# Fixing Reference Hallucinations of LLMs

November 29, 2024

Stephane H. Maes[1]

**Abstract:**

*In October and November 2024, using popular LLMs like OpenAI ChatGPT (4 and below), Azure OpenAI and its Copilot instantiations, Google Gemini and GenAI LLM tuned for scientific papers like Zendy, asking a question and references produces with every LLM fake references, well constructed, but with different titles or authors than the web or journal reference actually associated to the citation, or sometimes totally invented.*

*Prompting to ensure that the reference exists and is correct may help for some, but in general it does not. Others have reported similar issues when using these LLM/GenAI services to produce legal briefs, and other legal documents.*

*This paper suggest simple ways to address this, instead of trying to just improve the LLMs and hope hallucinations will be reduced; they won't, no matter what, they are inherent to LLM. It is very surprising and mindboggling that LLM providers have not been implementing these kind of solutions: just check if the references exist, are correctly cited, and relevant to the paper/context.*

*We also expand the approach with our MultiAI approach to improve on the previous approach, or address other hallucinations; actually eliminating in our tests.*

____

## 1. Introduction

LLM hallucination is a well-known problem [4,5] of LLM / GenAI [4]. It is usually due to the data or to the models. An interesting problem comes from hallucinating references and their citations [6,11,12]. This is the case for scientific publications, but also when generating legal documents [8,11], as for example anecdotally reported in [7], and in many other articles on the web.

We did some experiments with different leading LLMs, and all generated immediately (not later in the conversation) fake scientific references. It happened on the first draft. Ironically, asking Google Gemini on November 26, 2024: "*Give me some key references of LLM hallucination. Please check that they exist, not fake and are correct.*", the response included a key reference as follows:

> Gemini: ***"How Language Models Could (and Should) Hallucinate Scientific References" (Durmus et al., 2023):*** *This paper specifically examines the issue of LLMs generating fake scientific citations and explores potential solutions.*
>
> *https://arxiv.org/abs/2305.13983*

[1] shmaes.physics@gmail.com

Cite as: Stephane H. Maes, (2024), "Fixing Reference Hallucinations of LLMs", https://doi.org/10.5281/zenodo.14543939, https://shmaes.wordpress.com/2024/11/29fixing-reference-hallucinations-of-llms/, November 29, 2024.

No such paper exists, from this author. Worse the arguments of the citation (i.e., URL and arXiv code) proposed by Gemini is [9]. It is unrelated to this title, author or topic! It guarantees that is not an obscure reference not easily found. Instead we know for sure that it is fabricated. Checking it immediately identifies that there is a problem.

When pointed out Gemini and Meta AI service (in WhatsApp) apologized and admitted having invented them. It is not clear that they were actually an admission, or rather part of how it had be trained to interact with the end user, following a “the user/customer is always right” kind of script. On the other hand Microsoft copilot[2], reacted very badly! Instead of admitting anything, it anecdotally tends to accuse the user to have tricked him, then that the user lied. Interesting isn’t it… And then, a replacement reference may be as bogus as the first one.

All the LLMs out there that we tested have similar hallucination problems. Perplexity seems the best behaving, but it exists too.

As discussed in [10,21], these kind of hallucinations are a real a danger to science (and other disciplines), especially if part of document presents formally, with authority and confidence and sounding correct, with fabricated citations to justify its statements. It does not matter that such statements may be correct. Of course it is even worse if the statements are also incorrect. Surely, we can understand that it renders a paper with fabricated reference worthless, fake news and misleading. When it comes to the authors of the paper, the loss of credibility and implications can also be disastrous. The same is true for student homeworks, or papers, or legal documents that may be statements assumed made under oath, or for works done against (large) compensation, or on the basis of credibility, expertise and experience of the authors. If this happens, many should mistrust GenAI LLM tools, and avoid using them for anything serious output. And this adds to concerns about privacy and confidentiality with using these tools, unless if it is entirely in the control of the user, or its employer; something often prohibitively expensive by now.

There are lots of efforts and investments done trying to detect, predict and mitigate hallucinations[19], including for examples references [8,11,13,21,22]. They, and many others, rely on understanding what happens deep in in a LLM, and predicting or mitigating the effects to reduce hallucination. It is great, but we doubt that this will lead anywhere useful as we know that hallucinations are inherent to LLMs [14,15,19,20]. Indeed, LLMs are just predictors of new word patterns based on past history of such patterns encountered during their training or fine tuning. In the context of what a LLM does, generating a well-structured citation, fabricated to support a statement, is a normal output based as far as the LLM is concerned. Teaching it to generate, some but not other is a lost battle. Even constraining it to only use RAG content is tricky, and forces somebody to have already create, and maintain an always up-to-date database of all the relevant past and recent bibliography. Yes, it may work but it is significantly reducing the value proposition of document generation. Then again, the Google Gemini example shows that it probably still won’t guarantee no fabricated references

Contrary to what many seems to believe and argue in the press, on the web or in the literature, LLMs are stupid systems that can only learn how to construct sentences or text, detect, repeat, and predict best suited patterns in a given context. Reasoning capability with chain of thoughts type of approach [34-36], as now considered by some as part of LLM with, for example allegedly ChatGPT-4o, are after all mostly prompt chaining, which requires designs, and just as RAG amounts to using non-AI to make AI, better AI. They are in fact hybrid rule -based / AI systems. Which is fine but “learned all the way”. AGI needs to be able to learn more and better rules or algorithms, and do it by itself based on reasoning, common sense, evaluation of the result and its consequences, and experience kept in memory as life experience lessons.

[2] This was the case of early versions still named ChatGPT, offered through Skype. The same behavior was observed in Copilot, but a while ago. More recently the LLM seems to have been tamed when it comes to how it reacts to the user, but not with respect to preventing fabricated references.

No LLM architectural improvements, dataset enhancements, or factchecking mechanisms will solve the issues at the level of the text (or multimedia generation). It is laughable to see experts claim that they are leading us to AGI. They aren't! Many additional concepts must be introduced. In fact one may argue that while LLM/GenAI will become a useful tool, it is probably a distraction towards reaching AGI! Others corroborate our view like for example in [16].

We provide approaches to address the reference / citation fabrication problem. They are based on simple common sense. The first one is an example of good engineering of a LLM system or application, the other ones are examples of MultiAI [1,2]. Our proposals to address reference hallucinations, or even more generic hallucinations, are other variation on the theme of chain of thought, going beyond pre-canned prompting template, along with possibly dynamical use of AI to discover, or select, or optimize the prompts, iterations and algorithms [1,2]. But what is really important is that there are reliable simple ways to address hallucinations that any LLM provider could use today. Yet they don't.

# 2. Just Check The Citations!

## 2.1 Understanding Reference Hallucinations

As mentioned in section 1, LLMs are inherently hallucinating [14,15,19,20]. Nothing can be done about it within the LLM, despites all the research thrown at the hallucination problem.

Let us consider textual hallucinations. Hallucinations come from problems with model or data, so that some patterns of words or sentence become highly globally, or locally probable. It can lead to unrelated or garbage story or even garbage text. As GenAI using LLMs is about generating new text never seen before, it is very hard to learn "forbidden sequences" of text, especially if it applies to a certain new context but not others: we can only teach after if an output is ok. Sure, some fine tuning and reinforcement learning can teach them later after they have been discovered, but then again, in another context or for another paper, the combination may be OK.  Adding more training data, however well cleaned or selected, and building new or larger models is of little help. Addressing these issues require solutions external to the LLM.

[11] identifies a possible way by asking questions about a text segments, in that case references, and checking inconsistencies between original text and the answers. They assert that when inconsistencies appear, e.g., the list of authors changes, they are probably hallucinations, and in this case fabricated references. We will take advantage of this later.

In the case of papers with references, it is easy to see why they are natural targets for high rate of hallucinations. Anecdotally, playing with Gemini (up to 2.0), ChatGPT (multiple versions), Microsoft Copilot, Zendy and others like perplexity, and asking scientific questions, gets resulting text that is usually correct, while fabricated references creep in very often. Indeed, they are, by definition, related to, detailing, justifying or historically tracking a very specific topic, or assertion. The LLM has learned how they are associated to the assertions, their format, and the list of journals, publications, or repositories that are used in a given context (e.g. legal vs medical vs Physics or mathematics). Also, it has learned through its training, or fine tuning data that it is good practice to associate references whenever something is referred to, asserted or summarized that is not original to the present paper and research. Therefore it seeks to do so. When it can't find an optimal reference, it is therefore pushed by its model to create one that seems plausible: i.e., with a related title and journal, publication name, authors, or repository. If authors are missing they will be selected optimally as best suite pattern. This is also why consistency checks as in [11] seem to provide a possible mitigation strategy: ask question and spot inconsistencies. They may be on the list of authors or the details of the publications: these were made up based on the context. Under new context they are new optimal combinations pop up and the citations change.

The previous paragraph implies that just asking the LLM questions about text may not work for more generic hallucinations. Indeed any text will probably change with the context, and a LLM has limited reasoning capabilities. We need to be more clever to address those. We will take a stab at it later.

Note also that relying on RAG does not help. RAG provides a repository of relevant / related knowledge relevant to a domain, which is then pulled based on the user's prompt, to be part of the prompt built and then actually fed to the LLM [23-25]. It is a non AI step, and it will not work for new document generation if one want to use references beyond what the knowledge base contained. Of course if one intends to limit the references to the repository then it is possible to reject any other or fabricated references. Such a system will work, and can be seen as a particular case of the system described in figure 1. Such a system may be sufficient for journalists, blogs and web site that only reference their own past publications. Fine-tuning does not help either as again one expects that the references are not from the fine tuning data, and, in any case, we now know that new references will be created by the LLM.

## 2.2 Search For It

Instead of pestering the LLM with questions about its references, which in fact does not lead it to fix the problem but often to just created another hallucinated one, the first approach that we proposed is deceptively simple, yet apparently never implemented. It also does not involved LLMs in the second phase, although it might subsequently. It is based on the idea already mentioned that AI, AGI or even GenAI should not just be a LLM, and LLM/GenAI tools like prompt generators, training and tuning tools, LangChain, MLOps or RAG, and many others, but instead requires much more [23-30].

In the cases that are the focus of our paper, it is easy to identify the references, citations and bibliography in a paper.

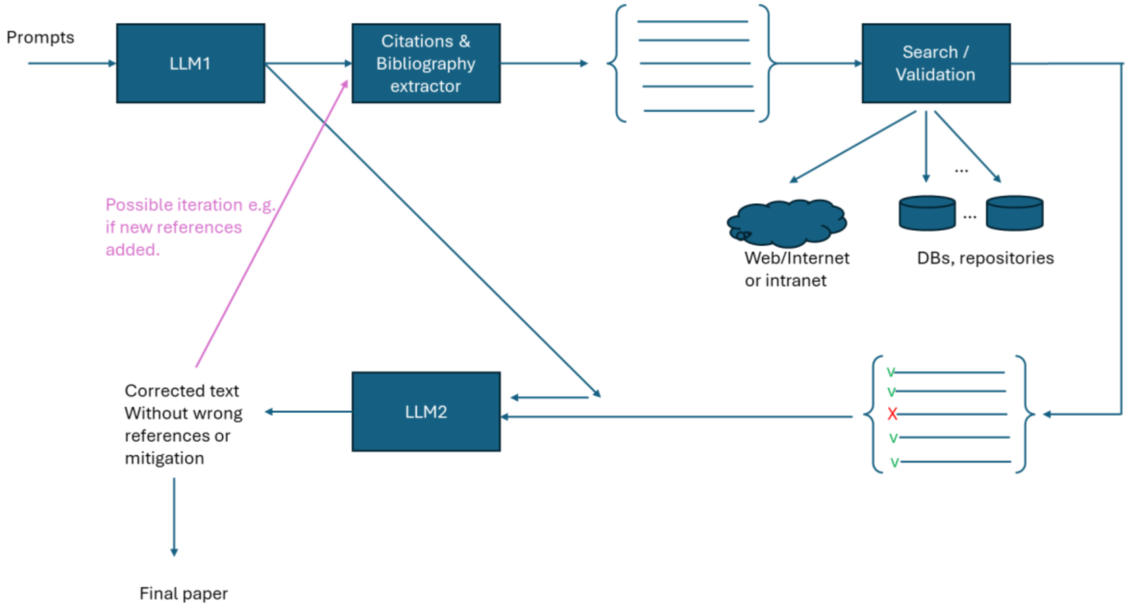
Prompts
LLM1
Citations & Bibliography extractor
Search / Validation
...
...
Web/Internet or intranet
DBs, repositories
Possible iteration e.g. if new references added.
Corrected text Without wrong references or mitigation
LLM2
V
V
X
V
V
Final paper

*Figure 1. A document is generates based on prompts. References and their citations are extracted, this may not require AI. The references are validated by checking their existence, by searching the web and paper databases. The paper is updated to handle problematic / unverified references. The process can iterate. Different LLMs can be used, albeit as same LLM is also working. LLM2 is optional, and it could be LLM1.*

Then, it is trivial to execute a search (Web, Citation Databases, etc.) of the references, using the citation and determine its existence and If the citation is adequate. Especially in the case of Google Gemini, Google with its indexing of the internet, patents, court  cases, Google scholar, library book indexing etc., is certainly capable to ascertain the correctness of most references. The same holds or Microsoft with Bing. Perplexity does the same but it seems constrained in terms of the literature that it can use, i.e., most preprint, public chats and web articles, not paywall articles, and even less books. That is another issue with current LLM: most may not have seen the main source of serious say scientific content contained in (non-digitized) books and beyond paywalls.

Unknown citations, which are not found by search, can then be filtered out, marked as non-verified (possibly versus the others marked as verified), or follow other steps as in section 3.

Anyway,  the proposed workflow is represented in figure 1.

The search may be limited to a certain repository to implement one of the cases discussed in section 2.1. You would think that a service like Zendy aimed at offering access to scientific publication would take advantage of this. It does not because, rightfully in our view, it also linked other references not accessible directly through the service.

## 2.3 Why not?

LLM service providers do not implement such a solution either because they never thought about it[3]. Or rather because of: (Arguments received from Google)

[3] Really? That may be the case but it would be somehow surprising.

> *A lack of real-time access, or because the LLM relies solely on it's internal knowledge and patterns to construct a plausible-sounding reference, failing to cross-check it with external sources, or more probably because of the cost of such a extra step.*

For providers like Google or Microsoft, the cost or access to indexes of the web and publications or publication repository seems hardly an "external access issue". Frankly, a provider like Zendy must have its own repository also. Sure the cost of the systems (more servers), and the extra response time due to executing searches and analyzing the outcome of some form of search query may be an issue. But so what? After all this may be a "do or die" situations for LLMs, or at least certain use cases. It is certainly frustrating to have to manually check any references, although we would argue that it is what would have to be done anyway by serious authors: LLMs should be a tool to prepare research, or finalize paper writing, not a replacement to it! Yet the risk of introducing fabricated references, possibly unbeknown to an author in the last pass of a paper, undermines confidence in LLM tools for serious purposes. The credibility impacts and legal risks are just too big.

In any case, our view here is that it behooves to Google and other AI LLM providers to address reference hallucination, and at, the minimum, performs checks as we proposed in section 2.2, now that we see that it is easily achievable, and resolving the issues.  Not performing such check, should lead to liabilities of the service providers both for producing irresponsible poor service, and for false advertisement of capabilities, but also for the possible longer terms consequences of their incorrect references. If they do not do the check, they should explicitly mark all (or any unchecked) references as unverified, in their output. This would warn less knowledgeable users, and avoid some of these potential liabilities. Today, as for many other parts, the way that LLMs are presented is really false advertising selling broken product to subscribers. Alternatively, a service provider implementing the proposal of section 2.2. would be able to differentiate in trust and credibility, and should win business for serious GenAI. Perplexity.ai seems to take that road, although apparently not yet implementing what we propose.

All this is also important not only in order to maintain confidence in science and LLM, but for security / safety reasons when it comes to medical documents and legal liability reason when it comes to legal and court papers.

But, we have an additional concern. It really matters because if we allow LLMs and LLM Service providers to generate more goop as bibliography, it will become harder to detect these fabricated references, e.g., if they appear in other papers, and weaken a solution as proposed here, bringing everybody to the starting block with no straightforward solution. Of course, one could then put in place credibility systems, that not only verify correctness of the citations, but also vet the references based on say the credibility of the sources, publications, number of paper referencing this reference, etc. But this would have the problem to disenfranchise other authors like experts in another field, or amateurs researchers, especially if they do not belong to academic or well-known corporate research institutions. It may stifle innovation[4], especially in an (academic) world where publications play such an important role in hiring, or to careers and promotions; a wrong and perverse role in our view [37]. Sabine Hossenfelder has a strong view, and published a lot about it. A starting point can be: [38,39], but she has many other recent blogs and videos on the subject[5].

[4] And yes, it might help with crackpots...

[5] Note that she is controversial, and she may have shift to a too extreme point of view in some of her latest publications. Citing her does not mean endorsement of all aspects, just referring an author who has pointed out problems with Academia and its (sole) emphasis on publications.

# 3. MultiAI to the rescue

In [1,2], we proposed MultiAI, combining multiple LLM engines, rules base systems, and other AI engines to implement optimal document / media generation in a particular domain or context, and systematically beating the best LLMs out there in terms of quality of the outcome.

The main reason is that determining whether a reference exists may not yet be sufficient. I would be even better if we could determine if the reference is useful or relevant, and details or support the statements made in the paper.

In the present case, we have a particular case of MultiAI, where:

- A first LLM generates a document based on prompt and context, including possible fine-tuned LLM or RAG.
- The same or another LLM is used to identify references/citations and bibliography.
    - It could also be a non-AI system as in section 2, but using AI allows us to be better track the reference, and the context in which it was used or statement tat it is supposed to support.
- The web and DB repositories are searched for each reference, as in section 2:
    - Authors, and citation are checked to see if they are correct and exist
        - At the difference of section 2, AI can also be used to compile a DB of references / bibliography (and papers referencing them), a bit à la Google Scholar.
    - Then a LMM is used to understand the reference abstract, if the information is available, or the whole document (or a summary), if available to determine if it relates and discusses the subject of the reference in the original output, and supports or details the related statement. This is way harder to do with a solution à la section 2.2. It can include inferring if the references are fabricated or not by asking questions about them as in [11].
- Verified and deemed suitable references are kept, and possibly marked as verified (if output of AI system, vs. a homework, scientific or legal paper, for which they probably should just be kept in the paper, without additional annotations).
- Unverified references should be removed and replaced by others, verified references, for example obtained by querying another LLM, or text detailing what they were supposed to describe instead of a reference.
    - In some cases unverified references could be highlighted, and marked separately as unverified or untrusted to let the user decide or investigate.
    - If a Google Scholar-like DB has been compiled but the solution, unverified references, i.e., not encountered explicitly as part of the search, could be considered as verified if they have been cited by other reputable (often cited) papers[6].

The approach is summarized in figure 2.

[6] Of course this may be fooled if many other authors are using fake references from LLM... However one can expect that the hallucination does not consistently fabricated the same fake references, even in a related context. It is related to [11]. As we discussed earlier, if AI generates a plethora of fake references that make their way to the literature, such kind of check will be harder to rely on.

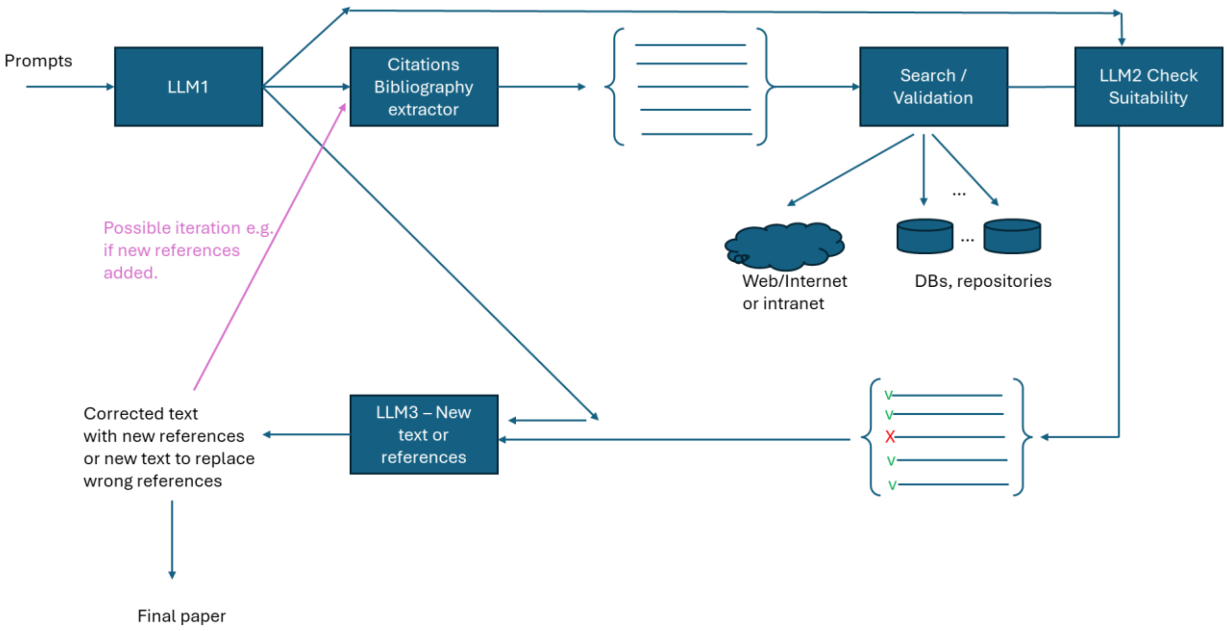
Prompts
LLM1
Citations Bibliography extractor
Search / Validation
LLM2 Check Suitability
...
Web/Internet or intranet
DBs, repositories
Possible iteration e.g. if new references added.
Corrected text with new references or new text to replace wrong references
LLM3 – New text or references
v
v
X
v
v
Final paper

*Figure 2. It shows an evolution of the proposal of figure 1, where multiple LLMs are used. LLM2 can be used to determine suitability of the reference for the statements they support, and ask questions about them. LLM3 can be used to rewrite with new references or replacement text any problematic references. Iterations can take place, possibly with different LLMs at different iterations. Only new, or initially problematic references need to be verified at a next iteration.*

As one can see, the main differences between the options of sections 2 and 3, is the use of MultiAi to find references in the original output, to understand the reference abstract or content and to estimate it suitability as reference to a particular aspect of the original output, and to rewrite the text with replacement text, or new references. Otherwise, both similarly verify the existence of the reference. Except that this may also include querying the LLM about the references to find inconsistencies per [11].

A priori, the above can be done with just one LLM, but mixing multiple allows “a different pair of eyes”, and typically no two LLM hallucinate the same way, except ironically when it comes to the citations for hallucinated references... The improvement obtained with different LLMs, especially if they can be specialized for a specific task, or for what they are the strongest at, or been trained / fine-tuned for.

# 4. Handling Broader Hallucinations

The MultiAI presented in section 3, and based on [1,2,53], can be used to prevent or mitigate other types of Hallucinations about other parts of a document or media. The approach goes as follows:

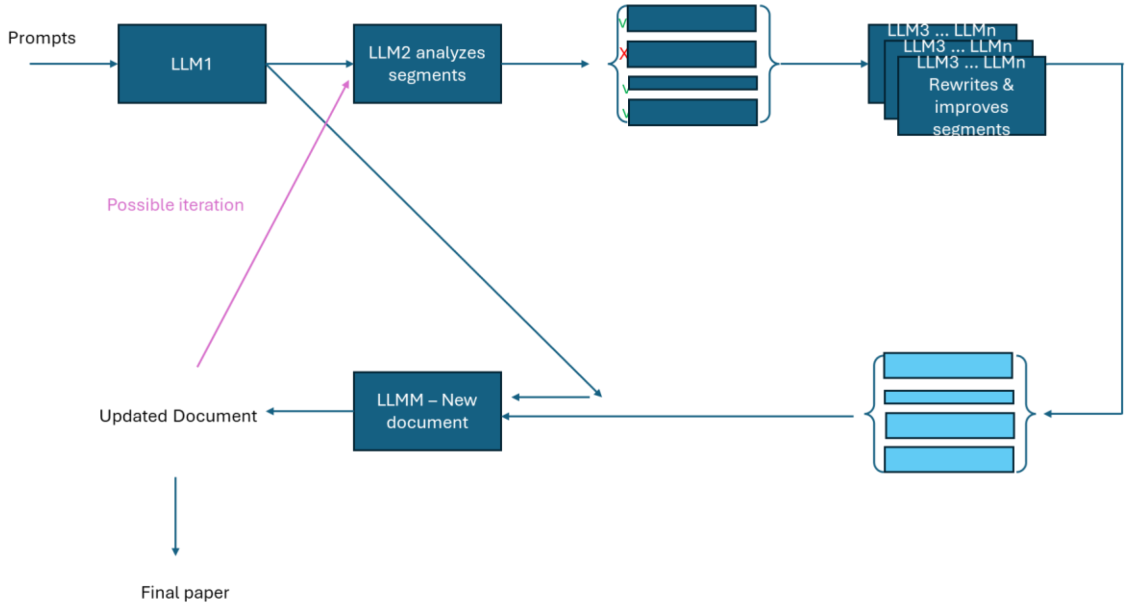

*Figure 3 shows a MultiAI version [1,2][7] of fixing the hallucinations in the document, involving multiple LLMs possible specialized for certain tasks or trained or fine-tuned for certain domains.*

- The output of the first LLM is passed through one or multiple LLMs to check items like:
    - Style, spelling and Grammar
    - Consistency of the content of a paragraph in the broader context
    - Plagiarism, which can be a sign of plagiarism, or a sign of having fabricated content based on paragraphs seen in the training data.
    - Consistency of the content of a paragraph with the knowledge in the literature, in the context of the output content:
        - Segments corroborated on the web, past papers etc., and consistent in the context of the paper will be considered as strong.
        - Segments apparently inconsistent with the literature will be flagged as “to be investigated”.
        - New segments may be checked for their logic, if within the capabilities of one of the available LLMs, and flagged or improved otherwise.
    - Segments to be investigated:
        - Can be further processed by other LLMs
        - Can be presented to the user for confirmation:
            - This can be done via a chatbot / conversational interface [31,33,43-52] or a Copilot à la Microsoft Windows Copilot, asking questions about the segment to confirm or change its content, or obtain approval of the segment by the user.

This kind of process allows reducing the risk of outputting inaccuracies, lies as well as non-sensical content. Fixing references is part of the process.

[7] Here the overall flow and LLMs are a priori fixed. In MultiAi, AI could also dynamically optimize them [1,2].

# 5. Cost and complexity

Searches, multiple AI engines, and iteration adds to the cost and complexity of the resulting system, but only as a multiple of what is obtained with a single LLM.

Using opensource LLM, like the Llama family [40], or SLM (small LM), are a way to reduce complexity, execution time and cost. In fact, we know that for knowledge distillation can best performance (for a particular domain) of LLMs (see [41,42], and references therein).

# 6. Conclusions

The approaches that we propose rely on the fundamental fact that hallucinations are inherent to LLMs, and that LLMs are not the way to AGI and human level intelligence (MultiAI may be [1,2,53]). We essentially do not believe that fine tuning, training or adding more LLM refinements to a LLM will solve anything. Math-based and step-by-step reasoning promised by OpenAI with ChatGPTo3 etc. is also wishful thinking, may better improving math problem resolution, but not reasoning *(Note added on December 22, 2024: [54])*. These are just way to navigate and decide the chains of thoughts, not much more.

Instead we put LLM is a broader framework that iteratively processes the output of a LLM to verify references or broader content, by checking existence of these references, their consistency, credibility, and adequacy for the domain, context. This can be done by simple search, or with other AI engines.

These approaches are particular cases of our MultiAI approach that mixes and matches, statistically or dynamically optimized, LLMs and other AI algorithms  (e.g. à la old AI), to iteratively improve content generated by (Gen)AI [1,2]. Future (AGI) reasoning capabilities will then come from being able to see the outcome of a reasoning and evaluating it suitability, and consequence. Ideally it will also contribute to a memory bank of lessons learned ,for the next time, to help in such evaluations, end evaluating the impact of an output.

These solutions can be implemented Today, and immediately significantly reduce the references, and other, hallucinations. Searches, multiple AI engines and iterations adds to the cost and complexity of the resulting system. However using open sources LLM, SLM, and knowledge distillation can help.

____

**References**

[1]: Stephane H. Maes, (2024), "MultiAI Document Generation Optimized for task and context", Internal Report, Factwise.ai, August 12, 2024.

[2]: Stephane H. Maes, (2024), " MultiAI Document Generation Optimized for task and context at Intelligine.ai", Intelligine.ai, October 1, 2024.

[3]: Wikipedia, "Hallucination (artificial intelligence)", https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence). Retrieved on November 26, 2024.

[4]: Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al., (2023), "A survey of large language models", arXiv:2303.18223v15.

[5]: Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen, (2024), “The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models”, arXiv:2401.03205v1.

[6]: Ayush Agrawal, Mirac Suzgun, Lester Mackey, Adam Tauman Kala, (2023-2024), “Do Language Models Know When They're Hallucinating References?”, arXiv:2305.18248v3.

[7]: Will McCurdy, (2024), “Stanford Professor Allegedly Includes Fake AI Citations in Filing on Deepfake Bill. Professor Jeff Hancock from the Stanford Social Media Lab submitted a legal argument in support of a Minnesota deepfake bill, but it reportedly includes citations made up by AI.”, PCMag, https://www.pcmag.com/news/stanford-professor-allegedly-submits-fake-ai-citations-in-argument-on-deepfake, November 24, 2024.

[8]: Abe Bohan Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, Benjamin Van Durme, (2024), “Gaps or Hallucinations? Gazing into Machine-Generated Legal Analysis for Fine-grained Text Evaluations”, arXiv:2409.09947v2.

[9]: A A Zhukov, I E Batov, (2023), “Regimes of electronic transport in doped InAs nanowire”, arXiv:2305.13983v1.

[10]: Oxford University, (2023), “Large Language Models pose risk to science with false answers, says Oxford study”, Oxford University News and Events, https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study. Retrieved on November 16, 2024.

[11]: Ayush Agrawal, Mirac Suzgun, Lester Mackey, Adam Tauman Kalai, (2023-2024), “Do Language Models Know When They're Hallucinating References?”, arXiv:2305.18248v3.

[12]: Walters, W.H., Wilder, E.I., (2023), “Fabrication and errors in the bibliographic citations generated by ChatGPT”, Sci Rep 13, 14045.

[13]: Farquhar, S., Kossen, J., Kuhn, L. et al., (2024), “Detecting hallucinations in large language models using semantic entropy”, Nature 630, 625–630.

[14]: Sourav Banerjee, Ayushi Agarwal, Saloni Singla, (2024), “LLMs Will Always Hallucinate, and We Need to Live With This”, arXiv:2409.05746v1.

[15]: Sahin Ahmed, (2024), “Hallucination in Large Language Models: What Is It and Why Is It Unavoidable?”, https://medium.com/@sahin.samia/hallucination-in-large-language-models-what-is-it-and-why-is-it-unavoidable-d9ddc1ebc29b.

[16]: Yann LeCun, (2014), “"LLMs are an offramp on the path to AGI" […]”, Twitter, https://publish.twitter.com/?url=https://twitter.com/ylecun/status/1801018194121118103#, June 12, 2024.

[17]: Francois Chollet, (2024), “OpenAI has set back the progress towards AGI by 5-10 years because frontier research is no longer being published and LLMs are an offramp on the path to AGI”, Twitter, , https://publish.twitter.com/?url=https://twitter.com/ylecun/status/1801018194121118103#, June 12, 2024.

[18]: Thomas Macaulay, (2024), “Meta’s AI chief: LLMs will never reach human-level intelligence. Sorry, Elon — AGI won't arrive next year”, TNW, https://thenextweb.com/news/meta-yann-lecun-ai-behind-human-intelligence, April 10, 2024.

[19]: Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, Jian Guo, (2024), “A Survey on Large Language Model Hallucination via a Creativity Perspective”, arXiv:2402.06647v1.

[20]: Muru Zhang, Ofir Press, William Merrill, Alisa Liu, Noah A. Smith, (2023), “How Language Model Hallucinations Can Snowball”, arXiv:2305.13534v1.

[21]: Mittelstadt, B., Wachter, S. & Russell, C., (2023), “To protect science, we must use LLMs as zero-shot translators”, Nat Hum Behav 7, 1830–1832.

[22]: Amos Azaria, Tom Mitchell, (2023), “The Internal State of an LLM Knows When It's Lying”, arXiv:2304.13734v2.

[23]: Charles Sprinter, (2024), “Evolving RAG Systems for LLMs: A Guide to Naive, Advanced, and Modular RAG”, ISBN-13 : 979-8346280682.

[24]: Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang, (2023-2024), “Retrieval-Augmented Generation for Large Language Models: A Survey”, arXiv:2312.10997v5.

[25]: Denis Rothman, (2024), “RAG-Driven Generative AI: Build custom retrieval augmented generation pipelines with LlamaIndex, Deep Lake, and Pinecone”, Packt Publishing.

[26]: Ben Auffarth, (2024), “Generative AI with LangChain: Build large language model (LLM) apps with Python, ChatGPT, and other LLMs”, Packt Publishing.

[27]: Noah Gift, Alfredo Deza, (2021), “Practical MLOps: Operationalizing Machine Learning Models”, O'Reilly Media.

[28]: Raschka, Sebastian, (2024), “Machine Learning and AI Beyond the Basics”, No Starch Press.

[29]: Giuseppe Bonaccorso, (2020), “Mastering Machine Learning Algorithms”, Packt Publishing.

[30]: Jeremy Watt, Reza Borhani, Aggelos Katsaggelos, (2020), “Machine Learning Refined: Foundations, Algorithms, and Applications”, Cambridge University Press.

[31]: Stephane H. Maes, (2023), “Comments to MPAI-MMC V2 Draft – Request for Public Comments”, https://shmaes.wordpress.com/2023/09/16/comments-to-mpai-mmc-v2-draft-request-for-public-comments/, https://osf.io/hj6pw, September 16, 2023.

[32]: Stephane H. Maes, (2023), “Comments to MPAI-MMC V2 Draft – Request for Public Comments”, viXra:2310.0015v1, September 16, 2023.

[33]: Stephane H Maes, (2023), “A GUIDE TO Building Your Own Service Desk Virtual Agent and simulating human conversations”, https://www.researchgate.net/publication/370100099_Building_Your_Own_Service_Desk_Virtual_Agent_and_simulating_human_conversations_A_Guide_to_Building_Your_Own_Service_Desk_Virtual_Agent_and_Simulating_Human_Conversations, April 18, 2023.

[34]: Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, (2022-2023), “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, arXiv:2201.11903v6.

[35]: Liam Sturgis, (2024),”Chain of Thoughts (COT) Prompting for LLMs: 7 Powerful Prompting Techniques to Get the Results You Want”, IBSN: 979-8884109520.

[36]: Lijie Hu, Liang Liu, Shu Yang, Xin Chen, Zhen Tan, Muhammad Asif Ali, Mengdi Li, Di Wang, (2024), “Understanding Reasoning in Chain-of-Thought from the Hopfieldian View”, arXiv:2410.03595v1.

[37]: Stephane H. Maes, (2022), “Why is The Multi-fold Theory on viXra, and not peer-reviewed (yet)?”, https://shmaesphysics.wordpress.com/why-is-the-multi-fold-theory-on-vixra-and-not-peer-reviewed-yet/, July 8, 2022.

[38]: Sabine Hossenfelder, (2018), “Lost in Math: How Beauty Leads Physics Astray”, Basic Books.

[39]: Sabien Hossenfelder (2024), “My dream died, and now I'm here”, https://www.youtube.com/watch?v=LKiBlGDfRU8, April 5, 2024.

[40]: Wikipedia, “Llama (language model)”, https://en.wikipedia.org/wiki/Llama_(language_model). Retrieved on November 29, 2024.

[41]: Wikipedia, “Knowledge distillation”, https://en.wikipedia.org/wiki/Knowledge_distillation. Retrieved on November 28, 2024.

[42]: Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, Pengfei Liu, (2024), “O1 Replication Journey -- Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson?”, arXiv:2411.16489v1.

[43]: Stephane H. Maes, (2023), “AI-ITSM: ChatGPT and IFS assyst”, IFS report.

[44]: Stephane H. Maes, (2022), “Smart ITSM, ESM, ITOM: adding AI, including Conversational AI, Beyond it, AIOPs and Beyond towards the Autonomous Enterprise”, IFS report.

[45]: Daniel Coffman, Liam D Comerford, Steven DeGennaro, Edward A Epstein, Ponani Gopalakrishnan, Stephane H Maes, David Nahamoo, (1998 to 2011), “CONVERSATIONAL COMPUTING VIA CONVERSATIONAL VIRTUAL MACHINE”, US patent US 8,082,153 B2.

[46]: Daniel Coffman, Liam D Comerford, Steven DeGennaro, Edward A Epstein, Ponani Gopalakrishnan, Stephane H Maes, David Nahamoo, (1998 to 2011), “CONVERSATIONAL COMPUTING VIA CONVERSATIONAL VIRTUAL MACHINE”, US patent US 8,082,153 B2.

[47]: Stephane H. Maes, TLV Raman, (1997 to 2010), “Methods and systems for multi-modal browsing and implementation of a conversational markup language”, US Patent 7,685,252.

[48]: Daniel Coffman, Liam D Comerford, Steven DeGennaro, Edward A Epstein, Ponani Gopalakrishnan, Stephane H Maes, David Nahamoo, (1998 to 2011), “CONVERSATIONAL COMPUTING VIA CONVERSATIONAL VIRTUAL MACHINE”, US patent US Patent 7,137,126.

[49]: Stephane H. Maes, (2000), “Elements of conversational computing-A paradigm shift”, International Conference on Spoken Language.

[50]: A Tiwari, RA Hosn, SH Maes, (2003), “Conversational multi-modal browser: an integrated multi-modal browser and dialog manager”, 2003 Symposium on Applications and the Internet. Proceedings, 348-351.

[51]: J Gergic, J Kleindienst, S Maes, T Raman, J Sedivy, (2001-2003), “Systems and methods for providing conversational computing via javaserver pages and javabeans“, US Patent App. 09/837,024.

[52]: Stephane H. Maes (1999), “Conversational biometrics”, The European Conference on Speech Communication and Technology.

[53]: Stephane H. Maes, (2024), “From MultiAI to SingleAI honed to a specific domain, user/enterprise, and/or specialized for a specific type of document or media”, Intelligine.ai, November 2024.

[54]: Will Knight, (20240), “OpenAI Upgrades Its Smartest AI Model With Improved Reasoning Skills. A day after Google announced its first model capable of reasoning over problems, OpenAI has upped the stakes with an improved version of its own.”, https://www.wired.com/story/openai-o3-reasoning-model-google-gemini/, December 20, 2024.