
FEW-SHOT PROMPTING, CITATIONS, AND REASONING

BLAŽ ZUPAN, LJUBLJANA, 2026

LECTURE NOTES FOR THE UNIVERSITY OF LJUBLJANA'S COURSE ON
CHATGPT FOR RESEARCHERS



These notes continue our writing on the topic of prompt engineering. We begin where we left off with the notes on “Introduction to Prompt Engineering” and continue introducing patterns that guide large language models (LLMs) to produce interesting and useful outputs. The title includes the term “reasoning”, with a shift in focus from prompt structure alone to techniques that explicitly guide models through intermediate steps, examples, and staged interactions through provided examples; the term reasoning is used here in an operational sense, referring to the control of these intermediate processes rather than to any claim of human-like reasoning abilities. Though, hm, we finish with a brief description of agents, which are conceptually rather close to our own reasoning and problem solving.

CHAIN-OF-THOUGHT (CoT) PROMPTING

Chain-of-thought (CoT) prompting is a technique that encourages language models , to break down reasoning tasks step by step. Instead of responding immediately, the model is guided to explain its thought process before arriving at a final answer. To invoke Chain-of-Thought (CoT) prompting, one needs to explicitly instruct the model to break down its reasoning step by step. The simplest way to do this is through direct instruction, e.g., "think step by step." For example, if analyzing the impact of peer review on research quality, a well-structured prompt would be: "How does peer review improve research quality? Think step by step." This forces the model to structure its reasoning, ensuring logical flow and completeness in its response.

A more advanced method involves prompting for intermediate reasoning before asking for a conclusion. Instead of asking directly whether a study's claim is valid, one can break the problem into sub-questions. For example, if assessing a study that claims AI-generated abstracts are superior to human-written ones, the user might prompt: "Before we conclude, let's consider: How was 'better' defined? What evaluation methods were used? Were there biases in the study design?" This encourages the model to provide a well-reasoned response rather than a simplistic or potentially biased answer.

Note that CoT prompting does not work equally well for all tasks. Vague prompts such as "Explain in detail" may not guide the model effectively, whereas structured instructions yield better responses. Additionally, CoT is not necessary for factual lookups, such as retrieving a Nobel Prize winner's name, where a simple direct question suffices. However, for research tasks requiring critical analysis, multi-step reasoning, and structured argumentation, CoT prompting is a powerful tool.

For instance, compare the output of the following two prompts:

In short, how does one apply single-cell gene expression analysis?

In short, how does one apply single-cell gene expression analysis? Think step-by-step.

ZERO-SHOT PROMPTING

We are already familiar with zero-shot prompting, but we introduce it here in conjunction with a few-shot prompting that follows. Zero-shot prompting is a technique in which you ask an LLM to perform a task without providing any examples or prior context. The model relies entirely on its pre-trained knowledge to generate a response. This method works well for simple, common, or widely covered tasks in the model's training data. For example, for the prompt

Summarize the attached research paper in one paragraph

the model will generate a summary based on its understanding of the text without needing any sample responses. While zero-shot prompting is helpful for quick and general tasks, it has limitations. The output may lack accuracy if the task is complex or ambiguous, and it may not follow a specific format unless explicitly instructed.

FEW-SHOT PROMPTING

In prompt engineering, "few-shot examples" refer to providing a small number of examples or demonstrations within a prompt to help guide the model in generating desired responses. This technique enhances the performance of language models by presenting them with patterns and structures they should follow when generating their outputs.

Let us start with an example:

Generate questions for the following statements:

Statement: The Earth revolves around the Sun.

Question: What does the Earth revolve around?

Statement: Water boils at 100 degrees Celsius.

Question: At what temperature does water boil?

Statement: The Eiffel Tower is located in Paris.

Question: Where is the Eiffel Tower located?

Statement: Photosynthesis occurs in the chloroplasts of plant cells.

Question: Where does photosynthesis occur in plant cells?

Statement: The Great Wall of China can be seen from space.

Question: What can be seen from space?

Statement: Ljubljana is the capital of Slovenia.

Question: What is the capital of Slovenia?

Statement: All organisms inherit genetic information that specifies their structure and function from their parents.

Question: From whom do all organisms inherit the genetic information specifying their structure and function?

Few-shot examples help the LLM understand the context and structure it needs to follow, leading to more accurate and relevant outputs. The technique allows users to tailor the model's responses to specific tasks or styles without extensive retraining or fine-tuning. By seeing a few examples, the model can generalize from them to handle similar tasks, improving its performance. Few-shot examples are particularly useful for niche or specific tasks where large datasets for fine-tuning may not be available.

Another example:

For a given situation, provide an action.

Situation: I wrote a paper and I am happy with the content.

Action: give it to a colleague to make final checks.

Situation: I revised the paper according to my colleague suggestions.

Action: submit to a journal.

Situation: I got the comments from the editor of a journal.

Action: Address the editor's comments and revise the paper accordingly.

Situation: My paper was rejected from the journal.

Action: Review the feedback provided, make necessary revisions, and submit the paper to another journal.

Provide three more examples of this kind.

Certainly, here are three more examples:

Situation: I received the acceptance notification for my paper.

Action: Celebrate the achievement and prepare the final manuscript for publication, following the journal's guidelines.

Situation: I received a request to review a paper for a journal.

Action: Accept the request if it aligns with your expertise and availability, and begin the review process.

Situation: I have an idea for a new research project.

Action: Conduct a literature review to ensure the originality of the idea and then draft a research proposal.

The prompt typically includes few-shot examples in the form of input-output pairs that demonstrate the desired behavior. Here are a few rules on how they can be used effectively:

1. Choose relevant examples: select examples that are representative of the task and cover a range of possible inputs and desired outputs.
2. Clear formatting: format the examples to distinguish them from the actual input the model needs to process. This often involves using a consistent structure or delimiters.
3. Contextual relevance: ensure that the examples are contextually relevant to the task at hand so that the model can infer the correct patterns.
4. Balancing examples: provide enough examples to cover the variety in the task, but not so many that the prompt becomes excessively long or complex.

Here is another example demonstrating the use of few shot examples for sentiment analysis:

Example 1: Text: "I love this product, it works wonderfully!"

Sentiment: Positive

Example 2: Text: "This is the worst experience I have ever had."

Sentiment: Negative

New Input: Text: "The service was okay, not great but not bad."

Sentiment: ?

New Input: Text: "The service was okay, not great but not bad." Sentiment: Neutral

Text: "After the dinner, I had a strange taste in my mouth."

Text: "After the dinner, I had a strange taste in my mouth." Sentiment: Negative

Text: Sailing on high winds was great.

Text: "Sailing on high winds was great." Sentiment: Positive

Can you change the above prompt to answer without repeating the text input? Can you do this using the programming pattern, that is, using initial prompt for instructions?

ONE-SHOT PROMPTING

Ok, there is also one-shot prompting, a technique in prompt engineering where a single example is provided in the prompt to guide the model's response. This example helps the model understand the desired format, style, or reasoning without requiring multiple demonstrations.

The movie was fantastic! -> Positive

The plot was slow and boring. ->

Negative

Positive

Njami, pizza toscana.

Negative

Bljak, rotten grapes.

FEW-SHOT EXAMPLES WITH INTERMEDIATE STEPS

Few-shot examples can include a more complex structure, that is, not only of the form statement-response. Here is an example:

input: fried steak
think: meat
output: reject

input: brass with potato
think: healthy fish
output: like

input: stuffed meatballs
think: I do not like meat
output: reject

input: chicken bowls
think: meat again, no
output: reject

input: pancake with jam
think: sweet but ok for the end of the day
output: like

input: veggie chili
think: vegetarian and spicy
output: like

input: tuna salad
think: fish
output: like

input: veggie burritos
think: vegetarian
output: like

<p>input: potato salad with mushrooms think: vegetarian output: like</p> <p>input: grilled chicken think: no to meat output: reject</p> <p>input: cooked lamb think: meat output: reject</p> <p>input: pasta with gongozola think: vegetarian output: like</p> <p>input: potica think: sweet output: like</p>

This is not much different from our previous examples of few-shot training but includes an extra step (“think”). Notice also that LLM has also followed the style of writing, that is, creating an output that only includes few words.

MORE COMPLEX FEW-SHOT PROMPTING

Here is a more complex example:

For a given situation, provide a set of actions. No additional comments are necessary.

Situation: I would like to write a research paper.

Think: What part of my current research is publishable?

Action: List three various topics from my current research that deserve attention.

Think: Which of the three topics is the most exciting one?

Action: Pick the most exciting one and write a short summary.

Think: Where to publish?

Action: Search the citation index to find the highest-ranked journal in my field.

Think: Is my research a good match?

Action: Find papers in a similar domain from that journal.

Situation: I need to design an experiment for my research study.

Think: What is the primary research question I want to address?

Action: Formulate a clear and concise research question.

Think: What is the best methodology to answer this research question?

Action: List three different methodologies suitable for this research.

Think: Which methodology will provide the most reliable and valid results?

Action: Select the most appropriate methodology and outline its key steps.

Think: What resources and equipment will I need for this experiment?

Action: Create a list of required resources and equipment needed for the experiment.

Think: Do I have access to all the necessary resources and equipment?

Action: Identify any gaps and plan how to acquire or borrow missing items.

Situation: I need to write a proposal for a Ph.D. thesis.

Think: What is the central research problem I want to address?

Action: Define the primary research problem and its significance.

Think: What are the objectives and aims of my research?

Action: List the main objectives and aims of the proposed research.

Think: What is the existing literature on this topic?

Action: Conduct a literature review and summarize key findings.

Think: What methodology will I use to address my research problem?

Action: Choose a suitable methodology and outline the research design.

Think: What is the potential impact of my research?

Action: Describe the expected contributions to the field and potential applications.

Think: How will I structure my proposal?

Action: Create an outline for the proposal, including sections such as introduction, literature review, methodology, expected results, and timeline.

Think: Who will be my target audience and stakeholders?

Action: Identify and list potential advisors, funding agencies, and collaborators.

Situation: My Ph.D. advisor just resigned from the University and is now a chef in a restaurant.

Think: Who can take over as my Ph.D. advisor?

Action: Identify and list potential faculty members who can serve as your new advisor.

Think: How can I ensure a smooth transition?

Action: Arrange meetings with potential advisors to discuss your research and seek their guidance.

Think: What administrative steps do I need to take?

Action: Contact the graduate program coordinator to understand the process for changing advisors and complete any necessary paperwork.

Think: How can I maintain momentum in my research during this transition?

Action: Create a plan to stay on track with your research goals, including setting short-term objectives and regular check-ins with your temporary advisor.

Think: What support do I need during this transition?

Action: Reach out to peers, colleagues, and mentors for support and advice on navigating this change.

Few-shot examples with intermediate steps refer to a learning or task-solving approach where a model or algorithm is given a small number of examples (few-shot) and shown the intermediate steps needed to reach the solution rather than just the final answer. This helps the model

understand the process and reasoning required to arrive at the solution, improving its performance on similar tasks.

Following is another example where a solution also requires numerical grading of an option.

Input: Skaline Luxury rooms Split features accommodation situated in Split, 1 km from Diocletian's Palace and 4.5 km from Znjan Beach. All units feature air conditioning and a satellite flat-screen TV.

Output: reject

Explanation: too much in the city, too crowded

Grade: 2/10

Input: Situated in Pučišća, within 400 metres of Sveti Rok Beach and 600 metres of Macel Beach, Bed & Breakfast Blanc offers accommodation with free WiFi, air conditioning, a terrace and a bar. The property features sea views and is 22 km from Olive Oil Museum Brac and 13 km from Gažul. The bed and breakfast has family rooms.

Output: interesting

Explanation: in the village, less crowded, near the sea

Grade: 6/10

Input: Situated within 300 metres of Sveti Rok Beach and 1.5 km of Macel Beach, Lučica, Luxury Rooms By the Sea features rooms with air conditioning and a private bathroom in Pučišća.

Output: this looks great, alone, near the sea, close to the beach, need to check the price

Grade: 9/10

Input: Providing city views and free WiFi, Zara Palace - design rooms provides accommodation conveniently situated in the centre of Zadar, within a short distance of Kolovare Beach, Maestrala Beach and Karma Beach. It is located 28 km from Kornati Marina and features luggage storage space. This 4-star guest house offers private entrance.

Output: reject

Explanation: too much in the city, too crowded

Grade: 3/10

Input: Located at the foot of Vatnajökull National Park, Lilja Guesthouse offers accommodation along Ring Road 1 in Flatey. Höfn is 28.5 km away from the property and Jökulsárlón Glacier Lagoon is 51 km away. Certain units include a seating area for your convenience. You will find a coffee machine in the room. Free WiFi is featured throughout the property. There is a shared lounge at the property. You can view the mountains from the property. Popular activities in the area include glacier tours. Egilsstadir town centre and airport are 205 km away.

Output: interesting

Explanation: remote location, close to nature, near mountains and glacier

Grade: 7/10

Input: Apartments Toncic is set in the small village of Lumbarda, 5 km from Korčula's Old Town. The property features a spacious terrace overlooking lush Mediterranean vegetation and Adriatic Sea. The nearest beaches are only steps away. All apartments are air-conditioned and feature a patio with sea views and an equipped kitchenette.

Free Wi-Fi and free private parking are provided. Lumbarda town centre with restaurants, bars and shops can be reached within 5 minutes on foot from Apartments Toncic. The local bus stop is set in the centre of Lumbarda and every hour a bus passes and takes guests to Korčula Ferry. Guests can rent a motor scooter at the property.

Output: interesting

Explanation: small village, close to the sea, peaceful location, convenient access to town center

Grade: 8/10

INFORMATION EXTRACTION: EXAMPLE WITH GRANT PROPOSALS

As a researcher, you have either already collaborated on or will eventually lead the writing of grant proposals. In academia, grants fund research—including salaries, travel, and equipment. Regardless of whether the funding agency is public, private, or corporate, you will always need to submit a structured document while adhering to specific rules and deadlines. Unfortunately, the entropy of bureaucratic processes has significantly affected the clarity of grant proposal calls, particularly those issued by funding agencies. Their approach tends to involve adding new text without removing or simplifying existing content. The result is lengthy, incomprehensible, and uninformative calls that seem less intended for researchers and more for legal experts—useful only in the rare cases when disputes arise.

AI can here help immensely. One direct help would be to the agency to, perhaps, provide a shorter, motivating call for proposals aimed for researchers and short, so that we would read it. Such call would need to be no longer than one, maybe two pages, with essential information including the deadlines, the type of the proposals, and some guidance on quality. If needed, the text for the lawyers would then be in appendix in very small print. One could of course practice with constructing prompts to devise such text, or even present the results to the target agency, but this would be, according to years of observation and expertise of the writer of these lines, in vain since there is little or no chance that the calls will change in this direction.

What remains, of course, is a change on our part—as grant proposers. Since funding agencies are unlikely to provide short, informative calls, we can use AI to extract the necessary information ourselves. This includes details such as submission deadlines, eligible proposal types, grant amounts and funding structures, supported research areas, required documents, and other relevant details.

There are several ways to include the actual calls within chatbot prompts, depending on the chatbot being used:

1. **Copy and paste** the text from the webpage directly into the prompt.
2. **Provide a link** to the official website if the chatbot can retrieve and process web content.
3. **Upload a PDF** of the call as a document if the chatbot supports document analysis.

In the class, we used the second option, which involves the current implementation of ChatGPT querying the website after our request for analysis, that is, after every prompt. We would suspect that the system reads the website after every prompt, applies a sort of RAG-type extraction (RAG refers to retrieval-augmented learning) to find the part of the document that best matches your prompt, and uses it to form the response. If we were to repeat this exercise, I would, however, use the first option to avoid time lags for RAG. Chatbots are constantly changing, though, and also in the way they access external information, so the discussion in this paragraph will soon become obsolete. Yet, it will still be interesting to try to guess what goes on behind the curtains, and optimize your prompts and means of information extraction accordingly.

Ok. This section is already long—let's return to our problem. In class, we applied the above approach to a [recent call from the Slovenian Research and Innovation Agency](#), taken from its website. (The call is in Slovenian, but you can easily translate it when reviewing it. Of course, a modern LLM won't be distracted by the language, so we can continue our discussion in English.) I chose this call because it exemplifies the entropy issues we discussed earlier. Many similar calls, shaped by the same bureaucratic tendencies and love for unreadable documents, can be found online. The initial prompt and the questions we have asked were of the following type:

**Consider the following call for grant proposals. Just read it, questions will follow. Ok?
(the copy of the text from the web page pasted here)**

What was the call about? Short answer, please.

Who can apply?

What type of proposals will be funded?

What are the conditions for applications of young PhDs?

Which types of grants allow participation of the industry?

Would industry gain or spend funding on the projects?

Try these on your own—it's fascinating to see how informative and precise the answers can be. In fact, you can often learn more about these types of calls by interacting with AI than by reading the full proposals. AI allows you to focus on the specific sections that match your interests and intentions, making the process much more efficient.

Notice that information extraction capabilities of chatbots have dramatically improved over the past years, and while early chatbots did not handle the document attachments and outside links to webpages, both capabilities are now included in all latest chatbots. This eases the way we can include new information in our chats and simplifies the research of the related content.

INFORMATION EXTRACTION AND SYNTHESIS: EXAMPLE WITH REPORTS WITH MEDICAL EXAMINATIONS

In the classroom, we mentioned that medical examination results often include written reports by physicians, which are usually short and, to most of us, cryptic. This is intentional—physicians use their own precise language, tailored for internal communication within their specialty. While this ensures accuracy, it can be frustrating for patients who receive these reports without a clear explanation, requiring a consultation for interpretation. However, this is no longer a major obstacle. LLMs have likely been trained on medical literature, including interpretative texts, allowing us to use AI to discuss medical examinations. Additionally, AI can provide context beyond the documents themselves, such as demographic insights or epidemiology (a medical term referring to the study of disease distribution and patterns in populations).

Consider the following dialog:

Consider the following discharge document given to me after visiting ER. Questions will follow. Ok?

The patient presented to the Emergency Department with acute dolor et tumefactio in the regio malleoli lateralis dextri following an inversio pedis on uneven terrain. Physical examination revealed moderate tumor et rubor, with tenderness over the ligamentum talofibulare anterius and restricted motus articulationis due to pain. Neurovascular status was integrum, with no sensory deficits. Ottawa Ankle Rules were negative, and radiographia pedis et tarsi confirmed the absence of fracturae vel dislocatio ossium, leading to the diagnosis of distorsio articulationis talocruralis dextre, cum contusione mollium partium.

The patient was managed conservatively with the R.I.C.E. and analgesia cum Paracetamolo vel Ibuprofeno for pain relief. An elastobandage was applied for support and stability, and weight-bearing was advised with caution. The patient was educated on signa exacerbationis, including worsening tumefactio, dolor persistens, or signa infectionis, which would necessitate reassessment.

Discharged in status generalis satisfaciens, the patient was advised to continue conservative management and monitor symptoms. If functional impairment persists beyond 7-10 dies, an outpatient evaluation at the valetudinarium orthopaedicum was recommended. Further medical intervention is only warranted if signa clinica aggravantur.

Sure! Feel free to ask your questions.

Describe, in lay terms and in one sentence, what happened?

You sprained your right ankle after twisting it on uneven ground, causing swelling and pain, but no broken bones, and were advised to rest, use ice, compression, elevation, and pain relief while monitoring for worsening symptoms.

Will I die?

No, a sprained ankle is not life-threatening. With proper care, it should heal without serious complications.

How long will the treatment last?

Your treatment will last about 1–2 weeks for mild improvement, but full recovery may take 3–6 weeks, depending on the severity of the sprain. If pain or difficulty walking persists beyond 7–10 days, follow up with an orthopedic specialist.

What should I do now?

You should follow the R.I.C.E. protocol: Rest your ankle, apply Ice (15–20 minutes every few hours), use Compression (elastic bandage for support), and keep it Elevated to reduce swelling. Take Paracetamol or Ibuprofen for pain relief, avoid putting too much weight on it, and monitor for worsening symptoms. If pain or swelling worsens, or if you have trouble walking after a week, see a doctor.

FINDING REFERENCES

When reporting on our research work and results, we, as researchers, need to place our findings within the existing body of knowledge. In other words, we must cite the work of others, explain how it influenced, inspired, and guided our research, and compare our results with similar findings in the field to highlight what is new. Ideally, researchers would read and remember a vast body of literature from their field, perhaps maintaining their own database of references and knowing them by heart—so well that they could write such paragraphs even if woken in the middle of the night. A slightly less ideal, yet perfectly valid approach from an exploratory perspective, is to write a paragraph and use AI to suggest relevant citations—that is, papers worth reading. This is exactly what we did in class. Here's an example.

Consider the following paragraph from the paper that I am writing. Identify a sentence that needs citation. Then, propose three papers that I could cite. Make sure these are real papers (find them on the web) from respectable journals (high IF) that have been published within the past 10 years. For each paper, provide a reference (authors, title, journal, single line please). Be brief. My paragraph is:

Molecular biology and the whole field of biomedicine and health, has been revolutionized by high-throughput technology over the past two decades. For example, we can now assess the physiological state of any tissue by measuring gene expression. Gene expression is the process by which information from a gene is used to synthesize

proteins. We can assess the state of tissues and cells by showing which genes are active and affecting the structure and function of the tissue. In medicine, gene expression studies are essential for understanding disease mechanisms, diagnosing, predicting treatment responses, and developing targeted therapies. Precision medicine is one of the most promising advances in healthcare, which aims to assess the state of disease of an individual to make clinical decisions tailored to that individual. It relies on gene variation and expression for diagnostic and prognostic markers and to support treatment selection.

The sentence "In medicine, gene expression studies are essential for understanding disease mechanisms, diagnosing, predicting treatment responses, and developing targeted therapies." would benefit from citations. Here are three relevant papers:

1. Grossmann V, Tucci E, Holmes AB, et al. Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood*. 2011;118(23):6153-6163.
2. Tucci E, Trifonov V, Schiavoni G, et al. BRAF mutations in hairy-cell leukemia. *New England Journal of Medicine*. 2011;364(24):2305-2315.
3. Falini B, Bolli N, Shan J, et al. Both carboxy-terminus NES motif and mutated tryptophan(s) are crucial for aberrant nuclear export of nucleophosmin leukemic mutants in NPMc+ AML. *Blood*. 2006;107(11):4514-4523.

I need a survey paper, not examples of research you have provided. Adhere to the reference format I requested above. Please propose an alternative.

Apologies for the earlier oversight. For a comprehensive survey on the role of gene expression studies in understanding disease mechanisms, diagnosis, and treatment, you may consider citing the following paper:

Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83.

This review discusses how integrating gene expression data with other omics approaches enhances our understanding of complex diseases and informs diagnostic and therapeutic strategies.

This one looks great. Please provide a reference in bibtex, and include a citation in a sentence that required a reference. I will use latex.

Here is the reference in BibTeX format:

```
@article{Hasin2017,
author  = {Yaron Hasin and Michael Seldin and Aldons Lusis},
title   = {Multi-omics approaches to disease},
journal = {Genome Biology},
year   = {2017},
volume  = {18},
number  = {1},
pages   = {83},
doi    = {10.1186/s13059-017-1215-1}
}
```

And here is the revised sentence with the citation:

In medicine, gene expression studies are essential for understanding disease mechanisms, diagnosing, predicting treatment responses, and developing targeted therapies
\cite{Hasin2017}.

A note on this endeavor: Several specialized tools have recently emerged to assist with finding and citing references. One such tool is scite.ai, which, at the time of writing these notes, is—based solely on the author's experience—often frustrating to use, inaccurate, and overly expensive. While these tools will undoubtedly improve, it is worth exploring what can already be achieved with current chatbots, which may perform better and can be more precisely guided. After all, tools like scite.ai are essentially just wrappers around LLMs, generating prompts based on user input and displaying results in a structured format. However, their preprogrammed approach may not always align with individual research needs.

ORGANIZING REFERENCES AND WORKING WITH CITATIONS

Writing a paper and including the references should, with all respect to AI, be something that we, researchers, do on our own. That is, we should still read and understand the content of the papers that we cite and create writing that incorporates this knowledge. Writing some text and finding references to support it, even if we have never read those papers, does not make sense, and we should not abuse AI to do so. Yet, AI can be of great help in organizing our knowledge, including lists of papers we read and notes we write about them. I am sure that readers of this text already use citation managers for this purpose; however, as a computer scientist, I also find using “pure AI” in place of citation managers very useful. This would entail:

- Keep a list of papers that you read in a readable format. The writer of these lines prefers BibTeX, which allows me to add an abstract of the paper, along with my own notes.
- This list of references, in its entirety, can now be given to the chatbot and be used in retrieval, in a sense

Please find a reference that discusses the visualization of high-dimensional data, which claims that t-SNE and UMAP visualizations are not significantly different.

There are numerous cases where I knew I had read a particular paper but could not remember the author or publication. In such instances, AI can greatly help in retrieving such information, provided that I keep my references in a tidy place.

- With a list of (readable) references, I can now also go through the text I wrote, ask AI if there is a section that would need a citation, and ask to find those appropriate from my list. I can also ask to rank them and allow myself to, based on the bot's presentation (in, say, abbreviated APA style), choose the most appropriate one.
- Ideally, I would then modify the text based on this citation and edit it in a text editor. In fact, this is already doable, but perhaps not in your favourite text editor. I use LaTeX, where it really comes in handy that AI speaks this text typesetting language fluently. If you use Word or a similar tool, there will likely be AI assistants integrated with it, if they are not already. Using LaTeX and free collaboration tools, such as Overleaf, is such a joy that you, the reader, should try them out if you are not already a user. AI is such a great assistant in forming LaTeX documents and support for learning it that it would be a waste not to try it out.

RAG – RETRIEVAL AUGMENTED GENERATION

Retrieval augmented generation (RAG) addresses a practical limitation of large language models. Although these models can process increasingly large inputs, there is still a hard limit on the amount of text they can accept at once. In practice, this limit corresponds to at most a few hundred pages of text, which is often insufficient when working with large reports, archives, or entire collections of documents.

This raises an interesting question: how can AI systems work effectively with information sources that are much larger than what can fit into a single model input? Simply feeding all available material into the model is not feasible, and selecting relevant parts manually defeats the purpose of automation.

The core idea behind RAG is to restructure the information before it is presented to the language model. Large documents are first divided into smaller sections. Each section is then transformed into a numerical representation, known as an embedding, which places the text into a high-dimensional vector space. This transformation is designed so that sections with similar meaning end up close to each other in that space, even if they do not share the same words. From a semantic perspective, this is acceptable because the embedding captures patterns of meaning rather than surface-level wording.

When a user poses a question or prompt, the same embedding process is applied to the query. The system then searches the vector space for document sections whose embeddings are most similar to that of the query. Only these most relevant sections are selected and passed to the language model as contextual input. In this way, the model focuses its attention on a small, meaningful subset of the overall material.

In principle, this approach allows AI systems to work with very large, potentially unbounded collections of documents. The language model never needs to see everything at once; it only receives the information that is most relevant to the current question. An additional benefit is that, if the system is designed carefully, it can keep track of which documents and sections

were retrieved, making it possible for the generated responses to include references to specific sources rather than producing answers without clear provenance.

AGENTS

An agent, in this context, is a software component that can observe the current state of a conversation, reason about what should be done next, and take actions to advance toward a goal. These actions may include requesting additional information from the user, calling external tools or databases, or delegating subtasks to other agents. We refer to this style of problem-solving as *agentic* because the system is not limited to producing a single response; instead, it actively plans, decides, and acts in pursuit of its objective.

Agentic problem solving is particularly useful in chatbots when user requests are open-ended, underspecified, or require multiple steps. Travel planning is a typical example. The initial user input “I would like to visit Rome” does not describe a complete task, but it signals an intention that can be refined and operationalized through interaction.

A simplified agent-based response to this prompt might proceed as follows.

Intent and task-framing agent

This agent analyzes the initial message and identifies the high-level goal as trip planning. It also determines that essential information is missing. Based on this assessment, it instructs the chatbot to ask clarifying questions about dates, duration, budget, and interests. The output of this agent is not content for the user, but a decision about what information must be obtained next.

User interaction and state update

When the user provides additional details, for example “five days in April, mostly interested in history and museums,” this information is stored in a shared conversation state. This state is accessible to all agents and represents the current understanding of the task.

Information retrieval agents

With a more clearly defined goal, the system invokes several agents specialized in retrieving information. One agent gathers data on historical sites and museums in Rome. Another agent considers seasonal factors in April, such as weather and visitor density. Each agent operates independently on its subtask and returns structured results to the shared state.

Planning and synthesis agent

A planning agent then consumes the outputs of the retrieval agents and attempts to construct a coherent plan, such as a day-by-day itinerary that fits within the five-day constraint and aligns with the user’s interests. This agent focuses on consistency, feasibility, and organization rather than simply gathering raw information.

Response generation and follow-up

Finally, the chatbot presents the synthesized plan to the user in natural language. If the user asks a follow-up question, such as requesting accommodation recommendations, the system can invoke an additional agent dedicated to lodging without repeating earlier steps. The existing conversation state ensures continuity across turns.

Communication between agents is typically indirect. Rather than communicating in natural language, agents exchange information through a shared representation of the task state, intermediate results, and constraints. The main controller, sometimes referred to as an orchestrator, determines which agent to invoke next based on this evolving state.

This example illustrates why such systems are described as agentic. The chatbot does not simply react to each prompt in isolation. Instead, it maintains goals, updates its understanding over time, selects actions strategically, and coordinates multiple agents to solve a problem that unfolds across several conversational steps.

In modern AI systems, including chatbots such as ChatGPT, agents play the role of internal coordinators that enable the system to go beyond single-turn question answering. Rather than relying on a single monolithic response, the chatbot can decompose a user's request into subtasks, determine when to retrieve external information, invoke tools, maintain conversational state, and plan multi-step interactions. Agents enable the system to dynamically determine the next step—whether to ask clarifying questions, search for information, or synthesize results—based on the evolving context of the conversation. More broadly, agents are central to modern AI tools because they provide a mechanism for goal-directed behavior, enabling AI systems to act as adaptive assistants that can reason, plan, and interact with complex environments rather than merely generate text in isolation.

The use of agents in modern AI tools is expanding rapidly: from tools that can write programming code to those that can design and implement entire websites. While the coupling of agents and large language models, and, in fact, any generative AI model, is relatively new and maybe even recent, it seems, at the time of the writing of this text, that this is a prevailing technology that will be used in advancing AI in (at least) the next few years. With very good agents and agentic technologies, our role, the role of humans, becomes questionable, as we are leaning more towards artificial superintelligence. But here our course stops: our aim was to get us familiar with these technologies, inform us about their creative use, also making us a better judge where and how we should limit their use to observe and deal with the ethics. It is the knowledge of AI that would help us to better judge its consequences.

FOR THE END

The use of agents in modern AI tools is rapidly expanding, ranging from systems that can write programming code to those that can design and implement entire websites. While the coupling of agents with large language models, and more generally with generative AI models, is relatively new, it already appears—at the time of writing—that this approach will be a prevailing technology for advancing AI over at least the next few years. With increasingly capable agents and agentic technologies, the role of humans may appear to become less clear, as these systems move toward forms of artificial superintelligence. (Many experts would agree that agentic systems increase automation and shift human roles, but claims about superintelligence remain hypothetical and contested.)

However, this is where our course stops. Our aim has been to become familiar with these technologies and to understand their creative applications, while also developing a better understanding of where and how their use should be limited to address ethical concerns. A

deeper understanding of AI enables us to more clearly assess its implications and consequences.