education, avoiding risky applications Overreliance on model responses Overreliance, no fine first proved interface Training Training		Prompt sensitivity	•	Add more context and more instruction
Verbose Incorrect factual recall Fabricated or non-useful references Factual errors Coherent fiction Out-of-date responses Overconfidence in certainty Explicit bias Implicit bias Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias No transparency about role of AI Application Not collecting signals from users No rigorous evaluation protocol Low AI literacy among users Overreliance on human-in-the-loop No user training RAG, knowledge graphs, pre-check response RAG, tools Disallow expressions of uncertainty Improve alignment and system prompt Improve alignment, avoid leading signals A pre-check response Copyright infringement A knowledge graphs Fine-tuning, alignment Be clear Collect signals Add references Add references Create testing and reporting pipeline Education, avoiding risky applications Consultation, training, improved interface No user training Training		Overthinking	•	Provide more context about problem domain
Incorrect factual recall		Distraction		Use neutral wording
Factual errors Coherent fiction Out-of-date responses Overconfidence in certainty Explicit bias Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias Not collecting signals from users Not providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on human-in-the-loop No user training RAG, knowledge graphs RAG, knowledge graphs, pre-check response of uncertainty Implication RAG, knowledge graphs, pre-check response of uncertainty Improve alignment and system prompt Improve alignment, avoid leading signals Pre-check response Knowledge graphs, pre-check response Knowledge graphs, pre-check response Fine-tuning, alignment Be clear Collect signals Create testing and reporting pipeline Education and culture Overreliance on human-in-the-loop No user training Training		Verbose	\blacksquare	Specify minimal requirements
Factual errors Coherent fiction Out-of-date responses Overconfidence in certainty Explicit bias Implicit bias Improve alignment and system prompt Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training RAG, knowledge graphs, pre-check response RAG, knowledge graphs, pre-check response of uncertainty Improve alignment and system prompt Improve alignment, avoid leading signals Fre-check response Knowledge graphs Fine-tuning, alignment Be clear Collect signals Add references Add references Education and culture Education, avoiding risky applications Consultation, training, improved interface Training	Factual errors	Incorrect factual recall		RAG, knowledge graphs, pre-check response
Out-of-date responses Overconfidence in certainty Explicit bias Improve alignment and system prompt Implicit bias Improve alignment, avoid leading signals Improve alignment, avoid leading signals Pre-check response Copyright infringement Cultural or ideological bias Fine-tuning, alignment No transparency about role of AI Be clear Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training RAG, tools Disallow expressions of uncertainty Improve alignment and system prompt Fine-check response Knowledge graphs Fine-tuning, alignment Be clear Collect signals Create testing and reporting pipeline Education and culture Consultation, avoiding risky applications Overreliance on human-in-the-loop Training Training		Fabricated or non-useful references		RAG, knowledge graphs
Overconfidence in certainty Explicit bias Implicit bias Implicit bias Improve alignment and system prompt Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias No transparency about role of AI Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Improve alignment and system prompt Improve alignment and system proded in protocol Improve alignment and system		Coherent fiction		RAG, knowledge graphs, pre-check response
Explicit bias Improve alignment and system prompt Implicit bias Toxic or offensive responses Copyright infringement Cultural or ideological bias No transparency about role of AI Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Improve alignment and system prompt Improve alignment, avoid leading signals Pre-check response Fine-tuning, alignment Oclect signals Collect signals Add references Add references Education and reporting pipeline Education and culture Consultation, avoiding risky applications Consultation, training, improved interface Training		Out-of-date responses	•	RAG, tools
Implicit bias		Overconfidence in certainty	•	Disallow expressions of uncertainty
Bias & ethics Toxic or offensive responses Copyright infringement Cultural or ideological bias No transparency about role of Al Application Not providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Pre-check response Knowledge graphs Fine-tuning, alignment Be clear Collect signals Add references Add references Create testing and reporting pipeline Education and culture Education, avoiding risky applications Consultation, training, improved interface Training	Bias & ethics	Explicit bias		Improve alignment and system prompt
Copyright infringement Cultural or ideological bias No transparency about role of AI Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Knowledge graphs Fine-tuning, alignment Be clear Collect signals Add references Add references Education and reporting pipeline Education and culture Consultation, avoiding risky applications Training		Implicit bias		Improve alignment, avoid leading signals
Cultural or ideological bias No transparency about role of AI Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Fine-tuning, alignment Be clear Collect signals Add references Create testing and reporting pipeline Education and culture Education, avoiding risky applications Consultation, training, improved interface		Toxic or offensive responses		Pre-check response
No transparency about role of AI Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training No transparency about role of AI Be clear Collect signals Add references Create testing and reporting pipeline Education and culture Education, avoiding risky applications Consultation, training, improved interface		Copyright infringement		Knowledge graphs
Application Not collecting signals from users Not providing references No rigorous evaluation protocol Low Al literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Not collect signals Add references Add references Education and reporting pipeline Education and culture Consultation, avoiding risky applications Consultation, training, improved interface Training		Cultural or ideological bias	•	Fine-tuning, alignment
Not providing references No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Add references Create testing and reporting pipeline Education and culture Education, avoiding risky applications Consultation, training, improved interface Training	Application	No transparency about role of AI		Be clear
No rigorous evaluation protocol Low AI literacy among users Overreliance on model responses Overreliance on human-in-the-loop No user training Create testing and reporting pipeline Education and culture Education, avoiding risky applications Consultation, training, improved interface Training		Not collecting signals from users		Collect signals
Low Al literacy among users		Not providing references	•	Add references
Overreliance on model responses Overreliance on human-in-the-loop No user training Description Consultation, avoiding risky applications Consultation, training, improved interface Training	Implementation	No rigorous evaluation protocol		Create testing and reporting pipeline
Implementation Overreliance on human-in-the-loop Consultation, training, improved interface No user training Training		Low Al literacy among users	•	Education and culture
Overreliance on human-in-the-loop No user training Consultation, training, improved interface Training		Overreliance on model responses		Education, avoiding risky applications
		Overreliance on human-in-the-loop	•	Consultation, training, improved interface
No consideration of security or ethics A Proper governance and oversight		No user training	•	Training
		No consideration of security or ethics		Proper governance and oversight

CC BY Matt Hall & contributors 2024 $\,/\!/\,$ Re-use & adapt

ntha@equinor.con

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!