Human-centred (HCI) xAI - Hallucinations and user trust

Thomas Joyce * **Oisín Frizell** * **Andrew Jaffray** *
Edmund Phelan * **Ruairí Glackin** *

* *Immersive Software Engineering, University of Limerick, Castletroy, Limerick (email: 23367857@studentmail.ul.ie, x@studentmail.ul.ie, x@studentmail.ul.ie, x@studentmail.ul.ie, x@studentmail.ul.ie)*

Abstract: Large Language Models (LLMs) exhibit impressive fluency but are prone to hallucinations—outputs that are coherent yet factually incorrect or fabricated. These hallucinations pose significant risks in high-stakes domains, including healthcare, law, and scientific writing. Hallucinations broadly arise from two sources: prompting-induced errors and model-internal tendencies. Distinguishing these sources is essential for developing effective mitigation strategies. Current approaches, such as RLHF, improve behavior but do not fully eliminate hallucinations. In Human-Computer Interaction (HCI) contexts, hallucinations can directly undermine user trust and decision-making. To investigate this, we benchmarked five LLMs across standardized hallucination evaluation datasets, including TruthfulQA, HallucinationEval, and RealToxicityPrompts. Our study builds on prior analyses while extending evaluation to more models and scenarios. We examine whether recent LLM improvements reduce hallucinations and their impact on user trust. Findings aim to guide both technical mitigation strategies and HCI-informed interface design to maintain reliability and trust.

1. INTRODUCTION

Large Language Models (LLMs) have impressive fluency and ability to perform complex tasks but a critical challenge persists: hallucinations. Hallucinations occur when models generate output that appears coherent and convincing but is factually incorrect, fabricated, or logically inconsistent (Ji et al., 2023; Maynez et al., 2020; Kazemi et al., 2023). Numerous studies have documented this phenomenon, highlighting its prevalence and the risks it poses in high-stakes contexts with serious consequences, such as misdiagnoses in healthcare, fabricated citations in academic writing, and erroneous case references in legal documents [1], [2].

Broadly, hallucinations in LLMs can be divided into two primary sources: (1) Prompting-induced hallucinations, where ill-structured, unspecified, or misleading prompts cause inefficient outputs (Reynolds and McDonell, 2021; Zhou et al., 2022; Wei et al., 2022), and (2) Model-internal hallucinations, which caused by the model's architecture, pretraining data distribution, or inference behavior (Bang and Madotto, 2023; Chen et al., 2023; OpenAI, 2023a). Distinguishing between these two causes is essential for developing effective mitigation strategies.

Currently, there is no definitive method to prevent hallucinations. While RLHF has shown promise in improving model behavior, challenges remain in ensuring the accuracy and reliability of LLM outputs. To address this issue, we are conducting a literature review to study the interaction between LLMs and hallucinations. Our research involves benchmarking 5 different LLMs. We try to reproduce findings of Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior by Hoang, except on more models than they just did. Using standardized hallucination evaluation benchmarks [e.g., TruthfulQA (Lin et al., 2022), HallucinationEval (Wu et al., 2023), RealToxicityPrompts (Gehman et al., 2020)]. The conclusions drawn from this analysis aim to determine whether advancements in LLMs have mitigated the occurrence of hallucinations in previously problematic scenarios, and how this impacts user trust.

2. BACKGROUND AND RELATED WORK

Next we see a few subsections.

2.1 Explainable AI (xAI) in context

For submission guidelines, follow instructions on paper submission system as well as the event website.

Note that conferences impose strict page limits, so it will be better for you to prepare your initial submission in the camera ready layout so that you will have a good estimate for the paper length. Additionally, the effort required for final submission will be minimal.

2.2 Equations

Some words might be appropriate describing equation (1), if we had but time and space enough.

$$\frac{\partial F}{\partial t} = D\frac{\partial^2 F}{\partial x^2} \tag{1}$$

See Able (1956), Able et al. (1954), Keohane (1958), and Powers (1985).

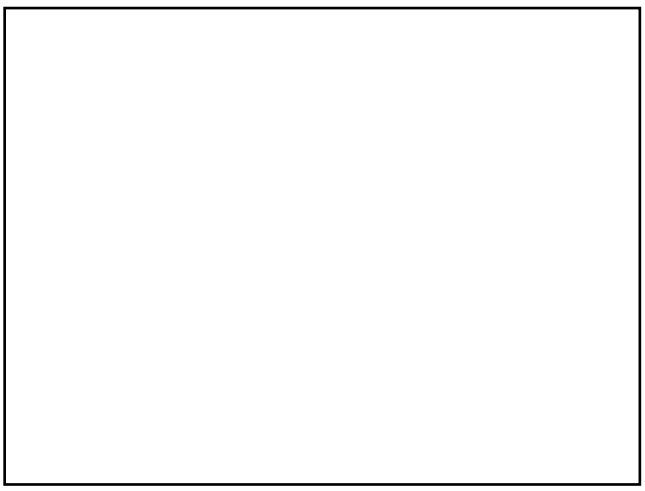

Fig. 1. Bifurcation: Plot of local maxima of x with damping a decreasing

Example. This equation goes far beyond the celebrated theorem ascribed to the great Pythagoras by his followers.

Theorem 1. The square of the length of the hypotenuse of a right triangle equals the sum of the squares of the lengths of the other two sides.

Proof. The square of the length of the hypotenuse of a right triangle equals the sum of the squares of the lengths of the other two sides. □

2.3 Figures

To insert figures, use the `#figure` function. See Fig. 1 for an example which was generated by the following code.

```
#figure(
  image("bifurcation.jpg", width: 8.4cm)
  caption: [Bifurcation: ...]
) <bifurcation>
```

Figures must be centered, and have a caption at the bottom.

2.4 Tables

Tables must be centered and have a caption above them, numbered with Arabic numerals. See Table 1 for an example.

Table 1. Margin settings

Page	Top	Bottom	Left/Right
First	3.5	2.5	1.5
Rest	2.5	2.5	1.5

2.5 Final Stage

Authors are expected to mind the margins diligently. Papers need to be stamped with event data and paginated for inclusion in the proceedings. If your manuscript bleeds into margins, you will be required to resubmit and delay the proceedings preparation in the process.

Page margins. See Table 1 for the page margins specification. All dimensions are in *centimeters*.

2.6 PDF Creation

All fonts must be embedded/subsetted in the PDF file. This is handled by Typst.

2.7 Copyright Form

IFAC will put in place an electronic copyright transfer system in due course. Please *do not* send copyright forms by mail or fax. More information on this will be made available on IFAC website.

3. UNITS

Use SI as primary units. Other units may be used as secondary units (in parentheses). This applies to papers in data storage. For example, write "15 $\mathrm{Gb/cm^2}$ ($100\ \mathrm{Gb/in^2}$)". An exception is when English units are used as identifiers in trade, such as "3.5 in disk drive". Avoid combining SI and other units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation. The SI unit for magnetic field strength $\mathbf{H}$ is A/m. However, if you wish to use units of T, either refer to magnetic flux density $\mathbf{B}$ or magnetic field strength symbolized as $\mu_0\mathbf{H}$. Use the center dot to separate compound units, e.g., "$\mathrm{A \cdot m^2}$".

4. HELPFUL HINTS

4.1 Figures and Tables

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity "Magnetization", or "Magnetization M", not just "M". Put units in parentheses. Do not label axes only with units. For example, write "Magnetization (A/m)" or "Magnetization ($\mathrm{Am^{-1}}$)", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Multipliers can be especially confusing. Write "Magnetization ($\frac{\mathrm{kA}}{\mathrm{m}}$)" or "Magnetization ($\mathrm{10^3 A/m}$)'. Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the axis label means 16000 A/m or 0.016 A/m.

4.2 References

Use Harvard style references (see at the end of this document). With Typst, you can process an external bibliography database in the BibTeX format (`.bib`) or Hayagriva (a Rust-based bibliography management system based on YAML) formats. Footnotes should be avoided as far as possible. Please note that the references at the end of this document are in the preferred referencing style. Papers that have not been published should be cited as "unpublished". Capitalize only the first word in a paper title, except for proper nouns and element symbols.

4.3 Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as IFAC, SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.", not "C. N. R. S." Do not use abbreviations in the title unless they are unavoidable (for example, "IFAC" in the title of this article).

4.4 Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$\int_0^{r_2} F(r,\phi)\mathrm{d}r\mathrm{d}\phi = [\sigma r_2/(2\mu_0)] \cdot \int_0^{\mathrm{inf}} \exp\bigl(-\lambda\ |z_j - z_i|\bigr)\lambda^{-1} J_1(\lambda r_2) J_0(\lambda r_i)\mathrm{d}\lambda \tag{2}$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (T might refer to temperature, but T is the unit tesla). Refer to "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is …".

4.5 Other Recommendations

Use one space after periods and colons. Hyphenate complex modifiers: "zero-field-cooled magnetization". Avoid dangling participles, such as, "Using (1), the potential was calculated" (it is not clear who or what used (1)). Write instead: "The potential was calculated by using (1)", or "Using (1), we calculated the potential".

A parenthetical statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.) Avoid contractions; for example, write "do not" instead of "don't". The serial comma is preferred: "A, B, and C" instead of "A, B and C".

5. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

ACKNOWLEDGMENTS

Place acknowledgments here.

REFERENCES

Able, B. (1956). Nucleic acid content of microscope. *Nature*, *135*, 7–9.

Able, B., Tagg, R., & Rush, M. (1954). Enzyme-catalyzed cellular transanimations. In A. Round (Ed.), *Advances in Enzymology* (3rd ed., Vol. 2, pp. 125–247). Academic Press.

Keohane, R. (1958). *Power and Interdependence World Politics in Transitions*. Little, Brown & Co.

Powers, T. (1985). Is there a way out?. *Harpers*, 35–47.

Appendix A. SUMMARY OF LATIN GRAMMAR

Appendix B. SOME LATIN VOCABULARY