

1 Supplementary materials

2 Examples of hallucinations in the medical output text:

3 Fabricated Facts/Negations

4 Facts (diagnostics, plan, communication, ...) that are completely fabricated (were never mentioned)

Type	Reason	Sentence
Minor	Clinical assessment not explicitly mentioned but accurate	Impression or clinical assessment: Chronic cough with sputum production and exertional dyspnea, possibly exacerbated by smoking, with a history of asthma.
Major	This is not in the input text shown	Follow up: Urgent referral to emergency department for further evaluation and management.
Minor	Not in the transcript	Treatment plan: Rest, continue with Advil as needed for pain relief.
Minor	Only neck mentioned, no other joints	Pertinent review of systems mentioned: No stiffness or loss of motion in joints.

5

6 Negated Events

7 Clinical events (e.g., symptoms, negations, surgeries, diagnostics) that were mentioned but not
8 recorded

Type	Reason	Sentence
Major	Patient reports tiredness and feeling hot	No other significant symptoms.
Minor	Has had attendance due to peanut allergy	She has not had any surgeries or hospitalisations.

9

10 Context Conflation

11 Topics that were mentioned but mixed up in the note.

Type	Reason	Sentence
Minor	Exposure to mold and dust is at home and asbestos exposure in living and workplace.	Family/Social History: Works in crowded, poorly conditioned workplace with potential mold, dust, and asbestos exposure.

Minor	The cough preceded the shortness of breath and has been over three weeks.	History of presenting complaint: Shortness of breath and coughing started two weeks ago, worsening over time.
Minor	The past medical history is epilepsy. The missed medication is related to the current presentation.	Past Medical History: Epilepsy, seizure disorder, infrequent seizures; missed medication dose.

12 Assumed causality

13 Made-up causality between two things that were mentioned independently

Type	Reason	Sentence
Major	I can't see in the text that the physician ascribes the pain as 'stress related'.	Impression or clinical assessment: Chest pain of unknown origin, possibly stress related.
Major	Shouldn't make associations without being explicitly mentioned in the transcript	Impression or clinical assessment: Possible upper respiratory tract infection or COPD exacerbation, considering significant smoking history and environmental exposures.
Minor	No mention of smoking exacerbating dyspnoea	Impression or clinical assessment: Chronic cough with sputum production and exertional dyspnea, possibly exacerbated by smoking, with a history of asthma.

14 Examples of omissions in the medical output text and the section it occurred:

15 Current Issue

Type	Reason	Sentence
Major	Palpitations not mentioned. This points the diagnosis towards Pulmonary embolism/infective cause rather than cardiac	Yes, since this morning, it's felt like my heart is racing. Okay. Have you noticed any skin changes at all? Any rashes? No rashes. Okay.
Minor	Blood in sputum important	And has there been any kind of green, slimy mucus in your cough at all? Uh, no, it's been a dry cough. Dry cough. Okay. And have you noticed at all coughing up blood?

Minor	Need to mention if any change in smell/taste as may indicate COVID infection	Uh, no, I haven't noticed that. What about your sense of taste? Are you being, are you able to taste things? Yeah, I think so. Okay.
-------	--	--

16

17 Information and Plan

Type	Reason	Sentence
Major	Advice: if the pregnancy test is positive, call straight back to organise be seen. Important safety net	So if you can do that today and if it's positive we need to see you straight away so give us a call back. Okay. Have you got any other questions? Right.
Minor	Explanation of shortness of breath important to include to gauge severity of this	Oh yeah, it's like when I'm doing walking, and I do exercise now and then, so it hasn't been... Super bad, like it hasn't stopped me from doing any of those things, but it's been pretty unusual and I'm a little bit worried.
Minor	Discussing diagnosis	So yeah, based on what we discussed today, it seems that you may have an upper respiratory tract infection.

18

19 PMFS Issues

Type	Reason	Sentence
Major	No allergies is an important negation to record	Okay. Right. Any allergies? Nothing. Yeah. Nothing. No allergies at all. Right here. Fine.
Major	Need to mention smoking history as this is important in someone presenting with chest pain. Is a risk factor for myocardial infarction etc	And do you use any recreational drugs like marijuana? Never. Okay. And do you smoke? Yes, I smoke about two packs of cigarettes per day.
Minor	Smoking status in an asthmatic should be recorded	No, I don't smoke. Do you drink any alcohol regularly? Yeah, one glass. How often is it? One glass of what? One glass of wine. Every... how often did you say?

20

21

22

23 **Prompts and note generation**

24

25 We used various approaches and prompts in our experiments. Each approach contributed valuable
 26 insights into refining both the process and output of LLM-based summarisation, allowing us to use the
 27 model with the best output for our clinical product. All prompts we used are listed in supplementary
 28 data 1.

29

Initial prompts

Experiments 1 and 2 were the initial prompts we had prior to utilising the CREOLA framework provided in the table below. These prompts served as baselines to our experiments before they were replaced.

Structured prompts

These were organised into three components: the base (core query), the template (context or format instructions), and the style (tone or level of detail). The purpose of this structured design was to address feedback on our outputs by changing specific sections of our prompt.

Experiments 3, 7, and 8 utilised the structured prompt. In experiment 7, we improved the base note to generate content from a first-person perspective. In experiment 8, we revised the style section of our SOAP note by incorporating negations and other style modifications. Additionally, we included "unknown" as an option for the status of problems not described in the transcript. Experiment 4 used the baseline SOAP note with an additional sentence to improve medication records.

Atomisation

In this model, atomic facts were extracted from our input text, which enabled precise extraction and organisation of clinical details. In experiment 5, atomisation was carried out to extract facts from the transcript. This process involved breaking the transcript down into individual atomic facts, which were then used to create a clinical note. The description did not provide a specific example. The procedure combined an adverb system prompt with instructions written in the third person.

Atomisation of clinical facts		LLM note generation
Facts: - the patient has a cough. - the patient's cough is productive. - the patient brings up gunk when coughing. - the gunk is green. - the gunk is sometimes yellowish. - the patient has no blood in the cough. - the patient uses a couple of Kleenexes a day for the sputum. - the patient has had a cough for the past few years. - the patient feels the cough is getting worse. - the patient coughs throughout the day. - the patient's cough is worse in the morning. - the patient cannot think of anything that triggers the cough. - the cough has been getting worse over the last few years. - the cough's worsening has been slow. - the patient has had episodes of cough with sputum for the last few years. - taking a deep breath makes the patient's cough worse. - nothing helps the cough go away. - the patient has no runny nose. - the patient has no sore throat. - the patient has no difficulty breathing. - the patient has not traveled outside of the country. - the patient has no fever. - the patient has no chills. - the patient has no night sweats. - the patient has increased fatigue. - the patient has been feeling more tired than usual in the last few months. - the patient's hands are always a little swollen. - the patient's feet are always a little swollen. - the patient's ankles are always a little swollen. - the patient has gained a few pounds over the years. - the patient has no chest pain. - the patient has trouble breathing when walking a few blocks or going up a few flights of stairs. - the patient has urinary issues at nighttime.	- the patient has to get up a lot at night to use the bathroom. - the patient has no bowel issues. - the patient's sleep has been poor. - the patient does not wake up during the night to take a breath. - the patient uses one pillow. - the patient has lost five pounds in the last few months. - the patient has not been trying to lose weight. - the patient's appetite has been good. - the patient has no recent infections. - the patient has no runny nose. - the patient has no sore throat. - the patient has not been exposed to anyone with tuberculosis. - the patient had a normal tuberculosis test for work. - the patient's blood pressure has been high. - the patient was diagnosed with high blood pressure in the last few years. - the patient takes Lisinopril for blood pressure. - the patient has no known medication allergies. - the patient has not been hospitalized. - the patient had a c-section for the birth of her daughter. - the patient was in the hospital for a few days for the c-section. - the patient has had no other surgeries besides the c-section. - the patient's father had a heart attack at age 78. - the patient's father passed away from the heart attack. - the patient has no family history of cancer. - the patient has no known family history of asthma. - the patient lives in an apartment. - the patient lives with her husband and daughter. - the patient works at the local grocery store. - the patient eats one or two meals a day. - the patient sometimes snacks throughout the day. - the patient's diet includes a variety of meats and vegetables. - the patient does not exercise currently. - the patient does not use recreational drugs. - the patient drinks a glass of wine three or four nights a week. - the patient drinks five or six ounces of wine per glass. - the patient smokes cigarettes. - the patient smokes one to two packs of cigarettes per day. - the patient has been smoking for nearly 40 years.	History. - Allergies: No known medication allergies. - Medications: Lisinopril for high blood pressure. - History of presenting complaint: Chronic productive cough for years, worsening slowly, worse in the morning, exacerbated by deep breaths, no identifiable triggers, associated with increased fatigue and weight loss without trying. - Past Medical History: High blood pressure, c-section. - Family/Social History: Father had a heart attack and passed away at 78, no family history of cancer or asthma, lives with husband and daughter, works at a grocery store, diet includes meats and vegetables, no exercise, smokes 1-2 packs of cigarettes daily for nearly 40 years, drinks wine 3-4 nights a week. - Sensitive information: [NOT MENTIONED]. Observations: - Examination findings: [NOT MENTIONED]. - Investigation results: [NOT MENTIONED]. - Impression or clinical assessment: Chronic cough likely related to smoking, with symptoms suggestive of chronic bronchitis or early chronic obstructive pulmonary disease (COPD), considering the history of smoking and productive cough. Differential diagnosis includes heart failure, given peripheral edema, weight gain, and dyspnea on exertion. Plan: - Planned investigations: Chest X-ray, pulmonary function tests, complete blood count, basic metabolic panel, echocardiogram to assess for heart failure. - Follow up: Schedule follow-up appointment after investigations are complete. - New prescribed medication or therapies: [NOT MENTIONED]. - Communication, reassurance & patient understanding of care: Explained the need for investigations to determine the cause of the cough and associated symptoms. Discussed the impact of smoking on health and the potential benefits of cessation. - Actions for referrer/GP: [NOT MENTIONED]

Supplementary Table 1: Atomisation of clinical facts (a) and LLM note generation (b)

The table shows the extraction of atomic facts from the transcript, followed by note generation by the LLM.

Figures 9 and 10 illustrate how the use of atomic facts prior to note generation leads to a significant increase in hallucinations and omissions compared to our base and structured prompts (experiments 1 and 3, respectively)

Function call and updated style block

A **function call** is the process of executing a specific function in a program by providing its name and the required input arguments. When called, the function performs its defined task and returns a result or performs an action.

In these experiments, the LLM was instructed to generate responses in a structured JSON format rather than free text. This output format was chosen due to its compatibility with clinician note summarisation in primary care. Experiments 6, 9, 10, and 11 were conducted using this format.

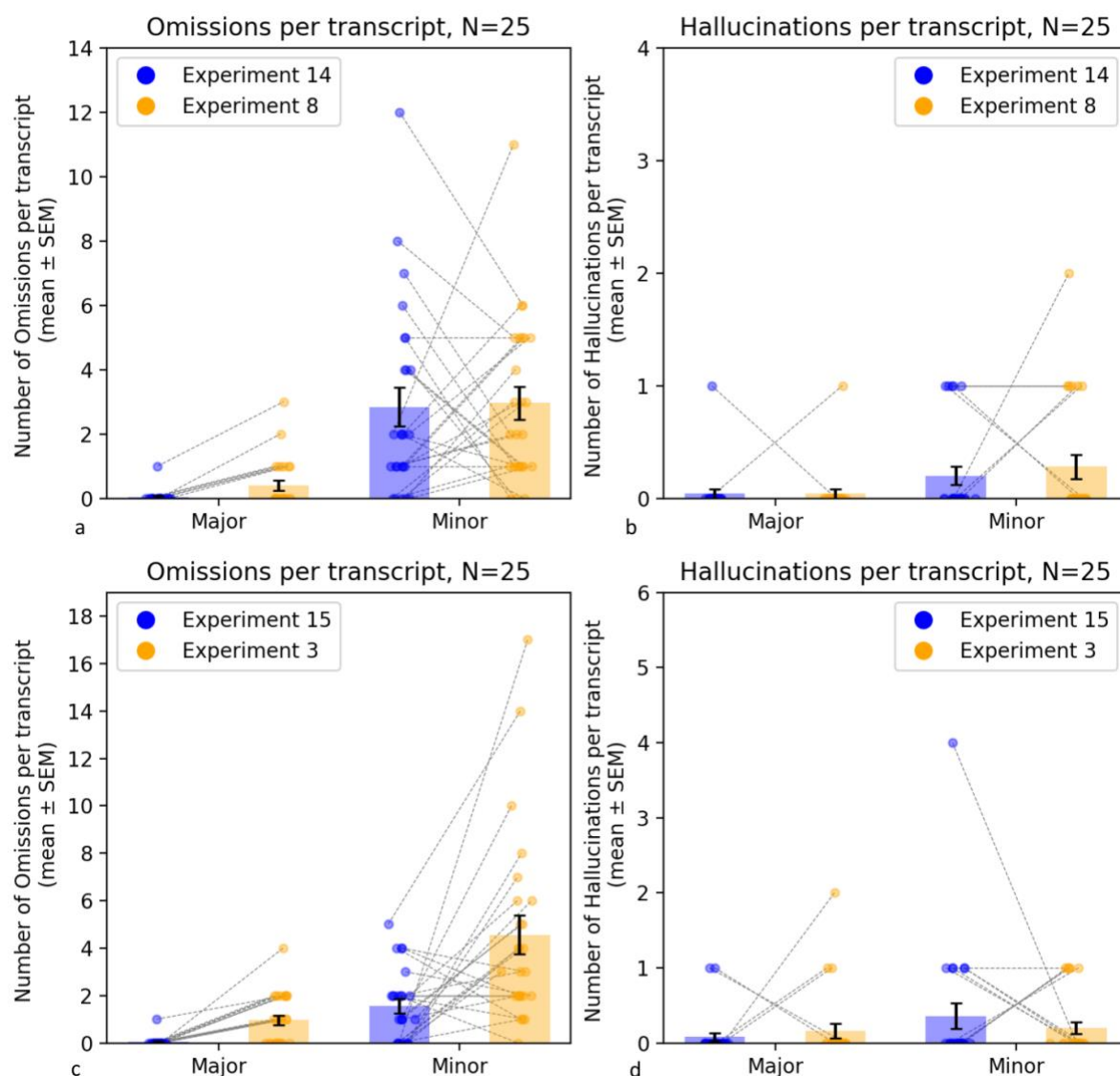
Experiment 9 aimed to test an updated style block for our EMIS (Educational Management Information System, which is an electronic health record system used widely in primary care settings in the United Kingdom) note, incorporating negations and other stylistic enhancements. We also revised several descriptions in the function schema. The block featured an EMIS note with the new style, which included negations and improvements such as conciseness. In experiment 10, we tested the EMIS function call with enhanced descriptions for examination and comments. This refinement aimed to ensure that only information present in the transcript was included. Experiment 11 was a continuation of experiment 10, and utilised the EMIS function call with improved descriptions and additional style guidance, specifically to avoid including information outside of the transcript.

Structured prompt + LLM revision step

In these experiments, we added an additional LLM call in order to review and refine its initial output to meet the requirements of a structured prompt, aiming to enhance the overall quality of the output.

In experiments 14 and 15, we applied this method. In experiment 14, we built upon the baseline established in Experiment 11 by adding a revision step after updating the state-of-the-art (SOTA) SOAP block. This revision included modifications to incorporate negations and other style enhancements, such as being more concise. Additionally, we introduced "unknown" as an option for the status of problems not described in the transcript. Experiment 15 involved adding a checking step to the 'Bad SOAP' note block, which was a SOAP note containing numerous hallucinations and omissions.

As noted in the supplementary Figure 1, experiment 14 had a slightly lower number of major omissions and minor hallucinations compared to experiment 8. Experiment 15 had fewer major hallucinations and omissions compared with experiment 3.

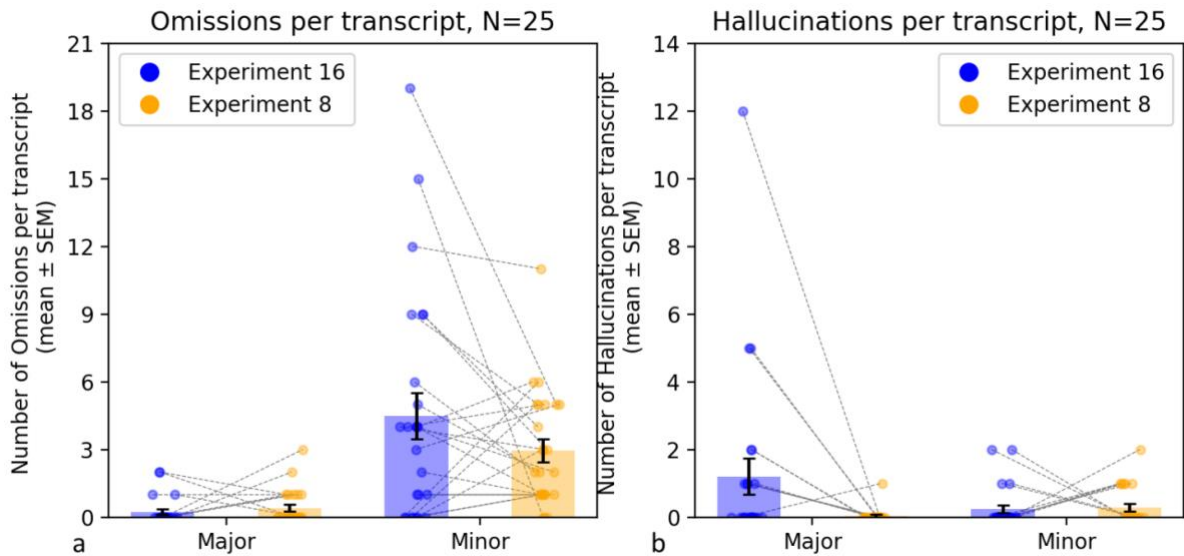


Supplementary Figure 1. Impact of adding an LLM revision step on the incidence of omissions and hallucinations

The figure illustrates the comparison between experiments 14 and 8 on the occurrence of omissions (a) and hallucinations (b), similar to the comparison between omissions (c) and hallucinations (d) between experiments 15 and 3.

New approach to note generation

In experiment 16, we developed a new method for generating customised outputs based on a specific template. We used our SOAP (Subjective, Objective, Assessment, Plan) note as the template and compared it to the baseline prompt output (from experiment 8). Interestingly, this comparison resulted in an increase in the number of major hallucinations and minor omissions, as shown in Supplementary Figure 2.

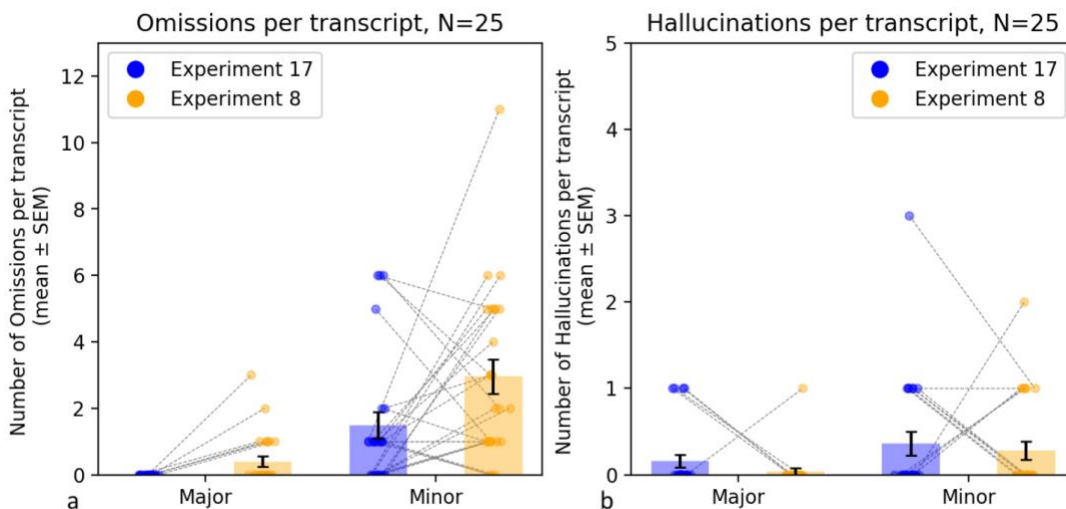


Supplementary Figure 2: Comparing hallucinations and omissions between our baseline prompt and customised output

The figure illustrates the changes in omissions (a) and hallucinations (b) between using the baseline prompt and a prompt designed to generate a customised output

Information gathering experiments

In experiment 17, we asked several clinicians to create notes based on medical transcripts. We then evaluated these notes for the presence of hallucinations and omissions. The goal was to understand the extent of inaccuracies in human-generated notes and to compare them with those generated by LLM. The clinician-generated notes exhibited slightly more major hallucinations and fewer omissions, as shown in Supplementary Figure 3.



Supplementary Figure 3. Comparing hallucinations and omissions between LLM-generated notes and human-generated notes

This Figure illustrates the number of major and minor omissions (a) and hallucinations (b) between LLM-generated notes in experiment 8 and human-generated notes in experiment 17

Supplementary Data 1: Rational for different experiments showing various workflows and prompts