Instructions for evaluation of artificial clinical notes

You will be asked to evaluate 30 clinical notes which were generated by a large language model (LLM). Per note, you get to see three portions of text:

- The **prompt** which was provided to the LLM. This tells the LLM the characteristics of the patient who is visiting and asks it to generate a clinical note to describe this patient encounter. This is the only input the LLM receives.
- The **generated note**. This note contains a "history" section, describing the symptoms experienced by the patient along with some additional context, and a "physical examination" section, describing the findings of a (fictional) physical examination.
- A **compact version** of the previous note, also generated by the LLM. After generating the first note, we tell the LLM: "Please write this note in more compact style (using abbreviations and shortcuts), while preserving the content."

We ask you to evaluate these notes against four general dimensions:

1. **Consistency with the prompt**: The description of the patient's symptoms must correspond with the instructions provided in the prompt.
2. **Realism and relevance**: The LLM is allowed to invent context and details in light of the symptom information it receives, but this must be realistic and relevant to the symptoms experienced by the patient.
3. **Clinical accuracy**: The "physical examination" portion contains many clinical details invented by the LLM. These must be clinically accurate, both in a standalone fashion and in congruence with the patient's symptoms described in the "history" portion of the note.
4. **Quality of compact version**: The compact version of the note must correspond well with the original generated note, and must have good readability.

We now dive into the evaluation methods for each dimension.

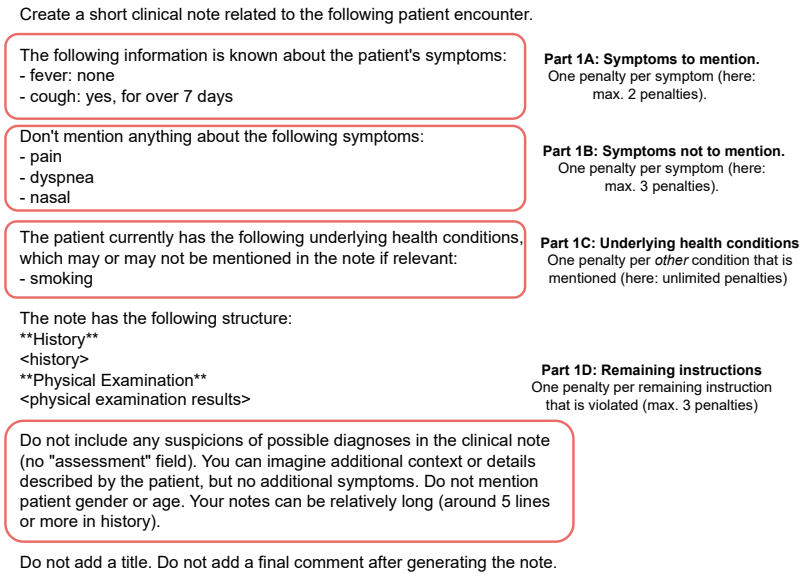

Figure 1: The different parts of the prompt for **respiratory notes (type 1)** and the potential penalties to assign. Not all parts are always present in the prompt, in which case you must assign 0 penalties for the missing part.

1 Consistency with the prompt

The description of the patient's symptoms in the note must correspond with the instructions provided in the prompt. There are two types of prompts, which we will evaluate a little differently.

1.1 Type 1: Respiratory note

Here, we generate a clinical note for a patient who is suffering from at least one respiratory symptom. There are three distinct parts in the prompt, as shown in Figure 1, which we will compare with the "history" portion of the note. We ask you to assign a penalty for each time the "history" portion of the note violates an instruction. You don't have to look at the "physical examination" portion for this part of the evaluation. We now describe what each part of the prompt means and how penalties are assigned, following Figure 1.

- **Part 1A: Symptoms to mention** This part of the prompt lists the symptoms that must be mentioned in the note. A symptom can either be present in the patient ("yes") or absent ("no"). If a symptom is present, the prompt may include a descriptor of the severity or nature of the symptom. We ask you to assign one penalty per symptom which is incorrectly

described in the note. A symptom is incorrectly described if its presence or absence is incorrectly listed in the text (e.g. the prompt says "cough: no" but the text mentions that the patient is coughing), OR the descriptor is not or incorrectly mentioned in the text (e.g. the prompt says "pain: yes, severe" but the text mentions only a mild pain). If both these things happen at once for the same symptom, you still assign one penalty. You can assign a maximum of one penalty per symptom that is listed in part 1A (i.e. if there are 4 symptoms listed in this part of the prompt, you can assign a maximum of 4 penalties).

- **Part 1B: Symptoms not to mention** This part of the prompt lists the symptoms that must *not* be mentioned in the note. We ask you to assign one penalty per forbidden symptom which is mentioned in the note in any way (even if the note mentions that the status of this symptom is unknown). Again, you can assign a maximum of one penalty per symptom that is listed in part 1B.

- **Part 1C: Underlying health conditions** Sometimes, the patient has underlying health conditions. If that is the case, this third part of the prompt lists these conditions. The LLM is not obligated to mention each health condition that is mentioned in this list. We therefore ask you to assign a penalty for each *other* health condition that is mentioned in the note, but which is not part of this list. With health conditions, we mean previous diagnoses or health issues, such as asthma, COPD, hay fever or smoking.

- **Part 1D: Remaining instructions** The last part lists some additional instructions: (1) no including suspicions of possible diagnoses, (2) no additional symptoms (but remember, adding context and descriptors for existing symptoms is allowed), and (3) no mention of patient gender or age. We ask you to assign one penalty per instruction in this list which is violated in the generated note, so a maximum of 3 penalties.

Not all notes contain all four parts. If the part is not present in the note, then you just assign 0 penalties.

1.2 Type 2: Unrelated note

Here, we generate a clinical note for a patient who is not suffering from any respiratory symptoms. The model is free to imagine an unrelated reason for why the patient might visit the doctor. Figure 2 shows the prompt, which now consists of only three parts.

- **Part 2A: Symptoms patient does not have** We tell the model that the patient is not experiencing any respiratory symptoms. We ask you to assign a penalty for each respiratory symptom which is incorrectly mentioned in the note (said to be experienced by the patient, when the

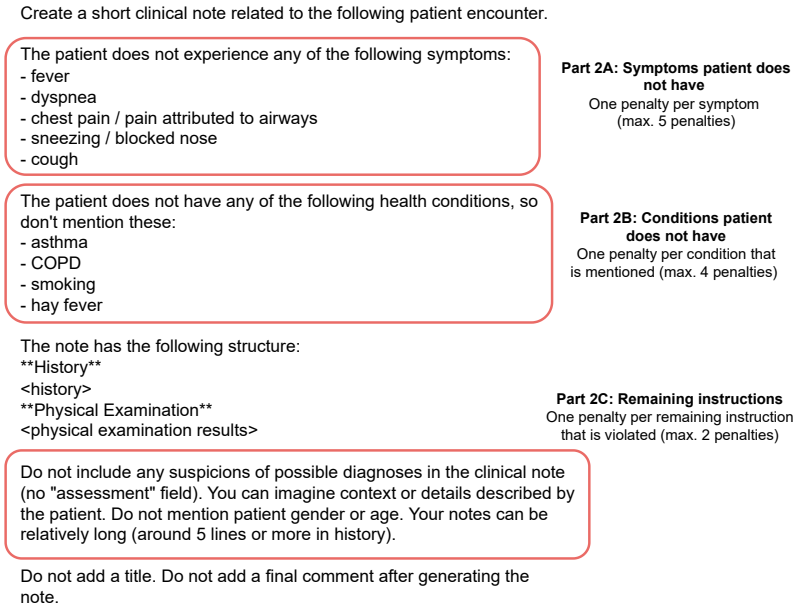

Figure 2: The different parts of the prompt for **unrelated notes (type 2)** and the potential penalties to assign.

prompt says it is absent). You can assign a maximum of 5 penalties for this part (one per symptom). The model is allowed to imagine other symptoms in this case (e.g. stomach pain, a sprained ankle...), so we don't assign penalties for those.

- **Part 2B: Conditions patient does not have** Here, we provide the model with a list of underlying health conditions that the patient does not have. We ask you to assign a penalty when one of these conditions is mentioned in the note, even though the prompt asked the LLM not to. You can assign a maximum of 4 penalties (one per condition).
- **Part 2C: Remaining instructions** The last part lists some additional instructions: (1) no including suspicions of possible diagnoses, and (2) no mention of patient gender or age. We ask you to assign one penalty per instruction in this list which is violated in the generated note, so a maximum of 2 penalties.

2 Realism and relevance

The LLM is allowed to invent context and details in light of the symptom information it receives, but this must be realistic and relevant to the symptoms experienced by the patient. While some clinical facts might not seem technically incorrect, you might not expect to see them in the note, or it might be unlikely that they would be written down by a real physician. For example, if a patient has a runny nose and no other complaints, most clinicians would not check for abnormalities in lung capacity. Another example is asking whether the patient has recently traveled to an exotic destination because they have a cough. These inconsistencies with reality make the note less useful and should be penalized. We ask you to give a score from 1-5 for the "history" and "physical examination" portions of the note separately. Keep in mind that we are specifically scoring the additional context and details imagined by the LLM on top of the specific symptoms mentioned in the prompt.

Scoring realism and relevance of "history", taking into account the information mentioned in the prompt:

5 – All pieces of additional context and details (i.e. outside of the symptoms and background provided in the prompt) are realistic and seem like they belong in the note.

4 – There are one or two pieces of additional context or details that I would not have mentioned as a physician, or that do not seem relevant (even though they do seem like they belong).

3 – There are one or two pieces of additional context or details that do not seem like they belong in the note, or do not seem relevant, given the symptoms and background provided in the prompt.

2 – There are multiple pieces of additional context or details that do not seem like they belong in the note, or do not seem relevant, given the symptoms and background provided in the prompt.

1 – (Almost) all of the additional context is nonsensical given the symptoms and background provided in the prompt.

Scoring realism and relevance of "physical examination", taking into account the information mentioned in the prompt and the "history" portion of the note:

5 – All elements in the physical examination are things I would check, given the history and symptoms of the patient, and no important elements are missing.

4 – There are one or two elements in the physical examination that I probably would not have checked, given the history and symptoms of the patient, but I could see it happen. Some minor elements might be missing, but nothing major.

3 – There are one or two elements in the physical examination that I would not have checked, or some important elements are missing, given the history and symptoms of the patient.

2 – There are multiple elements in the physical examination that make no sense given the history and symptoms of the patient, or many important elements are missing.

1 – The physical examination portion of the note seems totally unrealistic.

3 Clinical accuracy

The "physical examination" portion contains many clinical details invented by the LLM. These must be clinically accurate, both in a standalone fashion and in congruence with the patient's symptoms described in the "history" portion of the note. While the previous section talks about evaluating the realism of the presence of all elements in the "physical examination" section, here we talk about evaluating the clinical accuracy of these findings. Remember that the LLM is allowed to invent context and details, but only if this does not contradict the information in the prompt.

Clinical inaccuracies may depend on the context, like physical findings which are not congruent with the history and symptoms of the patient. For example, if the "history" portion mentions that the patient has a no fever, then this should not be contradicted in the "physical examination" portion with a temperature of 39°C. Clinical inaccuracies may also stand alone. For example, a blood pressure reading of 20/10 mm Hg is impossible to encounter in a patient.

Scoring clinical accuracy of the "physical examination" portion of the note:

5 – There are no mistakes, all reported clinical information is plausible in light of the patient's symptoms and history.

4 – There are one or two minor mistakes, or some details seem less plausible in light of the patient's symptoms and history, while the overall picture painted by the note is still correct.

3 – There are more than two minor mistakes, or multiple details which seem implausible in light of the patient's symptoms and history, but no major inaccuracies.

2 – There is a major mistake (on top of possibly some minor ones), or many details seem implausible given the patient's symptoms and history.

1 – There are multiple major mistakes and many details seem totally implausible given the patient's symptoms and history.

4 Quality of compact version

While all the previous evaluations concerned the original note, we also generated an additional compact version of the note. Here, we specifically asked the LLM to rewrite the note while using more abbreviations and shortcuts, using the following prompt: "Please write this note in more compact style (using abbreviations and shortcuts), while preserving the content." We purposefully want these notes to be harder to read and understand for both humans and machines, mimicking the complexity of some real doctor's notes.

The **content** of the compact version should convey the same information as the original text, albeit in a shorter format. We ask you to evaluate this using the scoring system below. You can assign only one score for the whole note (so jointly for "history" and "physical examination").

5 – The compact version conveys the exact same information as the original text.

4 – The compact version conveys all key points of the original text, leaving out some details here and there.

3 – The compact version conveys some of the key points of the original text, but misses some as well.

2 – The compact version conveys some of the same information as the original text, but misses many key points.

1 – The compact version does not convey the same information as the original text, leaving out almost all key points.

While we do specifically ask for the use of abbreviations, sometimes the language model uses so many that the text becomes unreadable. We evaluate **readability** using the scoring system below. You can assign only one score for the whole note (so jointly for "history" and "physical examination").

5 – The compact version seems understandable without seeing the original.

4 – The compact version seems mostly understandable without seeing the original, though there are some abbreviations that I would not immediately understand.

3 – Some parts of the compact version seem understandable without seeing the original, but other parts are not. There are some abbreviations that seem far-fetched or are used incorrectly (i.e. these are known to refer to other clinical terms than the way they are used in the text).

2 – Many parts of the compact version would not be understandable without seeing the original. Many abbreviations seem far-fetched or are used incorrectly (i.e. these are known to refer to other clinical terms than the way they are used in the text).

1 – The compact version is impossible to understand without seeing the original.

5 Frequently Asked Questions

We now answer a set of frequently asked questions to take away some confusion.

- **Why does the prompt explicitly ask not to mention gender and age?** We ask not to mention patient gender or age, because preliminary testing revealed that the LLM often used the same age and gender (34-year old woman), which could confound or bias the notes.
- **Why does the prompt ask for a patient without any symptoms?** Sometimes, it may occur that part 1 of the prompt (the list of symptoms to mention) only contains negative symptoms (all are "no"). These prompts are intentionally part of the experiment. While the notes resulting from such prompts might be confusing to read, since they don't mention an actual reason for the patient's visit to the doctor, we still ask you to evaluate the note in the same way as all other notes. Especially regarding the section "realism and relevance", we still ask you to **evaluate the realism of the additional context and details imagined by the LLM, apart from the symptoms requested in the prompt**, even if those requested symptoms seem unrealistic to start with.
- **If something is clinically inaccurate, doesn't that also make it unrealistic?** There is indeed some overlap between these concepts. In the "realism" evaluation, we ask you to evaluate the realism of the presence of all the elements in the "physical examination" portion of the note, taking into account the history of the patient. In the "clinical accuracy" evaluation, we ask you to assess the clinical accuracy of these findings. If something is unrealistic, it might still be accurate (e.g. it might not make sense to check for bloating if the patient has coughing as a primary complaint, but observing no bloating would not be clinically inaccurate in this case). If something is inaccurate (e.g. blood pressure being an implausible value), then the presence of that measurement in the "physical examination" may still be realistic, even if its value is incorrect. Please try to make this distinction in your evaluation.