
INFORMATION EXTRACTION, DISSECTING CALLS FOR PROPOSALS, WORKING WITH CITATIONS, AND FEW OTHER TRICKS

BLAŽ ZUPAN, LJUBLJANA, 2025

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 start where we have finished with the notes on “Introduction to Prompt Engineering” and continue introducing patterns that guide large language models (LLMs) to produce interesting and useful outputs.

PROMPTS HAVE LIMITS IN SIZE

Try pasting the text from a more extensive Wikipedia page or a large Word document. Older LLMs might have struggled with long prompts, sometimes displaying errors like “The message you submitted was too long,” freezing with “Something went wrong while generating the response,” or processing only part of the text. Newer LLMs handle longer inputs more efficiently, either due to increased token capacity or by using techniques to shorten prompts.

LLMs usually measure prompt size in tokens. Tokens can be as short as a single character or as long as a full word, with their size varying depending on the language and context. For example, GPT-3.5 has a maximum token limit of 4,096 tokens, meaning that the combined total of the input (prompt) and output (response) cannot exceed this limit. This equates to just over ten pages of text—not much, right? While later models may have different limits, it's important to remain aware that the amount of text an LLM can process at once is always constrained.

The number of tokens in the input of recent LLMs has increased substantially. For instance, GPT-4o can process up to 128,000 tokens in its input, equivalent to about 192 pages of text. While this is a massive amount of data, it is still a finite limit, and one must be mindful of it when handling extremely large texts or engaging in extended chatbot conversations.

There are ways to filter out the most relevant text from documents before including them in a chat. However, this is not done automatically by the LLM; it must be handled in the preprocessing phase

by the LLM application. While not the focus of these notes, it is worth mentioning that interested readers can explore this topic further by learning about retrieval-augmented generation (RAG).

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 that one, maybe to 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, for which the current implementation of ChatGPT queries the web site after our request for the 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 would go again with this exercise, I would, however, use the first option to avoid time lags for RAG. Chatbots are constantly changing, though, also in the way how do 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.

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 dextrae, 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 CITATIONS

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 a following paragraph from the paper that I am writing. Identify a sentence which needs citation. Then, propose three papers that I could cite. Make sure these are real papers (find them on the web) from respectful 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, Tiacci 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. Tiacci 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, Lusi 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 Lusi},
  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.

QUESTION REFINEMENT PROMPT

We can engage language models to help us write better prompts. Here is an example initial prompt:

Whenever I ask a question, suggest a better question and ask me if I want to use it instead.

Or a more specific one:

Whenever I ask a question about molecular biology, suggest a better version emphasizing machine learning and AI. Ask me for the first question to refine.

Or a bit longer one where we also define the structure of LLM's response:

Whenever I ask a question about molecular biology, please suggest a better version of a question that involves some computational aspect and a specific related algorithm. Present this revised question and provide an answer (Q: <revised question>**, **A**: <answer>). Be brief; answer with a single sentence. Now, ask me for the first question to refine.**

Here is a brief example of the chat that followed our last prompt above:

Sure, what's your initial question about molecular biology that you'd like to refine into a computational aspect?

How do I compare two DNA sequences?

Q: How can I align two DNA sequences using the Needleman-Wunsch algorithm?

A: Use the Needleman-Wunsch algorithm to perform global sequence alignment, which scores matches, mismatches, and gaps to find the optimal alignment between two DNA sequences.

What's the next question you'd like to refine?

Can I infer an evolutionary tree given a set of gene sequences from various organisms?

Q: How can I construct an evolutionary tree using the Maximum Likelihood method based on a set of gene sequences from various organisms?

A: Utilize the Maximum Likelihood method to infer an evolutionary tree, which calculates the tree that maximizes the probability of observing the given gene sequences under a specified model of sequence evolution.

What other question would you like to refine next?

Here, we have successfully combined several prompting techniques we have already reviewed: programming, that is, instructing LLMs from the start on, question refinement, and constraining the output of the LLM to a specific, pred-defined structure.

COGNITIVE VERIFIER PATTERN

Here, we will ask LLM to help us break our question or problem into a set of subquestions, or subproblems. Our hypothesis is that once we provide an answer to these subquestions, LLM can become more precise and can provide us with a better answer. Here is an example of such a conversation, with an all-important initial prompt that specifies the type of chat we would like to have.

When you are asked a question, follow these rules: (1) generate several additional questions, one by one, that would help you to answer the questions more accurately; (2) after you get my answers, combine them with the individual questions to produce the final answer to the overall question. Number the questions. Ask max three subquestions.

Sure, I can follow those rules! Could you please ask your main question so we can begin the process?

Can I use machine learning to analyze single-cell gene expression data?

Great question! Let's break it down to get a comprehensive answer:

1. What are the characteristics of single-cell gene expression data that make it suitable or challenging for machine learning analysis?
2. What are the typical preprocessing steps required before applying machine learning to single-cell gene expression data?
3. What are some specific machine learning techniques commonly used for analyzing single-cell gene expression data?

Once I have your answers to these questions, I can give you a detailed and accurate response about using machine learning for analyzing single-cell gene expression data.

1) large number of features and data instances. 2) normalization and batch-effect correction. 3) dimensionality reduction, t-SNE visualization.

Great, thank you for the answers! Now, let's combine them with the original question:

Main question: Can I use machine learning to analyze single-cell gene expression data?

Answers to subquestions:

1. What are the characteristics of single-cell gene expression data that make it suitable or challenging for machine learning analysis?

- Single-cell gene expression data typically have a large number of features and data instances, which can pose challenges for traditional machine learning algorithms due to scalability issues and high-dimensional data.

2. What are the typical preprocessing steps required before applying machine learning to single-cell gene expression data?

- Typical preprocessing steps include normalization of gene expression values to account for differences in sequencing depth and gene expression levels, as well as batch-effect correction to remove technical variations introduced during sample preparation and sequencing.

3. What are some specific machine learning techniques commonly used for analyzing single-cell gene expression data?

- Specific techniques include dimensionality reduction methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) for visualizing high-dimensional data and identifying clusters of cells with similar gene expression profiles.

Final answer: Yes, machine learning can be effectively used to analyze single-cell gene expression data. However, it requires addressing challenges such as handling a large number of features and instances, performing necessary preprocessing steps like normalization and batch-effect correction, and employing specific techniques such as dimensionality reduction and t-SNE visualization to extract meaningful insights from the data.

Can you modify the initial prompt for the ChatGPT to be a bit less verbose? Can you modify it so that besides the question, we can also get an explanation of why answering this question will provide additional insight and help in composing the final answer? Can you compose a prompt to request that questions and their motivations be printed out in a structured way?

SEVERAL USEFUL PROMPT PATTERNS

By “prompt patterns” we here refer to the partial structure of our prompt to additionally specify how the output should look. Here are some functional patterns that can constrain the type of the output:

always answer with yes/no

always limit the answers to one sentence

always structure the answer to include a title and a one-paragraph summary

always answer in Slovene language

always answer with mathematical equations only, no additional explanation

Patterns give us control over the output. Test the last prompt above, and, say, ask something technical, like

How do I determine the reaction speed of my chemical reaction?

What is principal component analysis?

Patterns in prompts allow us to instruct LLM to use the output of a specific type. You can, of course, invent any such patterns, but it is good to know that “always limit,” “always answer,” and “always structure the answer” are good phrases to start your prompt that constrain all future answers by ChatGPT in the current chat. Besides those above, there are, of course, other very interesting patterns, the persona pattern we will introduce next being one of them.