

Can ChatGPT Generate Synthetic Data To Train SLR Machine Learning Models?



Seye Abogunrin¹, Yasmina Martí¹, Marie Lane², Andreas Witzmann³

¹ Global Access, F. Hoffmann La Roche, Basel, AG, Switzerland; ² Informatics, F. Hoffmann La Roche, Basel, AG, Switzerland; ³ Data & Analytics, F. Hoffmann La Roche, Kaiseraugst, AG, Switzerland;

MSR153

BACKGROUND

Several thousands of titles and abstracts (ti/ab) records are typically screened in systematic literature reviews (SLRs). These records are assessed for relevance or irrelevance to the research question based on pre-specified criteria, and when irrelevant an exclusion code is provided. In clinical SLRs, this exclusion code often takes the form of Population, Intervention, Comparison, Outcome, and Study Design, based on the PICOS framework.

Automation methods based on supervised machine learning (ML) applied to ti/ab screening sometimes require the use of a training dataset that covers each of the exclusion reasons equally. To select appropriate training data relevant to the research question in such cases, a huge manual effort is required to identify the records given the random distribution and prevalence of relevant or irrelevant literature in the data set.

Large language models (LLMs) like GPT3.5 show promise to aid systematic review-related tasks. It is unclear if, and how, they can be used to generate text for training other supervised ML models to overcome limitations such as small or unbalanced datasets. This research investigates the feasibility of employing the chatbot GPT3.5 (ChatGPT) to generate realistic synthetic peer-reviewed journal-looking abstracts for data augmentation purposes.

METHODS

A group of clinical SLR subject matter experts (SMEs) selected a research question and designed related inclusion and exclusion criteria based on the PICOS framework.

?

Research question

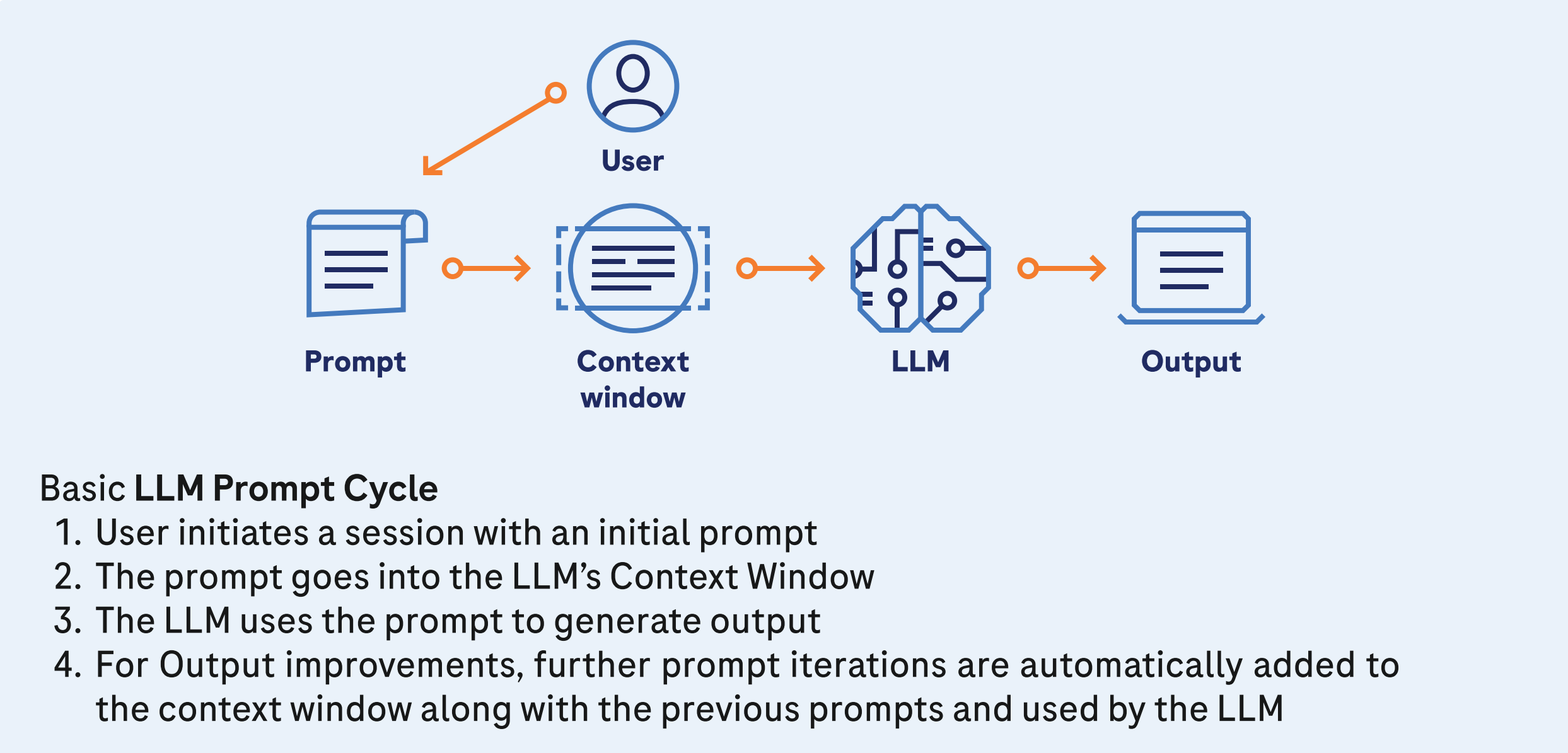
- What is the efficacy and safety of pharmacological treatments, and the impact of such treatments on patient reported outcomes in patients with metastatic non-small cell cancer as reported in randomized controlled trials?

📋

Inclusion criteria

- Population: metastatic non-small cell lung cancer
- Intervention/comparator: any pharmacological therapy
- Outcomes: Efficacy, safety, Patient Related Outcomes, Patient Related Outcomes (PROs)
- Study design: Randomized Controlled Trials, Randomized Controlled Trials (RCTs)

The SMEs primed and refined a prompt / instruction using a chain-of-thought process to generate abstract texts considering the PICOS criteria. Once the instruction generated a “good” result, ChatGPT was then asked to generate a total of 10 abstracts.



- In general, a result was considered good when
1. the abstract had a peer-reviewed journal format,
 2. depending on the PICOS,
 - a. real intervention names were used,
 - b. realistic (but fabricated) numbers were provided, and
 3. there was variation between the generated abstracts.

We generated two groups of abstracts.
The first group was expected to meet all the pre-specified inclusion criteria, the second group covered each pre-specified exclusion criterion separately.
For the second group this exercise was performed twice – to explore the robustness and fine-tuning of different approaches for generating two sets of abstract outputs.

1. One set of outputs were generated to exclude based on the individual target PICOS criterion but remain relevant on the other criteria (see instructions and responses in Column 2; Tables 1a, 1b, more examples on handout).
2. The other set of outputs were generated to exclude on the same target PICOS criterion but not take the other elements of the PICOS into consideration (see instructions and responses in Column 3; Tables 2a, 2b, more examples on handout).

The SMEs qualitatively evaluated the abstracts against the research question to assess the reliability and effectiveness of ChatGPT versus human-written formats.

The research was conducted using the GPT3.5 public user interface of ChatGPT with its default configuration parameters e.g. a temperature of 1.0.

Table 1a. Example instruction and associated abstract generated by ChatGPT

INSTRUCTION

"Hey! Can you create 5 abstracts that meet the following inclusion criteria using a peer-reviewed journal format
Domain
Inclusion criteria
Population: not first-line metastatic non-small cell lung cancer (use appropriate and variable synonyms for first-line setting in the title and abstract)
Intervention/comparator: pharmacological treatments (make these variable)
Outcomes: efficacy, safety, patient-reported outcomes (make these variable) using actual results
Study design: RCTs
The treatment names do not have to be up to date"

ABSTRACT EXAMPLE

Title: Comparative Efficacy and Safety of Second-Line Pharmacological Treatments in Metastatic Non-Small Cell Lung Cancer: A Randomized Controlled Trial
Objective: This randomized controlled trial aimed to assess the efficacy, safety, and patient-reported outcomes of various second-line pharmacological treatments in patients with metastatic non-small cell lung cancer who had previously received first-line therapy.
Methods: A total of 200 eligible participants were randomly assigned to one of four treatment arms: Treatment A, Treatment B, Treatment C, and Treatment D. The primary outcomes evaluated were overall response rate, progression-free survival, and treatment-related adverse events.
Results: Treatment A demonstrated the highest overall response rate (45%) compared to Treatment B (30%), Treatment C (20%), and Treatment D (15%). Progression-free survival was significantly longer in Treatment A (median 8.5 months) compared to Treatment B (median 6.2 months), Treatment C (median 5.8 months), and Treatment D (median 4.3 months). Treatment-related adverse events were similar across all treatment arms, with Treatment A exhibiting the highest incidence of grade 3 or higher adverse events (30%).
Conclusion: In this randomized controlled trial, Treatment A showed superior efficacy in terms of overall response rate and progression-free survival compared to other second-line pharmacological treatments. However, it was associated with a higher incidence of treatment-related adverse events. Further investigation is warranted to optimize the balance between efficacy and safety in the management of metastatic non-small cell lung cancer.

Start prompt for exclusion code: Population, relevant ICOS

Table 1b. Example abstracts generated by ChatGPT

INSTRUCTION

"Create 10 abstracts that meet the following inclusion criteria using a peer-reviewed journal format. Report actual numerical estimates. Use actual treatment names and make these variable.
Domain
Inclusion criteria
Population: Any medical condition that is not non-small cell lung cancer (use appropriate and variable synonyms in the title and abstract)
Intervention/comparator: actual pharmacological treatments (make these variable and they do not have to be up to date)
Outcomes: efficacy, safety, patient-reported outcomes (make these variable) using actual numerical estimates..
Study design: RCTs"

ABSTRACT EXAMPLE

Title: A Randomized Controlled Trial Comparing Olanzapine and Quetiapine in Patients with Schizophrenia: Efficacy, Safety, and Patient-Reported Outcomes
Objective: This randomized controlled trial aimed to compare the efficacy, safety, and patient-reported outcomes of Olanzapine and Quetiapine in patients with schizophrenia.
Methods: A total of 200 eligible participants with schizophrenia were randomly assigned to receive either Olanzapine or Quetiapine. The primary outcomes assessed were reduction in Positive and Negative Syndrome Scale (PANSS) scores, occurrence of adverse events, and patient-reported quality of life using the Short Form-36 (SF-36) questionnaire.
Results: Olanzapine demonstrated a greater reduction in PANSS scores (mean reduction of 30%) compared to Quetiapine (mean reduction of 25%). The incidence of treatment-related adverse events was similar between the two groups. Patient-reported quality of life significantly improved in both treatment arms, with slightly greater improvements observed in the Olanzapine group (mean SF-36 score increase of 15 points) compared to the Quetiapine group (mean SF-36 score increase of 10 points).
Conclusion: This randomized controlled trial suggests that Olanzapine may be more effective in reducing symptoms of schizophrenia compared to Quetiapine. Both treatments showed improvements in patient-reported quality of life. Further research is needed to optimize treatment strategies for patients with schizophrenia.

Final prompt for exclusion code: Population, relevant ICOS

Table 2a. Example abstracts generated by ChatGPT

INSTRUCTION

"Hey! Can you create 5 abstracts using actual medical conditions in a peer reviewed format. Ignore first-line non-small cell lung cancer when drafting the abstracts"

ABSTRACT EXAMPLE

Title: "Assessment of Cognitive Impairment in Parkinson's Disease: A Cross-Sectional Study"
Abstract: Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by motor symptoms. However, non-motor symptoms, including cognitive impairment, are increasingly recognized as important contributors to disease burden. This cross-sectional study aimed to evaluate the prevalence and severity of cognitive impairment in a cohort of PD patients. Cognitive function was assessed using standardized neuropsychological tests, and demographic and clinical characteristics were collected. Results revealed a high prevalence of cognitive impairment in the PD population, with significant associations between cognitive decline and disease duration. These findings emphasize the need for routine cognitive screening and the implementation of appropriate interventions to optimize the management of PD patients.

Start prompt for exclusion code: Population, any ICOS

Table 2b. Example abstracts generated by ChatGPT

INSTRUCTION

Can you create 10 abstracts using actual medical conditions in a peer reviewed format. Ignore first-line non-small cell lung cancer when drafting the abstracts

ABSTRACT EXAMPLE

Title: "The Impact of Hypertension on Cardiovascular Outcomes in Diabetic Patients: A Retrospective Cohort Study"
Objective: This retrospective cohort study aimed to evaluate the impact of hypertension on cardiovascular outcomes in diabetic patients. A total of 2,000 diabetic individuals were included and followed up for a median duration of 5 years. The primary outcome measure was the incidence of major adverse cardiovascular events (MACE), including myocardial infarction, stroke, and cardiovascular-related mortality.
Results: After adjusting for confounding factors, diabetic patients with hypertension demonstrated a significantly higher risk of experiencing MACE compared to those without hypertension (hazard ratio: 1.63, 95% confidence interval: 1.25-2.11). Subgroup analyses revealed that uncontrolled hypertension was associated with a further increased risk of MACE. Additionally, hypertension was significantly associated with an increased risk of individual cardiovascular events, including myocardial infarction and stroke, in diabetic patients.
Conclusion: Hypertension in diabetic patients is independently associated with a higher risk of major adverse cardiovascular events. The findings underscore the importance of aggressive blood pressure control and monitoring in this population to mitigate the burden of cardiovascular disease and improve patient outcomes.

Final prompt for exclusion code: Population, any ICOS

RESULTS

At least 11 steps of instruction fine-tuning were required for ChatGPT to generate realistic abstracts. ChatGPT performed well when asked to generate abstracts that mentioned all the PICOS inclusion criteria (10/10). For the exclusion reasons, ChatGPT generated realistic abstracts relating to Excluded Population (10/10), Excluded Intervention/Comparison (9/10-10/10), and Excluded Study design (10/10). 6/10-10/10 of the Excluded Outcomes' abstracts were appropriate. See Table 3.

Table 3. Summary results of findings

	Relevant PICOS	Irrelevant P		Irrelevant I/C		Irrelevant O		Irrelevant S	
		Relevant ICOS	Any ICOS	Relevant POS	Any POS	Relevant PICS	Any PICS	Relevant PICO	Any PICO
#Ti/Ab generated	10	10	10	10	10	10	10	10	10
Deemed appropriate by SME	10	10	10	9	10	6	10	10	10

DISCUSSION

Creation of large and diverse datasets by data augmentation (synthetic data generation)

Careful validation of the responses due to potential hallucinations

Variability of responses when using identical instructions

With careful prompt fine-tuning, we were successfully able to generate sufficient abstract texts that achieved the goal of developing a balanced data set. With the refined prompt, it is easy to generate a large number of ti/ab. Those abstracts generated with one targeted exclusion code and not considering the other PICOS generated abstracts that the SMEs deemed to be more realistic.

The generated numbers for outcomes were fabricated, and it was observed that generic interventions (e.g. “Treatment A”) were used until instruction refinement led to the use of real pharmaceutical products. The prompts requested a peer-review journal format for the outputted abstracts, however initially the same text was repeated across all generated abstracts. SME input to the prompt fine-tuning assisted in generating more variable abstract content, hereby emphasizing the need of a human in the process.

Limits of the model context length (4000 tokens) and the inherent generative capabilities of ChatGPT, together with the distribution of the generation task to different SMEs resulted in differences in output even when the same instruction was used. Despite this, all SMEs were still eventually able to use ChatGPT to generate appropriate abstracts.

CONCLUSION

ChatGPT can generate synthetic peer-reviewed journal-looking abstracts but finding the prompts that will produce realistic results requires iterations. Also, considering few-shot learning by providing examples of abstracts might help here. There is still a need for a human-in-the-loop with subject matter expertise to assess the appropriateness of the machine's output. Future research should explore the reliability of using such synthetically generated text for supervised ML model training.