

Performance of Specific Custom AI Models for German AMNOG-process Questions



C Minartz^{1, 2}, AS Neubauer²

¹ Institute for Health Economics (IfG), Muenchen, Germany; ² Neubauer Consult (NC) / Institute for Health- and Pharmacoeconomics (IfGPh), Muenchen, Germany

Objective

Artificial Intelligence (AI), specifically advanced language models such as ChatGPT, have the potential to revolutionize various aspects of healthcare. The German AMNOG assessment process is quite data intensive and all key documents are available online. Therefore the question was investigated, if customized AI can support specific process-related questions currently considered expert knowledge.

Methods

CustomGPT.ai is an AI model, which allows fast and easy setup based on website data and various formats such as PDF. For evaluation, 3 datasets were investigated:

- 1) Procedural documents on AMNOG process methodology,
- 2) All G-BA resolution documents for the last 2 years (2022+) and
- 3) All G-BA published documents for all assessments in ophthalmology (smaller bot as an example for full G-BA document coverage).

AI custom persona was adapted to focus on accuracy and give rigorous source citations. Multiple test queries were executed for all 3 custom bots to assess response quality. All queries were performed with documents in German language. Sample of information basis for evaluation 2 is in (Figure1):

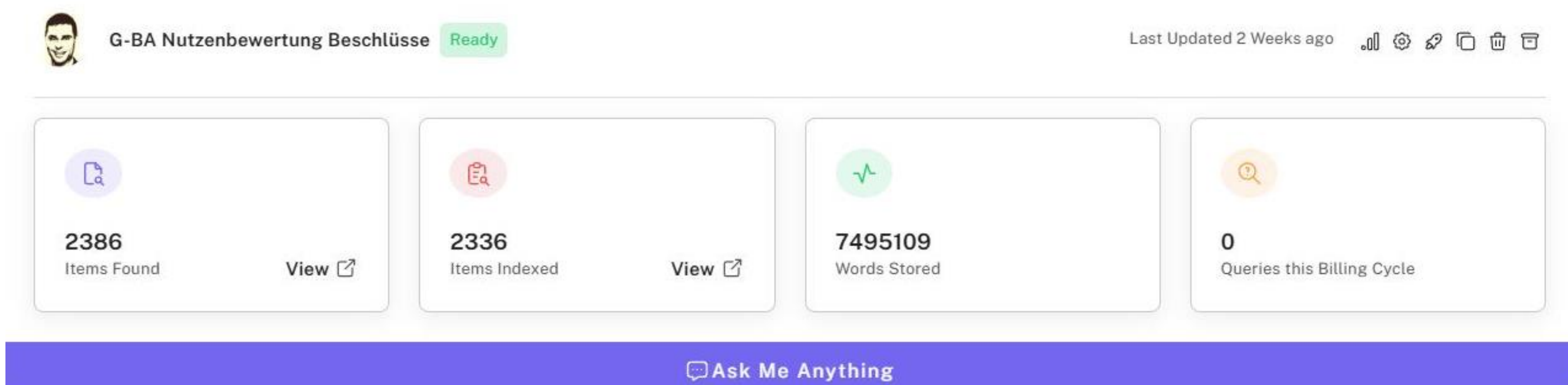


Figure 1: Database used to train bot 2 at the time of study execution - all G-BA resolutions from January 2022 to April 10th, 2024

Results

Responses to process-related questions (bot 1) were mostly accurate and included the relevant sources (e.g. Figure 2). In contrast, for all G-BA resolutions (bot 2) even for specific simple queries like on added benefit level for a specific drug or endpoint results, in most cases no relevant responses were received (Figure 3). By modifying custom persona “confabulation” could mostly be avoided and sources were cited correctly. Still, simple analytic queries like counting certain events were not performed adequately. With the disease area specific bot for ophthalmology based on all assessment-specific documents (bot 3) most specific queries could be answered to some degree. Still, for numeric specific questions – with clear answers existing in the available documents – in most cases no answer was given or answers were not fully related to questions (Figure 4). However, in such cases mostly the correct references were cited and linked.

Figure 2: Example of asking bot 1 for the assessment requirements in case of a type II variation: the provided answer was correct and specific, including timelines based on G-BA’s procedural rules.

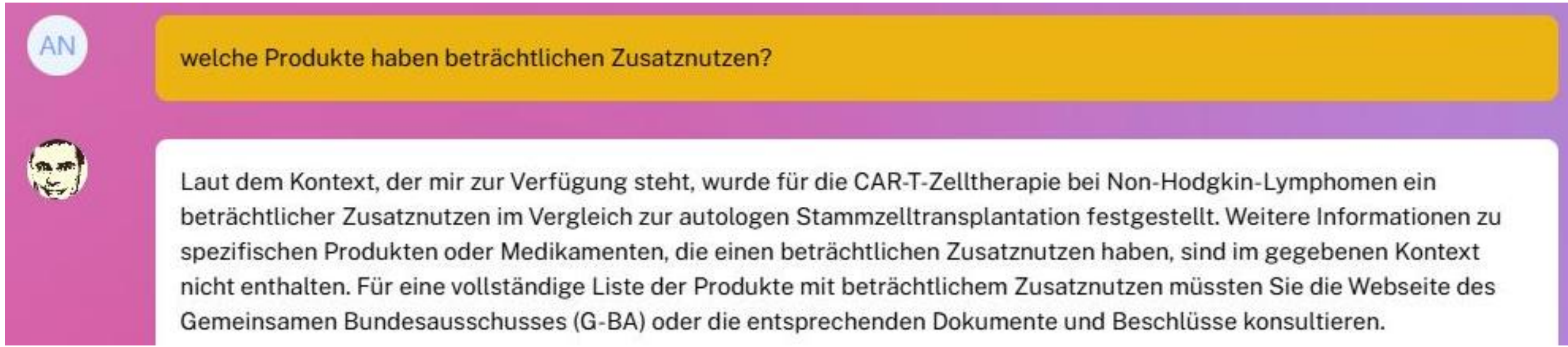
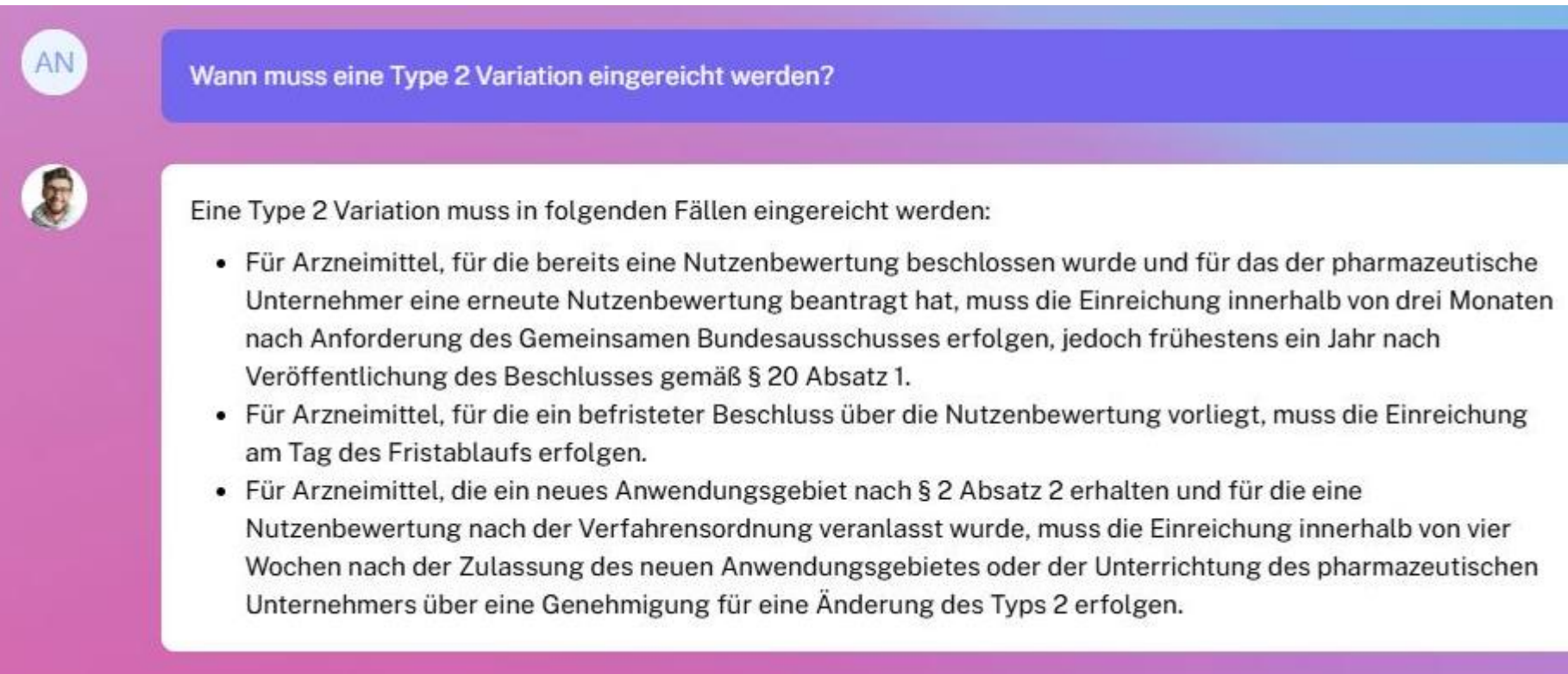


Figure 3: Asking bot 2 (all added benefit rulings) for products assessed to have a considerable added benefit: Answer was „CAR-T“ and asking the user to go the the G-BA website. No „confabulation“ as defined by custom persona, but a very incomplete response. By modification of search terms some more products could be identified, but no complete list could be generated. Changing persona did not change response behaviour relevantly in terms of analytical performance.



Figure 4: Asking bot 3 for the specific costs of a pharmaceutical (here: Ocriclasmin/Jetrea™): Answer was „no answer“, although the relevant G-BA resolution with costs was in the bot’s training PDFs. Same behaviour for several other questions and test queries for drugs. Also no listing of „products without added benefit“ could be generated – regardless of the limited number of resolutions to be processed in bot 3. In contrast, searching for persons in oral hearings yielded acceptable results and „maximum costs“ in the therapy field were identified correctly (i.e. a gene therapy).

Conclusion

Tested custom AI models were very precise for specific responses from texts. More complex and especially numeric analytic tasks performed suboptimal. This may be at least partly due to the AI engine used: case examples with ChatGPT – including PDF training documents – performed better in terms of analytic tasks, the aspect in need of improvement with CustomGPT.ai.