Accurate Description and Interpretation of Clinical Endpoint Results Using Commodity Large Language Models

Dr. Lydia Frick, Dr. Daniel Brand, Heike Kielhorn, Univ.-Prof. Dr. med. Matthias P. Schönermark
SKC Beratungsgesellschaft mbH

ISPOR acceptance code: HTA263
Poster presented at ISPOR Europe 2024
17-20 November 2024 in Barcelona, ES.

OBJECTIVES

With the EU HTA process, new challenges arise for the HTA dossier compilation due to the expected huge number of PICO schemes to be addressed in only 100 days from notification about the PICO schemes to dossier submission. Mastering this process operatively requires new and tech-enabled approaches. There is a high potential for large language models (LLM) to support dossier compilation, such as for description and interpretation of endpoint results.

METHODS

Ten quality criteria for the generated description were defined, such as “no hallucinations” and “all numbers are correct”. Additionally, five conventions regarding content and structure of the generated text were defined based on extensive experience in German HTA dossier compilation, e.g., “Always describe the first measurement”. For the development dataset, 1,264 tables from publicly available German AMNOG dossiers were catalogued and categorized resulting in 15 table types, such as “Change from baseline”. For each table type, a set of synthetic tables was generated to feed into a core algorithm operating PaLM 2 32k text-bison. This allows for basic table understanding, imitation of writing style and fine-grained control of the LLM output. 245 tables were transformed into machine-readable format used as input for the LLM algorithm. The LLM outputs were evaluated regarding the need for adjustment to identify and categorize mistakes. To this end, a Number Hallucination Score was calculated and feedback from users was collected on a 4-point scale. To capitalize on technological advances and test the hypothesis whether LLM may be used interchangeably, we switched from PaLM 2 32k text-bison to Google Gemini. This poster shows the results generated using the latest approach.

RESULTS

EVALUATION OF LLM-GENERATED OUTPUTS

We introduced a "Number Hallucination Score“, which is measured on a scale of 0-100%. This score reflects the ratio between numbers in the generated text that do not occur in the original table and those that do. This provides an objective measurement of the accuracy of the LLM-generated text as a score of >0% may indicate an issue with hallucinated numbers.

$$\text{Number Hallucination Score} = \frac{|Numbers\ in\ text\ that\ are\ not\ in\ orignal\ table|}{|Numbers\ in\ text\ that\ are\ contained\ in\ orignal\ table|}$$

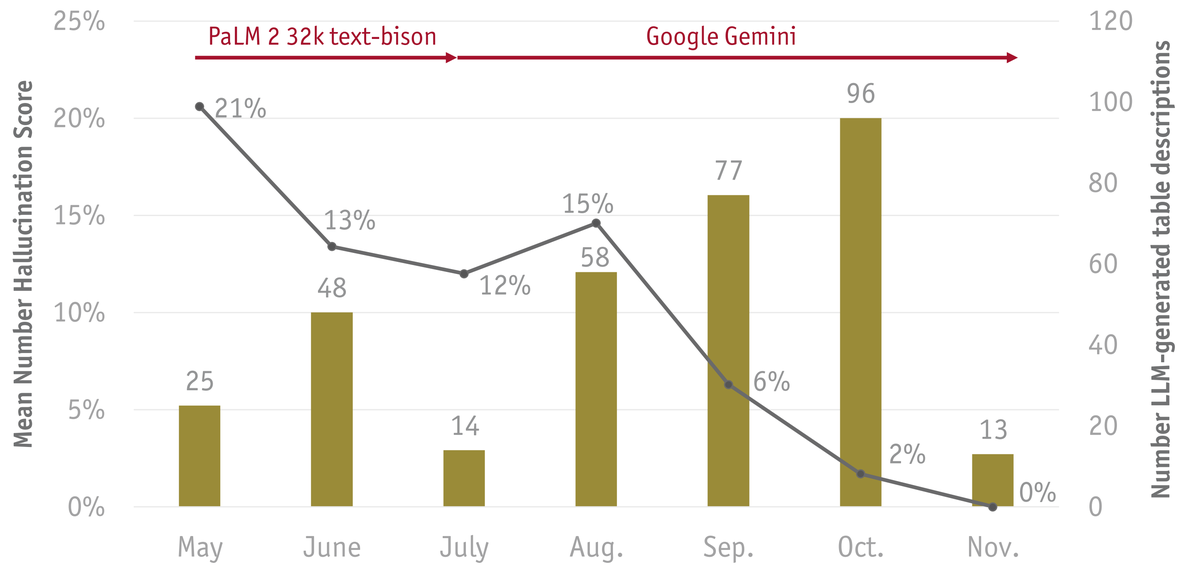

Fig.1: Mean Number Hallucination Scores (primary axis) and total number of LLM-generated descriptions (secondary axis) over the course of development in 2024

As shown in Fig. 1, number hallucinations occurred quite frequently with the early-stage algorithm (Number Hallucination Score = 21% in May 2024) and could be reduced to a reliable minimum by improving engineering of the prompts (e.g., 2% on average in 96 LLM-generated descriptions in Oct. 2024).

The LLM-generated descriptions of tables, which spanned a considerable length of up to four to ten pages, were subjected to a manual review and validation process. During the 5-week piloting phase, 47% of the generated results were directly usable or required minor adjustments, 35% of the generated descriptions required major adjustments but were still helpful, and 18% of the descriptions were not helpful (see Fig. 2). Most abundantly, data extraction from the table was incomplete or wording and writing style required adjustments. Importantly, users did not identify word hallucinations as a major concern. Note that the data shown in Fig. 2 was generated in the early stage of development and using PaLM 2 32k text-bison.

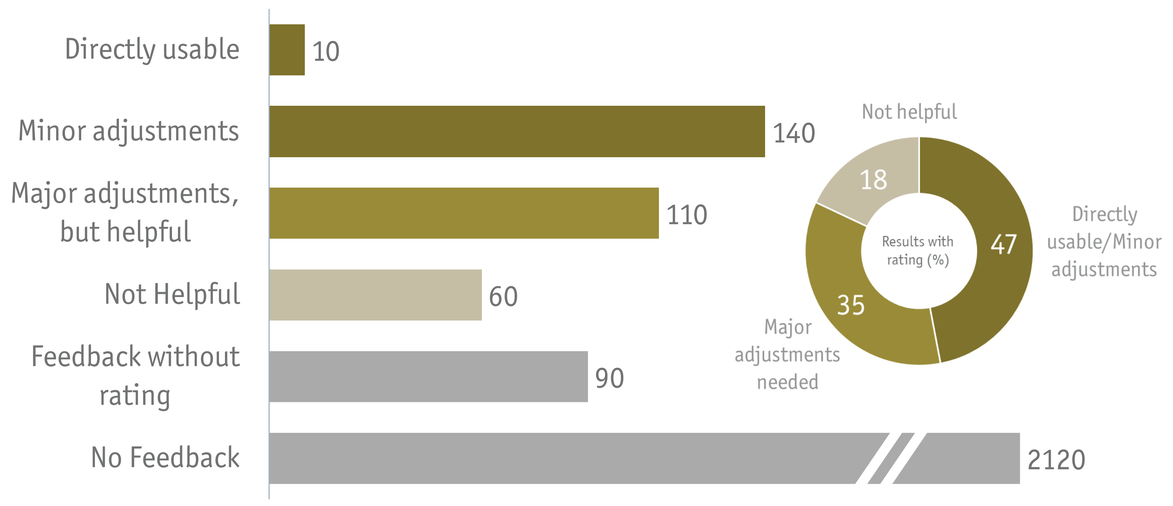

Fig.2: Total number (bar graph) and percentage (pie chart) of user feedback categories collected on the 4-point scale over the course of two months (April-June 2024).

The high number of LLM-generated table descriptions for which no feedback was given by the users (2,120 out of 2,530 generated texts) was mainly driven by erroneous submissions or, especially in the initial stages of testing, users attempting to familiarize themselves with the system without thoroughly reviewing the texts and providing well-founded feedback.

DATA TABLE TO GENERATED TEXT

*Table 1: Results for the outcome "change in PGIS - impact: social function" (ITT population) from RCT with the drug under evaluation (Asclepiopide) **

Treatment Week	n(a)/N(b)	Mean (SD)	Change from baseline: LS Mean (SE)	Asclepiopide vs. placebo: LS Mean Difference [95 % CI] p-value	Asclepiopide vs. placebo: Hedges'g [95 % CI]
Change in the PGIS - Impact: Social function					
Baseline					
Asclepiopide	23/23	40.4 (31.2)	-	-	-
Placebo	24/24	47.9 (40.53)	-		
Week 26					
Asclepiopide	21/23	12.4 (6.62)	-21.4 (4.05)	-9.1 [-23.9; 1.1] 0.0791	-0.22 [-1.04; 0.0607]
Placebo	21/24	23.6 (21.39)	-9.6 (6.53)		

The table shows the results for the endpoint "change in PGIS - impact: social function" (ITT population) from a randomized controlled trial (RCT) with the drug under evaluation (Asclepiopide).

The baseline values for the PGIS - Impact: Social Function averaged 40.4 (SD = 31.2) for Asclepiopide patients and 47.9 (SD = 40.53) on average for placebo patients.

After 26 weeks of treatment, the mean score of the PGIS - Impact: Social Function in Asclepiopide patients was 12.4 (SD = 6.62), which corresponds to a change of -21.4 (SD = 4.05) compared to the baseline. In placebo patients, the mean was 23.6 (SD = 21.39), which corresponds to a change of -9.6 (SD = 6.53) compared to the baseline.

The difference between Asclepiopide and placebo was -9.1 [-23.9; 1.1] with a p-value of 0.0791.

The Hedges’ g was -0.22 [-1.04; 0.0607], indicating a statistically non-significant advantage for Asclepiopide.

Table 1 shows an example endpoint table that served as the input for the LLM-generated description above (shown in italic). Google Gemini generated a 100% correct description of the clinical data including the correct interpretation of the p-value and direction of effect. Number Hallucation Score was zero, i.e., all numbers used in the text also appear in the original table.

* The Table was created for this poster; values are purposefully not correctly calculated as to demonstrate that the LLM does not calculate nor change values by itself. The generated text was automatically translated from German using DeepL Pro.

CONCLUSION

Commodity LLM can describe data tables showing clinical endpoint results accurately across a meaningful set of different table types and at a sufficient level of sophistication as required for HTA and other purposes. The use of LLM for table description can reduce the time required for generating the text and checking numbers by up to 50% and 80%, respectively. Smart prompt engineering based on solid experience in medical writing is key to generate reliable, accurate and correct outputs. Hallucinations can be minimized and controlled for if addressed early in the development process. The use of Standard Operating Procedures (SOP) regarding LLM-Input-Hygiene can help to reduce compounding errors in the LLM-generated text output. Nonetheless, quality check by human experts still is essential to ensure high accuracy and appropriateness of the text in each and every case. To this end, additional easy-to-use tools for the user’s quality check of the LLM-generated text are very helpful for detecting potential errors as it appears difficult to train an LLM to reliably detect erroneous or malformed data itself. The use of LLM for the description of clinical endpoint results has a huge leverage potential as HTA submissions and regulatory filings usually require the description of the results of the clinical trials in written form. In addition, the LLM approach is language-independent and makes it easy to generate descriptions of endpoint results in different languages.

SKC is a strategic consultancy focused on the increasingly challenging market access environment of innovative drug products based in Germany. We support the successful market access both on a strategic and an operational level. For nearly 20 years, our highly experienced team has been supporting our clients in solving their strategically complex questions.

SKC joined the MAP group, thereby broadening further the group’s European platform and combining crucial local expertise to drive our client’s global success. The MAP Group is a pan-European specialist strategic consultancy for pharmaceutical and biotechnology that has established operations across the UK, Ireland and Benelux, and has served more than 200 clients in 20 markets.

This project was conducted together with idalab GmbH and led to the development of the Endpoint Result Interpreter (EPRI), which is being continuously optimized by implementation of further features to generate improved output texts. It not only accelerates the time required to generate descriptions of trial data for regulatory or HTA purposes but also enhances reliable accuracy. Due to the modular tailor-made framework, it is adaptable for customization and can address potentially changing regulatory requirements.

schönermark kielhorn collegen

ACCELERATING PATIENT ACCESS

AxTalis

We are the market access special forces.

SKC Beratungsgesellschaft mbH | Hannover
www.skc-consulting.de | Email: frick@skc-beratung.de