# Appendix – Non-exhaustive list of risk factors in relation to the use of generative AI language models

The use of AI LMs involves various risks. Whilst multiple functionalities or risk mitigation measures can be adopted to enhance the utility, performance or safety of an AI LM, LCs should beware of potential limitations to their effectiveness. The non-exhaustive list of risk factors set out below should be read in conjunction with Core Principles 1 – 4 in the circular.

## Core Principle 1: Senior management responsibilities

*Over-reliance*

1. Researchers have found that whilst AI LMs can provide large productivity benefits, when some management consultants use AI LMs on complex tasks outside the AI's capabilities, a higher percentage of workers provide inaccurate recommendations than workers in the same organisation without AI support.

*Agentic AI*

2. A recent development is the integration of AI LMs into agentic AI workflows, where an AI LM is augmented with or incorporated into software programmes having the ability to make step-by-step plans and take sequences of actions (including using external tools) to pursue specific goals and accomplish a variety of multi-step digital tasks with limited direct human supervision. While such deployment is not always high-risk if properly designed and implemented, agentic AI currently presents significant governance challenges[1], for example, in striking the right balance between autonomy and control. LCs should therefore exercise extra caution when assessing the use of agentic AI LMs.

## Core Principle 2: AI model risk management

3. When validating AI LMs, monitoring and testing their performance, LCs should beware of the following issues:

   a. There may be multiple causes of hallucination. One possible reason is related to the inherent nature of the underlying AI LMs and how these are trained to generate outputs based on probabilities. Developed as general-purpose models to perform a variety of tasks, they are not necessarily designed with the primary purpose to generate factually accurate output. Another possible contributor to hallucination is noise in the training data such as inconsistent statements or ambiguous context.

   b. When assessing data quality, LCs should beware that biases may arise during:

      i. the data collection process when choosing what data to include (selection bias). Additionally, biases may arise from the way that questions are framed, ie, asking questions phrased in a certain way or sequence may elicit biased answers from respondents; and

[1] See for example Open AI, *Practices for Governing Agentic AI Systems*.

ii. the data cleansing and data processing stages when labelling the data, addressing gaps in the data or attempting to improve the signal to noise ratio within the data.

c. It has been shown that word embeddings in natural language processing can pick up racial and gender biases, and some debiasing techniques may mask and not completely remove biases in word embeddings. Biases embedded in the model may be more difficult to identify, and therefore extra care may be necessary when deploying AI LMs as components for classification.

d. As AI LMs may generate unsubstantiated output concerning other entities, LCs should beware of the reputational and legal risks associated with AI LMs outputting adverse comments about competitors.

e. Any apparent commitment made to users which is generated by an LC's public facing AI LM may expose the LC to legal risk even if the content is fabricated[2].

f. Hallucination in some solutions marketed as "eliminating" or "avoiding" hallucination is found to remain wide-ranging. LCs choosing to deploy such solutions should therefore critically review their efficacy.

g. Model drift has been observed in the performance of commercially available AI LMs in relation to, for example, mathematical questions. Even the same version of some commercially available AI LMs exhibited model drift.

h. AI LMs exhibit cognitive biases which are observed in human decision making, such as the framing effect (where people make a choice based on whether the options are presented with positive or negative connotations) and status quo bias (where an option presented as the status quo is more likely to be selected).

i. Even when AI LMs have been designed to be harmless and honest, misaligned behaviour such as AI deception has also been identified.

*Prompt engineering and content filtering*

4. Prompt engineering may reduce hallucination and cybersecurity risks by using system prompts to provide the AI LM with instructions on how to behave and respond to user enquiries, such as defining the types of questions it is allowed to answer and preventing it from replying to a question when it does not have the information to do so. However, users may still need to (a) test end-to-end performance after any prompt engineering to ensure desired behaviour, and (b) validate the model responses[3] as the response generated may still contradict the intent of the instruction in the system prompt[4].

5. Furthermore, instruction drift has been noted, ie, system prompts may not be stable over multiple turns of dialogue. As such, guardrails implemented through system prompts may degrade over the course of user interaction.

[2] In *Moffatt v. Air Canada*, the British Columbia Civil Resolution Tribunal found Air Canada liable for misinformation given to a consumer by an AI LM on its website.

[3] As Microsoft has reminded users, it is important that even when using prompt engineering effectively, users still need to validate the responses the models generate. Just because a carefully crafted prompt worked well for a particular scenario does not necessarily mean it will generalise more broadly to certain use cases.

[4] Microsoft has reminded users, even if users instruct a model in the system message to answer "**I don't know"** when unsure of an answer, this does not guarantee that the request will be honoured.

6. Whilst inputting high-quality, unambiguous user prompts may help reduce hallucination risk, LCs should beware that some AI LMs may exhibit recency bias, where the order or position of the information supplied within the user prompt affects model performance.

7. Content filtering may also enhance model performance and safety. For example, an input filter may screen out user instructions which are manipulative, malicious or outside the domain for which the AI LM is designed to be used, or an output filter may screen the output for inaccuracies or inappropriate language before providing a response to the user.

8. Whilst AI LMs may have multilingual capabilities, some models are mainly trained in the English language. Where content filtering and prompt engineering are implemented in an AI LM for use in multiple languages, their effectiveness may need to be tested for different languages.

*RAG*

9. Whilst RAG may help reduce hallucination, information integration poses a challenge for RAG. RAG may still produce inaccurate output when (i) the answer requires input from multiple sources and not all relevant sources are retrieved or appropriately prioritised, (ii) different factual data from longer text is not picked up correctly, or (iii) some retrieved sources are not effectively consolidated into the answer[5]. As such, it may be necessary to separately test the various components of the LC's implementation of RAG, such as retrieval accuracy and generation quality.

10. Citing the sources of the information to show the provenance of the data used to generate the response would enable the user to verify and fact-check the information outputted, but citation accuracy is not guaranteed. Some AI LMs have generated fake or non-existent citations.

**Core Principle 3: Cybersecurity and data risk management**

11. The trend of AI LMs having larger context windows can support more use cases. However, enabling users to input more information into the model may also offer increased scope for adversarial attacks through crafting long, complex user prompts.

12. System prompts which constrain the behaviour of AI LMs may still be at risk of being circumvented via direct and indirect prompt injection attacks. These attacks may affect output integrity or result in data leakage[6].

13. Retaining memory of a user's previous prompts may improve user experience for particular use cases. However, as the length of conversation with an AI LM chatbot increases, the opportunity for jailbreak attacks also increases, as pre-set system prompt guardrails may be diluted or distracted through multi-turn dialogue[7]. Whilst good system prompt design may be able to mitigate the impact of simpler attacks, additional risk mitigation measures may be necessary to defend against multi-turn jailbreak attacks[8].

[5] Further, complex questions have higher error rates, and AI LMs may not handle formatted tabular data well.
[6] As an example, an attacker with access to Slack AI was able to prompt Slack AI to exfiltrate data from private channels even if the attacker was not a member of the private channels.
[7] Anthropic, *Many-shot jailbreaking*.
[8] See for example the Crescendo attack outlined by Microsoft.

14. Thefts of users' sensitive information, such as facial images and fingerprints, using model-inversion and hill-climbing attacks against biometric systems have been reported before. Such attacks targeting AI LMs have also been successful. Membership inference, which is another type of attack, does not recover the training data, but, instead, recovers information about whether or not a particular individual was in the training set. In addition, recent research managed to circumvent privacy safeguards and extract a portion of an AI LM's training data by prompting it to repeat certain words. Moreover, data poisoning attacks attempt to trick an AI LM by maliciously manipulating the training data so that it learns to output the wrong result based on the tainted training dataset. Attackers can activate the backdoor by sending specific input to the corrupted model to obtain the desired (manipulated) output.

15. Although privacy enhancing techniques may protect AI LMs against data exfiltration, privacy side channel attacks are found to be possible when content filters are introduced to machine learning models.

16. The terms of use of some commercial AI LMs allow the Third Party Provider to use information supplied by the user for its own purposes (such as further training the AI LM to improve model performance). There have been reports that confidential information supplied by users in prompts to the AI LM were visible to other unrelated users of the same AI LM. As such, LCs using a Third Party Provider's AI LM should ascertain whether any information it supplies to the AI LM would be used by the Third Party Provider for its own purposes, whether such information is in the form of training data to fine-tune the AI LM or information included in the context window when prompting the AI LM. The LCs should also take appropriate measures to ensure the confidentiality of the relevant information.

17. When considering data privacy risks, LCs should beware that the applicable personal data privacy laws may include laws in the jurisdiction where personal data is collected or processed.

## Core Principle 4: Third Party Provider risk management

### *Contractual terms of use*

18. LCs should beware of any use restrictions stipulated in the terms of use for the Third Party Provider's AI LM. For example, the user may be prohibited from using the service to develop machine learning models or related technology[9].

19. AI LMs provided by Third Party Providers may require certain disclosure to users. For example, according to Anthropic's Commercial Terms of Service (effective 4 March 2024), the customer acknowledges, and must notify its users, that users should not rely upon factual assertions in output without independently checking their accuracy, as they may be false, incomplete, misleading or not reflective of recent events or information.

[9] See for example *Google's Generative AI Additional Terms of Service*, August 9, 2023 version.