1. **(Optional) Suggestions for turning it into a language exercise (e.g., gap fill)**
2. **Gap filling exercises as you mentioned, could be used in multiple ways**

* Remove the medical keyword or a key grammatical marker and let learners fill in the blank. e.g (Die \_\_\_\_\_\_ erhält \_\_\_\_\_\_ nach der Operation. (Expected: **Patientin**, **Pflege**))
* For Passiv: Remove the auxiliary **werden** or the past participle, e.g.,  
  Die Wunde \_\_\_ (**wird**/ist) gereinigt.
* Can be done similarly for other grammar structures.

1. **Multiple Choice Questions (Same as above but give them 3-4 options to choose from)**
2. **Sentence correction**

* Modify a correct sentence to include a grammar or vocabulary mistake, then ask learners to identify or correct it. (can be done for medical vocabulary or grammatical structure identifier words like replacing Konjunktiv I verb with indicative form)

1. **Sentence Reordering**

* Give them sentence but reordered to help improve sentence formation.

1. **How you would expand this script to work on a larger corpus**

* Batch process sentences instead of sending them one by one to the stanza **nlp** function.
* Parallel processing or multi-threading can be utilized.
* Spacy is much faster but we need to define words for grammar structure detection by ourselves, which is also an option.

**Q. What NLP tools you would use to improve grammar detection**

* I think there is no improving this other than having more grammatical structures that we can define (Aktivsatz, Nebensatz etc.). Stanza that I used is the best tool for morphology tasks, and I believe we can make it faster based on the answer to my last question. Could take help from LLMs using their APIs like OpenAI API for preprocessing of docs.
* Another approach is that we take a labeled dataset (Gemtex/Tiger Corpus) that contains labels for grammar structures, ideally medical keywords too and train/finetune a neural network/transformer and see if that provides us high metrics.

**Q. How would you test and evaluate the quality of the extracted sentences (and subquestions)?**

To evaluate the quality of the extracted sentences, I would start with manual sampling by randomly selecting examples to ensure they contain both the target medical keyword and the specified grammar structure. Then, I would involve language experts such as linguists or experienced German teachers for linguistic validation to confirm the correctness of these samples. Automated checks can also help by verifying sentence grammaticality and keyword presence using language tools. In terms of accuracy metrics, I would measure precision, which is the proportion of correctly identified sentences containing both the keyword and grammar structure, and recall, which reflects how many relevant sentences in the corpus were found. The F1 score would provide a balanced evaluation by combining precision and recall. Additionally, I would track the false positive rate to identify the percentage of irrelevant or incorrect matches and assess coverage to ensure a good variety of grammar structures and keywords are represented. To handle false positives, I would refine extraction rules (in case we decide to use spacy or could also deep dive into the stanza library and see what I can find there) to reduce partial or irrelevant matches. I would also implement error logging to keep track of common false positives and use thresholding techniques to discard weak or ambiguous matches, thereby improving overall extraction quality.

**Q. How would you ensure sentence variety and reduce repetition? (and subquestions)?**

To ensure sentence variety and reduce repetition, I would implement mechanisms to track and limit how often similar sentences or those with the same structures and keywords appear in the output. This can include measuring sentence similarity with embeddings (e.g., using sentence transformers) and filtering out near-duplicates or very similar sentences. Additionally, I’d diversify the selection by prioritizing sentences that cover a broad range of vocabulary items, grammar structures, and contexts. To prevent the system from extracting “safe but boring” examples repeatedly, I would design the extraction logic to penalize overly common or simplistic patterns, possibly by weighting rarer or more complex sentence constructions higher. Incorporating random sampling and periodic refreshing of the sentence pool also helps keep the examples fresh and engaging. Finally, we could gather learner feedback and think of a way to incorporate their feedback into our exercises for the better.

**Q. (Optional Bonus) Suggest an evaluation method….**

An effective human-in-the-loop evaluation would involve linguists or native speakers reviewing a sample of extracted sentences for correctness, relevance, and value, providing feedback to refine extraction rules. Learner feedback can also be collected through quizzes using these sentences to assess clarity and difficulty. For automated unit tests, I’d create synthetic sentences with known grammar structures and keywords to verify the extraction logic detects them accurately and rejects false positives, ensuring consistent, repeatable validation during development.

**Q. Prompt Design & Evaluation for LLMs**

**Prompt 1:**

**"Create 5 B1-level fill-in-the-gap sentences using the German Passiv voice focused on the medical topic 'OP-Vorbereitung'.**

**Risks:**

* The model might generate sentences that do not use Passiv correctly.
* Vocabulary could be too advanced or too simple for B1 learners.
* Sentences may include hallucinated or medically inaccurate details.

**Improvements/post-processing/prompt engineering:**

* Specify grammar clearly (e.g., "only use werden + past participle").
* Add a vocabulary list restriction or use a medical glossary to keep vocabulary appropriate.
* Use automated grammar checking tools or a rule-based filter to verify passive constructions.
* Have a domain expert review samples to filter out inaccuracies.

**Prompt 2:**

**"Generate 4 German sentences using Konjunktiv I for indirect speech, related to patient communication in a hospital setting, suitable for B2 learners."**

**Risks:**

* The model may confuse Konjunktiv I with Konjunktiv II or use incorrect verb forms.
* Sentences might be too complex or too simple for the target level.
* Context may drift away from hospital/patient communication.

**Improvements/post-processing:**

* Add explicit instructions about Konjunktiv I verb forms (e.g., "use verb forms from Konjunktiv I only").
* Include example sentences as a guide in the prompt.
* Use tools like stanza for manual checks for correct Konjunktiv usage.
* Filter outputs based on a medical vocabulary list. (regex as I did in the script)

**Prompt 3:**

**"Write 5 relative clause (Relativsatz) sentences related to medication instructions, aimed at A2-B1 learners."**

**Risks:**

* The model might produce overly long or complicated relative clauses unsuitable for A2-B1 level.
* Sentences might not clearly highlight the relative clause or use uncommon vocabulary.
* Possible confusion between restrictive and non-restrictive clauses.

**Improvements/post-processing:**

* Request sentences with clear, simple relative clauses and specify relative pronouns to use (e.g., der, die, das).
* Limit sentence length in the prompt.

All the prompts have similar risks and improvements; most can be fixed with prompt engineering so maybe this helps (https://github.com/jujumilk3/leaked-system-prompts).