Preventing AI Hallucinations with Effective User Prompts

Publication Date: 2025-12-02

Contents

WHAT?

AI hallucinations occurs when an LLM generates information that is not based on real-world facts or evidence. This can include fictional events, incorrect data or irrelevant outputs.

WHY?

Learn to create effective prompts that can help AI generate accurate and reliable content.

EFFORT

Less than 15 minutes of reading.

1 What causes AI hallucinations?

- **Ambiguous prompts.** Vague queries can lead to random or inaccurate answers.
- **Lack of clear context.** When the language model lacks context, it can fabricate answers.
- **Long generation length.** The longer the generated response, the higher the chance that hallucinations can happen.
- **No retrieval-augmented process.** LLMs without access to external sources—such as databases or search engines—can produce errors when they need to generate specific information.

2 How can I prevent AI from generating hallucinations?

2.1 Set clear expectations

The clearer the prompt, the less the LLM relies on assumptions or creativity. A well-defined prompt guides the model toward specific information, reducing the likelihood of hallucinations.

TECHNIQUES:

- Use **specific language** that guides the model.
- Focus on **known data sources** or real events.
- Request **summaries** or *paraphrasing* from established sources.

EXAMPLE

- **Ambiguous prompt:**“Tell me about space.”
- **Clearer prompt:**“Give me a summary of NASA’s recent Mars missions, including factual details from their official reports.”

EXAMPLE

- **Ambiguous prompt:**“What is quantum computing?”
- *Clearer prompt: *“Explain the basic principles of quantum computing, specifically how qubits work compared to classical bits.”

2.2 Break down complex prompts

Break down complex or broad prompts into manageable pieces. This keeps the language model focused on a narrower scope and reduces the chance of hallucination.

EXAMPLE

- **Complex query:**“Explain AI and how it can change the world.”
- **Broken down prompt:**“What are the most recent advancements in AI? How are these advancements being applied in the healthcare industry?”

2.3 Use retrieval-augmented generation (RAG)

When crafting prompts, encourage the model to retrieve relevant information instead of generating from scratch. Integrating a RAG system allows the LLM to query a specific database or resource.

TECHNIQUES

- Include context cues, for example, “Based on the following document” or “From the official Web site” to point the model toward facts.
- If using a tool like Milvus or ChromaDB, structure your prompt to refer to specific collections or documents. This reduces hallucination by grounding the LLM in real data.

EXAMPLE

- **Prompt without RAG:**“Tell me about the company’s AI products.”
- **Prompt with RAG:**“Based on the "`technical-info” collection in Milvus, provide details about the company’s AI product line.`"

2.4 Constrain the output

Limit the length or scope of the language model's response. Shorter, more direct answers reduce the chances of the model drifting off-topic or hallucinating extra details.

TECHNIQUE

- Use *tokens* or *word limits* where possible to enforce the output length.

EXAMPLE

- **Unconstrained prompt:**"Give me a detailed report on quantum mechanics."
- **Prompt with limited output:**"In 100 words or fewer, explain the main concept of quantum entanglement."

2.5 Prompt for verification

You can structure prompts to ask the LLM for clarification or to cite the source of its statements. This leads the model to produce more grounded and reliable responses.

EXAMPLES

- "Where did you find this information?"
- "Verify this answer against known historical facts about the event."

2.6 Use chain-of-thought (CoT) prompting

By guiding the model through logical steps, you can control the reasoning path and help the model arrive at accurate conclusions. This method is especially helpful when asking the model to explain complex processes.

EXAMPLE

- *Step-by-step prompt: *"Explain the following concepts step by step: 1. How do neural networks learn from data? 2. How is backpropagation used in this process?"

2.7 Use templates for complex tasks

For complex tasks, for example, answering requests for proposals or technical questions, templates help provide a structure that minimizes hallucinations. This is achieved by making the desired format and content explicit.

EXAMPLE

- "Based on the document provided, summarize the key technical features of the product. Format the response as: 1. Feature, 2. Benefit, 3. Use case. Use only factual information."

2.8 For more information

- Find good examples of system prompts in link:https://documentation.suse.com/suse-ai/1.0/html/AI-system-prompts/index.html.

3 Legal Notice

Glossary

AI, artificial intelligence

Refers to the simulation of human intelligence in machines that are designed to learn and solve problems like humans. Enables computers to understand language, make decisions and improve from experience.

Air gap

A security measure where a computer network is physically isolated from unsecured networks, including the public Internet.

Batch size

The number of samples processed simultaneously during model inference, affecting processing speed and resource utilization.

BYOC, bring your own certificate

A practice allowing users to provide their own SSL/TLS certificates for securing communications instead of using default or auto-generated ones.

CA, certification authority

An entity that issues digital certificates to verify the identity of certificate holders and ensure secure communications.

Chain-of-thought (CoT) prompting

A prompting technique that guides AI models to break down complex problems into step-by-step reasoning processes, improving response accuracy and transparency.

Chat template

A structured format for organizing conversations between users and AI models, defining how system prompts, user inputs, and AI responses are formatted and processed.

Context window

The maximum amount of text (tokens) that an AI model can process at once, including both the input prompt and generated response.

CRD, custom resource definitions

Extensions of the Kubernetes API that allow users to define custom resources and their controllers in a Kubernetes cluster.

CUDA, Compute Unified Device Architecture

NVIDIA's parallel computing platform and programming model used to accelerate AI workloads on GPU hardware.

Data leakage

The unintended exposure of sensitive information through AI model responses, potentially compromising data security and privacy.

Embeddings

Numerical representations of data (text, images, etc.) in a high-dimensional space that capture semantic relationships and enable AI models to process information effectively.

Fine-tuning

The process of further training a pre-trained AI model on specific data to adapt it for particular tasks or domains, improving its performance for targeted applications.

GenAI, generative AI

A type of artificial intelligence that can create new content such as text, images or music.

GPU, graphics processing unit

Specialized hardware designed for parallel processing. In AI applications, GPUs accelerate model training and inference tasks.

Hallucination

An AI behavior where the model generates false or unsupported information that appears plausible but has no basis in provided context or real facts.

Helm

A package manager for Kubernetes that helps install and manage applications. Helm uses charts to define, install and upgrade complex Kubernetes applications.

Helm chart

A package format for Kubernetes applications that contains all resource definitions needed to deploy and configure application workloads.

IaC, infrastructure as code

The practice of managing and provisioning infrastructure through machine-readable definition files rather than manual processes.

Inference

The process of using a trained AI model to make predictions or generate outputs based on new input data.

Kubernetes pods

The smallest deployable units in Kubernetes that can host one or more containers, sharing networking and storage resources.

LLM, large language model

An advanced AI model trained on amounts of text data to understand and generate human-like text. Can perform tasks like translation, summarization and answering questions.

Model weights

The learned parameters of an AI model that determine how it processes inputs and generates outputs. These weights are adjusted during training to optimize model performance.

NLG, natural language generation

A process of automatically generating human-like text from structured data or other forms of input. Designed to convert raw data into coherent and meaningful language easily understood by humans.

NLU, natural language understanding

A process AI uses to analyze and understand the meaning of the input query.

NVIDIA GPU driver

Software that enables communication between the operating system and NVIDIA graphics hardware, essential for GPU-accelerated AI workloads.

NVIDIA GPU Operator

A Kubernetes operator that automates the management of NVIDIA GPUs in container environments, handling driver deployment, runtime configuration, and monitoring.

Ollama

An open source framework for running and serving AI models locally. Ollama simplifies the process of downloading, running and managing large language models.

OpenGL

A cross-platform API for rendering 2D and 3D graphics, commonly used in visualization applications and GPU-accelerated computing.

Prompt Engineering

The practice of crafting effective input queries to AI models to obtain desired and accurate outputs. Good prompt engineering helps prevent hallucinations and improves response quality.

Prompt injection

A security vulnerability where malicious inputs attempt to override or bypass an AI model’s system prompt or safety constraints.

Quantization

A technique to reduce AI model size and computational requirements by converting model parameters to lower precision formats while maintaining acceptable performance.

RAG, retrieval-augmented generation

A technique that enhances AI responses by retrieving relevant information from a knowledge base before generating answers, improving accuracy and reducing hallucinations.

RBAC, role-based access control

A security model that restricts system access based on roles assigned to users, managing permissions and authorization in Kubernetes clusters.

Semantic search

A search method using AI to understand the meaning and context of queries rather than just matching keywords, enabling more relevant results.

System prompt

Initial instructions given to an AI model that define its behavior, role and response parameters. System prompts help maintain consistent and appropriate AI responses.

Temperature

A parameter controlling the randomness in AI model outputs. Lower values produce more focused and deterministic responses, while higher values increase creativity and variability.

Token

The basic unit of text processing in AI models, representing parts of words, characters or symbols. Models process text by breaking it into tokens for analysis and generation.

Top-K

A parameter that limits token selection during text generation to the K most likely next tokens, helping control output quality and relevance.

Top-P

Also known as nucleus sampling, a parameter that selects from the smallest set of tokens whose cumulative probability exceeds P, providing dynamic control over text generation diversity.

Vector database

A specialized database designed to store and efficiently query high-dimensional vectors that represent data in AI applications, enabling similarity searches and semantic operations.

Vector store

A specialized storage system optimized for managing and querying vector embeddings, essential for semantic search and RAG implementations in AI applications.

A Copyright

B GNU Free Documentation License

B1 0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

B2 1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

B3 2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

B4 3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

B5 4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

1. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
2. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
3. State on the Title page the name of the publisher of the Modified Version, as the publisher.
4. Preserve all the copyright notices of the Document.
5. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
6. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
7. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document’s license notice.
8. Include an unaltered copy of this License.
9. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
10. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

11. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

12. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

13. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

14. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

15. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

B6 5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

B7 6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

B8 7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

B9 8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

B10 9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

B11 1. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See https://www.gnu.org/copyleft/ ↗.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

B12 ADDENDUM: How to use this License for your documents

```
Copyright (c) YEAR YOUR NAME.
```

```
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.2
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with…Texts."" line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.