Benchmarking and Generating GP Consultation Summaries

1 Introduction

General practitioners (GPs) are the first point of contact for most patients in the Netherlands and play a central role in the healthcare system. However, rising patient demand combined with workforce shortages has increased the pressure on general practices [1]. Recent surveys show that 68% of Dutch GPs consider their workload too high [2]. A substantial portion of this pressure stems from administrative tasks, particularly the work surrounding electronic health records (EHRs) [3]. One time-consuming activity is documenting each consultation in the patient's record.

Ambient scribing: LLM-powered automation of clinical summaries In recent years, new tools have emerged to reduce the administrative workload associated with writing consultation summaries. These tools, known as ambient scribes, use speech-to-text models to transcribe a consultation in real time, after which a large language model (LLM) generates a summary of the transcript for the GP to review. Studies suggest that ambient scribing can reduce documentation time and improve patient–provider communication, as clinicians no longer need to type during the consultation [4].

SOAP-format This project focuses specifically on the LLM component of this pipeline: converting an automatic transcript into a structured consultation summary. In Dutch general practice, summaries are typically written in the SOAP format, a widely used clinical documentation structure that organizes information into four sections [5]:

- **Subjective** — The patient's reported symptoms, concerns, and medical history..
- **Objective** — Observable findings, including physical-exam results and other measurable data.
- **Assessment** — The clinician's diagnostic reasoning based on the subjective and objective information.
- **Plan** — Follow-up steps, such as treatment, tests, referrals, and patient instructions.

See Figure 1 for an example of a transcript with corresponding SOAP note.

LLM-based clinical note generation To generate clinical notes in structured formats such as SOAP, several strategies can be employed. A first route is to use general-purpose foundation models (e.g., Gemini, OpenAI models) and rely on prompt engineering, structured outputs, or agentic workflows to produce notes in the desired structure [6]. A second route focuses on adapting open-source models such as LLaMA. Approaches such as continued pre-training, supervised fine-tuning, and reinforcement-learning–based optimization can improve factual accuracy and consistency with clinical documentation standards [7], [8].

Benchmarking clinical note generation Evaluating the summarization pipeline is crucial for assessing model performance throughout both development and deployment. Existing evaluation approaches can be broadly categorized into automatic evaluation (using predefined metrics or LLM-based judges) and human evaluation [9]. Because human assessment is time-consuming, rapid iteration during the prototyping phase relies primarily on automatic methods. However, these automatic metrics often show limited correlation with expert judgments [10], highlighting the need for evaluation measures that more closely reflect human clinical assessment.

Goal The objective of this project is to develop an LLM-powered application that generates a SOAP note from consultation transcripts. In addition, you will design an automatic benchmark (based on LLM-driven evaluation) to assess the performance of your SOAP-note generation pipeline, and propose a method to validate the alignment between automatic and human evaluation.

Example: Transcript

Doctor: Well, what can I do for you today?
Patient: I've had such a sore throat for three days now, and this morning I also had a bit of a fever, so I thought: yes, I'd better call just to be sure.
. . . .
Doctor: I'll also measure your temperature. Thirty-eight point three. Oh yes. You do have a bit of a fever. Have you taken any paracetamol or anything else for this?
Patient: Yes. This morning I took two paracetamol.
Doctor: Okay. But now it's three o'clock. So that has already worn off. You do indeed have a bit of a fever.
. . .
Doctor: That really points to an infection. So I think you have a throat infection. But based on these symptoms I can actually say that it's most likely a virus you've caught. And unfortunately, but fortunately in a way too, it's really just a matter of letting it run its course.
Patient: And what should I do about the pain?
Doctor: Yes. Good question. You should keep taking paracetamol regularly, and if that's not enough, you can also take ibuprofen. Paracetamol up to four times a day, 1000 mg each time, and ibuprofen three times a day, 400 mg each time.

Example: SOAP-note

Subjective: Sore throat for 3 days. Fever this morning. Took 2 paracetamol in the morning.

Objective: Temperature: 38.3 °C.

Assessment: Throat infection, most likely viral.

Plan: Explanation that it is most likely a viral throat infection and needs to run its course. Pain relief advice: paracetamol 4x1000 mg/day and ibuprofen 3x400 mg/day.

Figure 1: Example of a consultation transcript (input) and its corresponding SOAP note (output).

2 Research Questions

The following are the research questions we will aim to answer in this project:

RQ1: How can a new automatic benchmark for SOAP-note generation be developed using LLM-based evaluation methods?
RQ2: How can a study be designed to examine the correlation between this automatic benchmark and human expert judgments?
RQ3: How can a SOAP note be generated from a doctor–patient conversation transcript?

3 Experimental Setup

We will provide you with ten representative consultations from general practice, covering some of the most common complaints encountered in primary care. For each consultation, the following data will be included:

- Full transcript of the doctor–patient conversation
- Example SOAP-note (reference summary)
- For each SOAP category (Subjective, Objective, Assessment, Plan), a list of key facts that should be included in the corresponding section of the summary

All examples will be provided in Dutch, but you are encouraged to translate them to English to implement and evaluate your system in both languages. This allows you to explore cross-lingual generalization and compare model performance across Dutch and English data.

To answer RQ1

1. Construct an automatic benchmark for SOAP-note generation, inspired by the FineSurE framework [11]. FineSurE is a recently proposed method for automated, fine-grained evaluation of text summarization quality. Unlike traditional metrics such as ROUGE or BERTScore, which rely on lexical or embedding similarity, FineSurE evaluates generated summaries at the sentence and key-fact level. It uses large language models to perform two subtasks:

   a. *Fact checking*: identify factual errors in each sentence of the generated summary;

   b. *Key-fact alignment*: measure how well key information from the original transcript is captured in the summary.

   Based on these subtasks, FineSurE provides three interpretable evaluation dimensions: *faithfulness*, *completeness*, and *conciseness*. Each evaluation dimension is expressed as a percentage score.

2. Adapt FineSurE's methodology to the clinical summarization domain. Implement the two FineSurE evaluation tasks (fact checking and key-fact alignment) and tailor them to medical consultation transcripts. Evaluate each SOAP-note section (Subjective, Objective, Assessment, and Plan) individually to capture differences per category. Combine the individual SOAP-category scores into one overall benchmark score.

3. Evaluate and test the reliability of the benchmark by applying it to the provided consultation examples and analyzing whether the resulting evaluation scores are stable, interpretable, and consistent across cases. You may compare the benchmark based on both the Dutch and the English datasets to assess cross-lingual consistency.

To answer RQ2

1. Design a framework to study the correlation between automatic evaluation scores (from the benchmark created in RQ1) and human expert judgments of SOAP-note quality. Define how human evaluators, such as clinicians or advanced medical students, would assess the same SOAP notes based on three criteria: *faithfulness*, *completeness*, and *conciseness*.

2. Determine how to collect human evaluation data in a standardized and reproducible way. Specify the number of evaluators, the rating scale (for example, a five-point Likert scale), and how to ensure consistent interpretation of each evaluation criterion.

3. Describe the statistical methods that could be used to measure the alignment between the automatic and human evaluations.

4. You do not need to execute this study, but the framework should be detailed enough that another researcher could implement it. Include references to existing work on automatic evaluation and human–machine alignment in text summarization to justify your design choices. You may use the correlation analyses from [11] or [12] as methodological starting points for your framework.

To answer RQ3

1. Develop a large language model (LLM) pipeline that generates structured SOAP notes from doctor–patient conversation transcripts. Focus on using inference-only methods (e.g., prompt engineering, few-shot exemplars, output-schema constraints), without any fine-tuning.

2. Compare at least one open-source and one closed-source model. Examples of open-source models include `Gemma-3` or `Qwen03`, while closed-source options include `OpenAI GPT-5` or `Gemini 3`. Closed-source models can be accessed via API, or smaller variants can be used in Google Colab if computational resources are limited.

3. Apply the benchmark developed in RQ1 to evaluate your pipeline. Evaluate the results for both Dutch and English data to explore cross-lingual generalization.

4. Reflect on how the results from the benchmark evaluation can guide further improvements. Discuss whether your benchmark successfully captures clinically relevant differences between model outputs and what limitations remain when applying automatic evaluation to this task.

References

[1] B. Baird, A. Charles, M. Honeyman, D. Maguire, and P. Das, *Understanding pressures in general practice*, 2016.

[2] E. Gumas, M. Gunja, A. Shah, and R. Williams, "Overworked and undervalued: Unmasking primary care physicians' dissatisfaction in 10 high-income countries—findings from the 2022 international health policy survey," *Commonwealth Fund. Aug*, vol. 10, 2023.

[3] B. G. Arndt, M. A. Micek, A. Rule, C. M. Shafer, J. J. Baltus, and C. A. Sinsky, "More tethered to the ehr: Ehr workload trends among academic primary care physicians, 2019-2023," *The Annals of Family Medicine*, vol. 22, no. 1, pp. 12–18, 2024.

[4] R. C. van Linschoten, I. M. van Loon, L. Joanknecht, and E. W. Bischoff, "Ambient scribe in general practice: A multi-perspective before-after longitudinal mixed-methods study," *medRxiv*, pp. 2025–10, 2025.

[5] V. Podder, V. Lew, and S. Ghassemzadeh, *Soap notes*, StatPearls [Internet], Updated 2023 Aug 28. Available: https://www.ncbi.nlm.nih.gov/books/NBK482263/, Treasure Island, FL, 2025.

[6] E. Vedadi, D. Barrett, N. Harris, *et al.*, "Towards physician-centered oversight of conversational diagnostic ai," *arXiv preprint arXiv:2507.15743*, 2025.

[7] H. Wang, C. Gao, B. Liu, *et al.*, "Towards adapting open-source large language models for expert-level clinical note generation," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 12 084–12 117.

[8] H. Y. Leong, Y. Gao, and S. Ji, "A gen ai framework for medical note generation," in *2024 6th international conference on artificial intelligence and computer applications (ICAICA)*, IEEE, 2024, pp. 423–429.

[9] H. Wang, R. Yang, M. Alwakeel, *et al.*, "An evaluation framework for ambient digital scribing tools in clinical applications," *npj Digital Medicine*, vol. 8, no. 1, p. 358, 2025.

[10] A. B. Abacha, W.-w. Yim, G. Michalopoulos, and T. Lin, "An investigation of evaluation methods in automatic medical note generation," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 2575–2588.

[11] H. Song, H. Su, I. Shalyminov, J. Cai, and S. Mansour, "Finesure: Fine-grained summarization evaluation using llms," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 906–922.

[12] Y. Lee, T. Yun, J. Cai, H. Su, and H. Song, "UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3941–3960. DOI: 10.18653/v1/2024.findings-emnlp.227. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.227/.