Dialogue Management

Written by Febogamedev

An important tool for analyzing, verifying and improving a language model's responses is **dialogue management** (cf. Madaan et al., 2023; Wu et al., 2022).
In this context, it should not be understood as a simple sequence of questions and answers, nor as a conversation equivalent to a human one.
Rather, it is a working space in which an initial response is clarified, broken down, subjected to critical scrutiny and progressively made more precise and easier to check.

The model's first response often represents only an initial plausible arrangement of the content.
It may be clear, fluent and apparently complete without yet being precise, stable or verifiable enough to be accepted as it stands.
Its value therefore also lies in providing an initial basis on which to work.

Dialogue extends the work begun with the prompt.
The prompt defines the task, context, purpose and, when necessary, the form of the output; subsequent exchanges allow us to correct what was not sufficiently clarified, such as ambiguous terms, implicit steps, excessive generalizations, omitted conditions or criteria left in the background (cf. Wu et al., 2022).

The value of dialogue does not depend on the number of exchanges or the amount of text generated. It depends on the function performed by each new question.
A methodical dialogue does not simply ask the model to elaborate, but introduces a precise operation: clarifying a definition, reconstructing a step, delimiting a statement, looking for a counterexample, distinguishing a fact from an interpretation or identifying the evidence needed for an external check.

If these operations do not occur, the conversation risks producing only an expansion of the initial response: more words, examples and explanations, but not necessarily greater understanding.

Many responses appear convincing because they compress the content into fluent, well-organized prose.
However, this compactness can conceal important distinctions, conditions of validity and intermediate steps.
Dialogue can make these elements explicit.

This does not mean observing the actual internal process through which the model generated the output: explanations produced afterward may be plausible reconstructions rather than faithful descriptions of the original computational procedure (Jacovi & Goldberg, 2020; Turpin et al., 2023).
Dialogue makes the argumentative structure of the content more visible, not the model's internal workings.

Besides clarifying a response, dialogue allows it to be **stress-tested**. A statement may appear solid in its initial formulation but reveal weaknesses when edge cases, different conditions, more precise definitions or alternative explanations are introduced.

- A robust response tends to retain a compatible core even when questioned from different perspectives (cf. Elazar et al., 2021).
- A fragile response, on the other hand, may fluctuate, contradict itself, introduce previously ignored conditions or abruptly qualify its conclusion.

These changes do not automatically demonstrate that the content is false, but indicate which parts require closer checking.
Dialogue is therefore the working environment in which the techniques analyzed in the previous chapters can be applied and coordinated:

- Domain knowledge guides questions, helping identify inappropriate definitions, implausible relationships and overlooked variables.
- Inter-response consistency can be examined by reformulating the problem or subjecting the same concept to different operations (cf. Elazar et al., 2021).
- Retrospective verification of criteria can be initiated by asking under which conditions, assumptions and criteria a claim should be considered defensible.
- Finally, cross-checking can be prepared by isolating critical statements and determining which sources, data or procedures are needed to check them.

Dialogue management therefore organizes verification, but does not replace external evidence.
It can help us understand what to look for, suggest keywords or indicate possible references, but sources proposed by the model must actually be consulted and verified (Walters & Wilder, 2023).
Obtaining a citation during the conversation is not equivalent to having checked the claim (cf. Walters & Wilder, 2023).

It is also necessary to distinguish dialogue within the same session from queries made in separate sessions.

Within the same conversation, the model has access to previous turns and can adapt to them, producing local consistency (cf. Laban et al., 2025).
This continuity is useful for developing and correcting the content, but does not demonstrate that the response would remain stable under independent conditions.
When we want to verify a claim's stability, it is therefore appropriate to supplement the dialogue with new queries that lack the previous context.

A methodical dialogue can serve five main functions.

- It can clarify vague terms and formulations.
- It can unpack overly compressed logical or causal steps.
- It can delimit a statement by identifying its conditions, limitations and exceptions.
- It can check consistency between different parts of the response.
- It can prepare external verification by isolating the claims and necessary evidence.

These functions do not form a rigid sequence, but help us understand what purpose each new question should serve.

Consider the following claim: “*PNG* is always better than *JPEG* because it does not lose quality.”

- An initial question could clarify the criterion: “What does ‘better’ mean in this case?” The model might distinguish visual quality, file size, transparency and intended use.
- A question that unpacks the claim could be: “Why should the absence of loss make PNG preferable in every situation?”
- A question that delimits it could add: “In which cases might JPEG be more suitable?”
- Retrospective verification of criteria could ask: “On what technical properties are you comparing the two formats, and what conditions make the judgment valid?”
- Finally, external checking could be prepared by asking: “What technical specifications or tests would allow these differences to be verified?”

Through these steps, the initial claim could be reformulated as follows:
“*PNG* uses lossless compression and is often suitable for graphics, text and images with transparency. *JPEG*, on the other hand, uses lossy compression but can produce substantially smaller files for photographs. The best format therefore depends on the type of image and its purpose.” (MDN contributors, 2026).

Dialogue management has turned an absolute statement into a set of more precise claims that can be compared with format specifications and practical tests.

However, dialogue also carries certain risks:

- Context drift: as the number of turns increases, the conversation may accumulate details, hypotheses and digressions that make the initial problem less clear and shift attention toward secondary aspects (cf. Laban et al., 2025).
- Dependence on previous premises: the model tends to build new responses on what has already been said. If an initial premise is weak or incorrect, it may therefore continue to treat it as valid rather than genuinely questioning it again (Laban et al., 2025).
- Apparent convergence: progressive agreement between user and model may create the impression that a conclusion has become stronger, when it may simply be the result of mutual adaptation during the conversation (cf. Sharma et al., 2023).
- Progressive rationalization: turn after turn, an initial response may be accompanied by increasingly elaborate and coherent explanations without an increase in the independent evidence supporting it (cf. Turpin et al., 2023).
- Overinterpretation of corrections: when the model changes a response after an objection, the change does not necessarily demonstrate that the new version is correct; it may simply indicate that the system has adapted to the new direction provided by the user (Huang et al., 2024; Sharma et al., 2023).

To avoid these problems, each new question should have a recognizable function and remain connected to the claim being analyzed.
Once the content is sufficiently precise, dialogue must give way to external verification, rather than indefinitely expanding the explanation.

Example

Someone asks a model whether studying while listening to music helps concentration.

The conversation might begin as follows:

- User: “Does studying with music help you concentrate?”
- Model: “Yes, in many cases music can improve concentration.”

The dialogue continues from here.

Context drift

The user starts asking:

- “Is classical or electronic music better?”.
- “What about headphones?”.
- “What about low volume?”.
- “What if I study in the evening?”.
- “What if I always listen to the same playlist?”.

After many turns, the conversation contains numerous details, but the initial question—whether music actually improves concentration—has not been verified. The dialogue has gradually shifted toward secondary aspects.

Dependence on previous premises

The model’s first response had already introduced the premise:
“Music can improve concentratio..”

Subsequent responses may therefore develop by implicitly assuming that this idea is correct:

- “Classical music might be better because…”.
- “A low volume might further enhance the effect…”.

The problem is that the model is refining a premise that has not yet been verified.

Apparent convergence

After a few turns, the user might conclude:
“So the best solution is probably instrumental music at a low volume.”

And the model responds:
“Yes, that seems like a good conclusion.”

User and model have reached a shared position, but this agreement does not constitute new evidence.
The conclusion may seem stronger simply because it was gradually constructed through dialogue.

Progressive rationalization

The model might then explain:
“Instrumental music reduces linguistic distractions, low volume avoids attentional overload, and a familiar playlist limits unexpected stimuli.”

The explanation has become much more elaborate than the initial response. However, if no studies, data or other independent references have been introduced, the evidence has not increased alongside the complexity of the explanation.

Overinterpretation of corrections

The user objects:

- “But music distracts m..”

The model then changes its position:

- “You’re right: music can actually worsen concentration, especially during complex tasks”.

This correction does not automatically demonstrate that the new version is more correct than the previous one. It may simply reflect the model’s adaptation to the new information introduced by the user (cf. Huang et al., 2024; Sharma et al., 2023).
At this point, continuing to ask for explanations risks producing only more variations of the same reasoning.

It is more useful to isolate the claim:
“Does music while studying improve or worsen concentration, and under what conditions?” and move on to verification through studies, data or other independent sources.

The problematic path would be:
Initial claim → More turns → More details → Greater agreement → More elaborate explanation

but not necessarily:
Initial claim → Better evidence → Greater reliability

For this reason, dialogue management is very useful for exploring, refining and testing a statement, but at some point it must stop and give way to external verification.

Dialogue therefore turns the first response into something progressively clearer, more delimited and easier to check.
It does not guarantee the correctness of the content and does not replace sources, data or tests, but allows us to organize the path through which we determine what needs to be clarified, challenged and ultimately verified (cf. Huang et al., 2024).

Dialogue management does not automatically make a response true: it turns it into a more precise formulation on which genuine verification can be performed.

Progressive rationalization

In dialogue with a language model, there is a subtle but important risk: **progressive rationalization**.

This expression describes the process through which an initial response, perhaps fragile, incomplete or overly categorical, is progressively made more coherent and convincing through subsequent explanations, without necessarily increasing the evidence in its favor (cf. Laban et al., 2025; Turpin et al., 2023).

The problem is not that the model consciously wants to defend what it said; the problem is structural.
Each new response is generated within the context built by the conversation: previous statements, terms used, requests for clarification and the direction taken by the dialogue become part of the material on which the model continues to work (cf. Laban et al., 2025).

When the user simply asks the model to explain more clearly, elaborate or give an example, the model is not being invited to verify the initial proposition, but to develop it.
It will therefore tend to produce a longer, more detailed and more plausible version of the previous response, without this constituting independent verification of its validity.

The conversation can thus create an illusion of deeper understanding.
The response appears richer, more nuanced and better argued, but that does not make it truer or better grounded.
A more detailed explanation does not automatically amount to a more reliable one (cf. Jacovi & Goldberg, 2020; Turpin et al., 2023).

The model may become more effective at connecting steps, making certain conditions explicit, introducing relevant concepts and constructing persuasive formulations.
However, all this may remain within the logic of the initial response without introducing a genuine check of its underlying assumptions.

The internal coherence of an explanation and its correctness are different properties. An explanation can be perfectly coherent while also being false, incomplete or based on questionable assumptions (cf. Elazar et al., 2021; Turpin et al., 2023).

A typical case might begin with a statement such as: “A sequence shot increases viewer immersion.”

- If the user simply asks for a clearer explanation of this sentence, the model might develop it by arguing that a sequence shot eliminates interruptions from editing, maintains perceptual continuity and strengthens the sense of presence within the scene.
- If questioned further, it might add that spatiotemporal continuity reduces the fragmentation of the visual experience and encourages more direct viewer engagement.

The response thus becomes progressively more elaborate.
However, none of these additions verifies that the relationship between sequence shots and immersion is actually well founded. The model is constructing a coherent explanation of the initial proposition, not necessarily demonstrating its validity.

The situation changes when we ask under what conditions the statement holds, in which cases it does not, what counterexamples could challenge it, what competing factors influence immersion or what evidence supports the proposed relationship.

At that point, the model is prompted to consider the limitations of its formulation:

- It might then acknowledge that immersion depends not only on the absence of cuts, but also on acting, pacing, staging, camera movement, narrative context and viewer expectations.
- It might also specify that a sequence shot can increase immersion under certain conditions, but can also produce distance, artificiality, boredom or disorientation.

This reveals the difference between developing a response and verifying it.

- Developing a response means making it longer, clearer, more precise or more elaborate.
- Verifying it means subjecting it to a check capable of revealing its limitations, assumptions, conditions of validity and possible alternatives.

Verification therefore requires a form of cognitive friction.
Questions must not merely help the response become more elegant; they must force it to withstand an attempt at refutation.

To make the checking process recognizable, we propose separating the production phase from the critical analysis phase. This separation alone does not guarantee reduced rationalization or reliable error correction (cf. Huang et al., 2024; Turpin et al., 2023).
An initial response is generated first; it is then subjected to critical questions. In brief: produce first, then verify.
Requests for expansion such as “explain more clearly”, “elaborate” or “give an example” can be useful, but are not sufficient.
They must be accompanied by questions capable of changing the cognitive function of the exchange:

- What conditions must hold for this statement to be valid?
- In which cases might it not hold?
- What counterexamples might challenge it?
- What assumptions are being treated as implicit?
- Is this demonstrated causation, a plausible correlation or an interpretation?
- What alternative explanations are possible?
- What evidence or sources support the statement?
- What information might show it to be false?

This does not mean that the model is incapable of correcting itself.
If explicitly invited to look for errors, counterexamples, alternative explanations or limiting conditions, it may even radically change its previous response (cf. Madaan et al., 2023).
The risk arises especially when all subsequent requests continue to develop the same proposition without changing direction.
When we want to seriously test whether a response holds up, it may be useful to reopen verification in a new conversation.

- Staying in the same session encourages continuity with the path already constructed. Previous statements become part of the context and may also steer subsequent responses (Laban et al., 2025).
- Starting a new session does not automatically guarantee a correct response, but allows a path less conditioned by the previous conversational context. This makes it possible to check whether the same conclusion also emerges through an independent path or depends on the framework built during the first conversation.

However, the strongest barrier against progressive rationalization remains comparison with external criteria.
A response that becomes increasingly persuasive within the same conversation is not, for that reason, a verified response.
Its strength can only be assessed by comparing it with counterexamples, explicit criteria, domain knowledge, observable data or independent sources.

The transition from internal explanations to external evidence marks the point at which plausibility can begin to turn into reliability.

Progressive rationalization is therefore not a mere stylistic flaw in dialogue, but an epistemic risk.
It shows that a longer conversation is not necessarily a better one, and that a more elaborate response is not automatically truer (cf. Laban et al., 2025; Turpin et al., 2023).

A reliable response is not one that merely explains itself better and better, but one that continues to hold up even when deliberate attempts are made to challenge it.

Iterating methodically

The techniques presented in the previous chapters should not be considered isolated tools. In practice, they are often used in sequence, progressively adapting them to the objective of the analysis.

Iterating does not simply mean repeating a prompt or correcting a response until the desired result is obtained.
It means using each new interaction to gather information about the model's behavior and decide which check to apply next (cf. Wu et al., 2022).

For this reason, it is useful to distinguish two phases of iteration.

- Exploratory iteration.
- Consolidation iteration.

Exploratory iteration aims to understand the problem and broaden the space of possible solutions.
At this stage, the prompt can remain relatively open, and the goal is not yet to obtain a stable response, but to observe which interpretations, strategies or perspectives the model tends to produce.

When a promising response emerges, we move on to consolidation iteration.
The aim is no longer to seek new possibilities, but to verify and strengthen those already identified. The prompt is progressively made more precise, while the various checking techniques are applied to analyze specific aspects of the response (cf. Wu et al., 2022).

The entire process can therefore be represented as follows:
Exploratory iteration
Open problem → alternatives → comparison → promising solution

Consolidation iteration
Promising solution → refinement → targeted checks → verification → stable version

The fundamental difference is that, in the first phase, the question is: "What possibilities exist?".

In the second, however, the question becomes: “Does this solution hold up when we check it more precisely?”.

Example
Let us consider this problem: “choosing how to organize a school presentation on climate change”.

The first request could deliberately be open-ended:
“How could I organize a ten-minute presentation on climate change for a class?”.

This is a phase of exploratory iteration.
The goal is not yet to obtain the final version, but to see what possibilities the model proposes.
The model might suggest, for example:

- A chronological structure, from the first scientific observations to the current situation.
- A structure of causes → consequences → possible solutions.
- A presentation built around three fundamental questions.
- A structure based on a few data points accompanied by concrete examples.

At this point, the user can continue exploring:
“Which structure would be easiest for students who know little about the topic?”
or:
“How would the presentation change if I wanted to avoid an overly alarmist tone?”
or again:
“Show me an example of the causes → consequences → solutions structure.”

Each new turn helps us better understand how the model interprets the problem, which strategies it proposes and which alternatives seem most promising.

After a few attempts, it may emerge, for example, that the “Causes → Effects → Possible responses” structure seems particularly clear.

Consolidation iteration begins at this point.
The question is no longer “What other structures could I use?”, but becomes:
“Let’s build a ten-minute presentation using the Causes → Effects → Possible responses structure. It must be understandable to 16-year-old students, contain no more than three numerical data points and clearly distinguish verified facts from possible future scenarios.”

The prompt therefore becomes more precise.
In subsequent turns, we no longer continually seek new ideas, but check specific aspects:

- “Are the three numerical data points actually necessary?”.
- “Are any statements too general?”.
- “Is the distinction between facts and predictions clear?”.
- “Is the planned duration of the different parts compatible with ten minutes?”.
- “Which claims require verification through external sources?”.

If a problem emerges, we address it in a targeted way.
For example, the model writes:
“Extreme weather events have doubled over the last twenty years”.

Instead of simply asking:
“Rewrite it better”.

we recognize that the problem concerns a quantitative claim and decide to verify it through an independent source.

In this sense, iterating does not simply mean producing versions ever closer to what the user likes, but using each response to decide what information is missing and which check should be applied in the next turn.

At this point, iteration takes on a methodical form.

- If a doubtful statement emerges, it can be subjected to retrospective verification of criteria.
- If it contains a causal relationship, it can be analyzed by reconstructing the causal chain.
- If we want to assess its stability, inter-response consistency can be applied.
- If the doubt concerns factual accuracy instead, cross-checking with independent sources becomes necessary.

Iteration therefore does not consist of continually adding new requests, but of choosing, at each step, which technique is most appropriate for the question we want to ask.

It may be useful to represent this process as a sequence of decisions.

- Explore the problem.
- Consolidate the prompt.
- Identify the relevant claims.
- Apply the most appropriate verification technique.
- Correct the prompt or claim when problems emerge.
- Repeat the cycle until the main issues have been sufficiently checked.

The goal of iteration is not to obtain an ever-longer response, but to progressively reduce uncertainty.
Each cycle should produce a better understanding of the problem or make it clearer which aspect still requires verification (cf. Wu et al., 2022).

In this sense, iteration is the common thread connecting all the techniques presented.
It is not an additional technique, but the method through which they are combined in a process of progressive analysis.

Supplementary material

Exercises

What is dialogue really for?
A student claims:
“If AI’s first response is not good enough, you just need to keep asking questions until it becomes complete.”

Critically analyze this statement:

- Explain why dialogue management is not the same as a simple sequence of questions and answers.
- Explain the relationship between the initial prompt and subsequent exchanges.
- Why can the first response be considered an initial plausible arrangement of the content?
- Explain why the value of dialogue does not depend on the number of turns.
- Distinguish these five possible functions of dialogue:
    - Clarifying.
    - Unpacking.
    - Delimiting.
    - Checking consistency.
    - Preparing external verification.
- Explain the difference between asking “Elaborate” and formulating a question with a precise checking function.
- Why is a longer, more elaborate response not necessarily more reliable?
- Explain how dialogue can function as a stress test of a claim.
- What behavior would you expect from a robust response when edge cases or counterexamples are introduced?
- Why must dialogue, at some point, give way to sources, data or other external evidence?

When a conversation convinces us too much
During a long conversation with a generative system, user and model gradually reach the same conclusion. The model provides increasingly detailed explanations and, when the user raises an objection, partially changes its position.
Analyze the situation using the concepts in the handout:

- Define context drift.
- Explain what is meant by dependence on previous premises.
- Explain the concept of apparent convergence.
- What distinguishes agreement between user and model from new evidence?
- Define progressive rationalization.
- Why can an increasingly coherent explanation strengthen only plausibility rather than reliability?
- Why does the model’s correction following a user objection not automatically demonstrate that the new response is more correct?
- Explain the difference between:

    - Developing a response.
    - Verifying a response.
- What does the handout mean by cognitive friction?
- Why might starting a new session be useful in some cases?
- Does even a new session constitute independent verification? Explain.
- Indicate what signal would suggest stopping the dialogue and moving on to external verification.

Explore or consolidate?
You need to use a generative system to design communications for a new school event.
Initially, you have not yet decided on:
- Priority audience.
- Channels.
- Main message.
- Communication style.
- Formats to produce.

After a few exchanges, however, a seemingly promising strategy emerges.
- Explain the difference between exploratory iteration and consolidation iteration.
- What is the main objective of the exploratory phase?
- Why might it be useful to keep the prompt relatively open at this stage?
- What, by contrast, is the objective of the consolidation phase?
- How should the prompt change when moving from the first phase to the second?
- Classify these requests as predominantly exploratory or consolidating:
- “What strategies could we use?”.
- “What audiences might be interested?”.
- “Check whether this strategy fits the available budget.”.
- “What problems does this proposal have?”.
- “Propose three very different visual directions.”.
- “Which claims in this campaign require sources?”.
- Explain why methodical iteration does not mean continuing to change the output until the user likes it.
- What role do intermediate responses play in choosing the next check?
- Explain why techniques such as inter-response consistency, retrospective verification and cross-checking do not necessarily have to be applied in every turn.
- Reconstruct the general process: Exploration → promising solution → consolidation → checking → revision.
- According to the handout, what is the real goal of iteration: increasing the number of versions or progressively reducing uncertainty? Explain.

Transforming an initial response through dialogue
Ask a generative system:
“What is the best way to present many projects in an online portfolio?”

Treat the first response only as a starting point.
Your task is to manage the dialogue independently until you obtain a more precise formulation that is easier to check.
During the work, you decide:

- Which terms need clarification.
- Which statements need to be delimited.
- Which steps need to be unpacked.
- Which counterexamples or alternative conditions to introduce.
- Which claims deserve external verification.
- When the dialogue is no longer producing useful information.

You are not required to use all the techniques studied: choose those you consider appropriate and justify your choices.

Submit:
- The initial response.
- The sequence of turns you consider truly significant.
- For each turn, a brief indication of the function performed by the question.
- The main changes that occurred in the response.
- Any claims you do not consider verifiable through dialogue alone.
- A final version of the response.
- A short reflection on the point at which you would have chosen to stop the dialogue and move on to external checks.

The aim is not to produce many turns, but to show that each new question performs a recognizable operation.

Exploration, consolidation and checking
With the help of a generative system, you must design a solution to this problem:
A school wants to reduce the number of students who fail to notice important announcements published on its website.

Do not start immediately with a solution.
Use dialogue in two phases.

First phase — Exploratory iteration
Freely explore the problem and possible solutions. You can question the model about users, causes, channels, information organization or alternative strategies.
When you believe you have identified a promising solution, stop exploring.

Second phase — Consolidation iteration
Put the chosen solution under pressure.
Decide independently which aspects to check and which techniques to use. You may discover that some parts of the solution need to be corrected, scaled back or abandoned.

Submit a report documenting:
- How the initial interpretation of the problem evolved.
- Which alternatives emerged.
- Why you chose a particular solution to consolidate.
- Which problems emerged during checking.
- Which questions actually changed your assessment.

- Whether any episodes of context drift, apparent convergence or progressive rationalization occurred.
- Which aspects would require verification with users, data or independent sources.
- The final solution, including its degree of uncertainty.

Bibliography

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics, 9*, 1012–1031. https://doi.org/10.1162/tacl_a_00410

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). *Large language models cannot self-correct reasoning yet*. International Conference on Learning Representations. https://openreview.net/forum?id=IkmD3fKBPQ

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4198–4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). *LLMs get lost in multi-turn conversation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.06120

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems, 36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html

MDN contributors. (2026). *Image file type and format guide*. MDN Web Docs. https://developer.mozilla.org/en-US/docs/Web/Media/Guides/Formats/Image_types

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). *Towards understanding sycophancy in language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.13548

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, Article 14045. https://doi.org/10.1038/s41598-023-41032-5

Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022*

CHI Conference on Human Factors in Computing Systems (Article 385, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3491102.3517582

Further reading

- Michael McTear and Marina Ashurkina, Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents, Apress, 2024 — Probably the closest overall match to this handout. It examines how conversational systems have changed with the arrival of LLMs, covering conversation design, prompt engineering, conversational agents, evaluation, safety, and the transition toward more autonomous systems. It is particularly relevant to understanding dialogue as an evolving interaction in which previous turns influence subsequent behavior rather than as a collection of isolated prompts.
- Michael McTear, Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots, Springer, 2020 — A strong theoretical foundation for dialogue systems and dialogue management. It covers rule-based, statistical, and neural conversational systems together with their evaluation. Although it predates the widespread use of modern generative assistants, it is useful for understanding dialogue state, context, response generation, interaction management, and why a multi-turn system must maintain coherence across successive exchanges. Springer Nature Link
- Chip Huyen, AI Engineering: Building Applications with Foundation Models, O'Reilly Media, 2024 — Particularly useful for the handout's emphasis on evaluation, iterative development, conversational feedback, monitoring, and the limits of treating model output as inherently reliable. Its discussion of application architecture and user feedback is especially relevant to the idea that dialogue should generate information about what must be clarified or checked next, rather than merely producing longer responses.
- James Phoenix and Mike Taylor, Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs, O'Reilly Media, 2024 — A valuable complement for the chapter's treatment of successive prompts as controlled interventions. The book emphasizes direction, examples, decomposition, evaluation, and iterative improvement. It is particularly relevant to the distinction between asking the model simply to "elaborate" and asking a new question that performs a specific operation such as clarification, decomposition, comparison, or evaluation
- Ben Shneiderman, Human-Centered AI, Oxford University Press, 2022 — The strongest complement for the human-control and reliability perspective. Shneiderman argues for AI systems that preserve meaningful human oversight, reliability, safety, and user control rather than maximizing automation for its own sake. This fits the handout's position that dialogue remains a human-directed verification process and that conversational agreement should never replace evidence or independent checks.

Editorial notes

In producing these materials, I have used generative artificial intelligence tools to support the writing process, particularly to improve the form of the text, reorganize content, refine wording, and speed up certain editorial tasks.

As I work independently on the production of these materials, I try to automate any activities that can reasonably be automated, so that I can devote more time to research, design, and content development.

However, artificial intelligence does not determine the content of this work: the choice of topics, structure, ideas, interpretations, examples, and teaching approach are developed by me. AI is therefore used primarily as a tool to support production and formal revision, while authorship and responsibility for the design and content remain mine.

Unless otherwise indicated, this material is distributed under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International license (CC BY-NC-SA 4.0).

You are therefore free to share, redistribute, adapt, and create derivative works based on this material, provided that appropriate credit is given, the material is not used for commercial purposes, and any modified or derivative versions are distributed under the same license.

