Master Thesis

# Generating Follow-up Questions in Health Conversations Using Fine-tuned Language Models

Xin Chen

| | |
|---|---|
| Supervisor | Piek Vossen; Luís Morgado da Costa |
| $2^{nd}$ reader | Lucia Donatelli |

*a thesis submitted in fulfillment of the requirements for the degree of*

**MA Linguistics**

(Text Mining)

**Vrije Universiteit Amsterdam**

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

| | |
|---|---|
| Date | June 27, 2025 |
| Student number | 2836834 |
| Word count | Word count |

# Abstract

This thesis proposes a novel approach to generating follow-up questions in health dialogues by fine-tuning large language models on synthetic, ICF-guided data. A customized Self-Instruct pipeline was implemented to construct a synthetic corpus of multi-turn patient–clinician conversations, grounded in four mobility-related ICF categories (D420, D445, D465, D470). Each dialogue incorporates randomized patient personas and severity levels to ensure diversity and clinical plausibility while avoiding privacy concerns. The Qwen3-8B model serves as the backbone, fine-tuned on the synthetic dataset and evaluated alongside the base model under zero-shot and few-shot prompting conditions. Automatic metrics (BLEU, ROUGE-L, BERTScore) show that the fine-tuned model produces follow-up questions more closely aligned with human-authored gold references. By contrast, both human evaluation and LLM-based G-eval assessments favored the base model, which tended to generate longer and more conversationally engaging questions despite being less clinically structured. This discrepancy reveals a tension between automatic metrics, which emphasize lexical and semantic similarity, and expert judgments, which prioritize naturalness, coherence, and clinical appropriateness, while also exposing the inherent subjectivity underlying human evaluation. Overall, the findings highlight both the potential and the limitations of synthetic, ontology-guided training for medical dialogue modeling, underscoring the need for more nuanced evaluation frameworks and hybrid training strategies in future research.

# Declaration of Authorship

I, Xin Chen, declare that this thesis, titled *Generating Follow-up Questions in Health Conversations Using Fine-tuned Language Models* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: <20/08/2025>

Signed: <Xin Chen>

# Acknowledgments

I would like to express my sincere gratitude to Piek Vossen and Luís Morgado da Costa for their supervision and insightful feedback throughout this journey. I am deeply thankful to the A-proof Team—Edwin, Sabina, Marike, and Jesse—for their invaluable assistance and patience during the human evaluation of both the synthetic training data and the final results. I also greatly appreciate the medical students from the Amsterdam University of Applied Sciences, whose efforts were instrumental in creating test data. Special thanks to my mom, who calls me every day to gently remind me not to overwork myself, and to my friends Shenglin, Shutao, Melina, Matt, and Naomi, who have shared in my struggles and provided me with incredible companionship. And to my boyfriend, Abel, who patiently listens to my endless rants while gaming. Your support has truly kept me sane.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

Effective health monitoring requires more than episodic check-ups; it depends on continuous, longitudinal tracking of patients' daily activities, symptoms, and emotional states. This kind of longitudinal monitoring is vital for preventing serious illness and enabling timely interventions. However, geographical and temporal constraints often limit access to frequent in-person consultations. Conversational agents (CAs) offer a promising solution to bridge this gap. Deployed via smartphone apps, websites, or messaging platforms, CAs are able to conduct natural multi-turn dialogues with patients, eliciting details about symptoms, medications, and functional ability to detect changes over timeTudor Car et al. (2020). By improving accessibility, personalization, and patient engagement, these agents have the potential to enhance healthcare delivery and patient monitoring. A central requirement for effective health-related dialogue is the ability to ask relevant follow-up questions (FQs). Follow-ups probe deeper into patient responses, clarify ambiguities, and ensure coverage of clinically important topics that may not be addressed by initial prompts.

However, despite the rapid progress in large language models (LLMs) for dialogue, generating high-quality follow-up questions in healthcare remains challenging. Off-the-shelf LLMs produce fluent and coherent outputs, but they often default to generic or overly verbose questioning styles that lack clinical focus. For example, Gatto et al. (2025) found that unadapted LLMs frequently failed to generate follow-up questions aligning with those written by real healthcare providers. Similarly, Abbasian et al. (2023) reported that current LLM-based medical chatbots tend to produce generic, non-individualized responses when they lack access to patient-specific data. Such generic outputs can miss important clinical details, limiting the dialogue's utility for monitoring patients. What is needed are methods to systematically adapt these models to healthcare dialogue demands so that follow-ups are personalized to the patient's context, grounded in relevant medical knowledge, and clinically meaningful.

In this regard, Vossen et al. (2024) introduced a proof-of-concept conversational agent designed to engage patients in multi-turn dialogues and construct structured diaries of their lives. In their framework, the diary is represented as an episodic knowledge-graph timeline of events, providing contextualized insights into patients' physical, social, and mental functioning for health monitoring. Although the prototype followed a scripted interaction strategy, the work demonstrated the feasibility of integrating dialogue with structured representations of personal health and functioning.

Despite recent progress, existing methods for generating follow-up questions in healthcare are limited in several ways. They often lack personalization and comprehensive coverage, focusing narrowly on immediate symptoms while overlooking broader aspects of a patient's health. Comparative analyses of prompt-based versus fine-tuned model adaptation remain scarce, leaving the relative benefits of these strategies underexplored. Moreover, structured clinical frameworks have rarely been integrated into question generation, leading to gaps in domain coverage. These limitations motivate the present study.

## 1.2 Research question

Building on these foundations, the present study is guided by the following central and subsidiary research questions:

Central Question: **Can a fine-tuned generative language model improve the quality and contextual relevance of follow-up question generation in health-related conversations, particularly in assessing physical function and emotional well-being?**

To address this overarching question, four sub-questions are examined:

- Can a non-fine-tuned generative language model generate contextually relevant and logically sequenced follow-up questions under zero-shot or few-shot prompting?
- To what extent does few-shot prompting improve question generation compared to zero-shot prompting, across both fine-tuned and non-fine-tuned models?
- Does fine-tuning on ICF-based synthetic health dialogue data significantly enhance the contextual relevance and personalization of follow-up questions?
- How do fine-tuning and prompting strategies jointly affect performance across different functional categories defined by the ICF?

## 1.3 Research Methods and Evaluation

This study investigates how synthetic data generation, model fine-tuning, and prompting strategies affected the quality of follow-up question (FQ) generation in health-related conversations. The research focuses on evaluating the contextual relevance, coherence, and personalization of FQs in structured health assessments. To address the research questions, the methodology includes the following components:

Synthetic Data Generation: As an important contribution of this project, a synthetic dataset of health-related dialogues was created. The dialogues were generated using prompt-based techniques with large language models (a local LLaMA-3.1-8B) and structured according to the International Classification of Functioning, Disability and Health (ICF) framework. The dataset covered four key functional categories: D420 transferring oneself, D445 hand and arm use, D465 moving around using equipment, and D470 using transportation. All dialogues included contextually relevant follow-up questions and represented both physical and emotional health dimensions. This dataset served as the foundation for training.

Fine-tuning on Synthetic Health Dialogue: A pre-trained generative language model(Qwen3-8b) was fine-tuned on the synthetic dataset to improve the contextual awareness and medical relevance of the generated follow-up questions. Fine-tuning enables the model to better learn domain-specific structures and to generate more personalized and clinically meaningful responses.

Prompting Strategies: To evaluate inference-time strategies, both zero-shot and few-shot prompting were tested. In zero-shot settings, the model received task instructions and dialogue history only; in few-shot settings, example input–output pairs were provided in the prompt. These prompting methods were applied to both the fine-tuned and non-fine-tuned models, resulting in four system configurations under a unified experimental setup. This approach enables a systematic comparison of how synthetic training, model adaptation, and prompt design jointly affected the generation of high-quality follow-up questions in the context of AI-driven health dialogue systems.

Evaluation: To assess the quality of the generated follow-up questions, multiple evaluation strategies were employed. Automatic evaluation compared model outputs to reference questions created by medical students (and their paraphrased variants), using metrics such as BLEU, ROUGE, and BERTScore to measure lexical and semantic similarity. Model-based evaluation was conducted with a large language model, which rated each question for relevance and faithfulness. Finally, a small-scale expert evaluation was carried out, in which medical professionals assessed the clinical appropriateness, relevance and ICF-alignment of the generated questions. This expert review ensures clinical reliability and alignment with real-world health assessment practices.

## 1.4 Organization of Chapters

The remainder of this thesis is organized as follows. Chapter 2 reviews related work on follow-up question generation and medical dialogue systems. Chapter 3 outlines the methodology, including dataset construction, model fine-tuning, and evaluation design. Chapter 4 presents the experimental results from automatic metrics, model-based evaluations, and expert assessments. Chapter 5 discusses the limitations of this study and provides some wrong case analysis. Finally, Chapter 6 concludes with key insights and directions for future research.

# Chapter 2

# Related Work

## 2.1 Follow-up Question Generation in NLP

Recent advances in large language models (LLMs), such as GPT-3, LLaMA, and Alibaba's Qwen, have dramatically expanded the capabilities of natural language generation; these models are trained on massive text corpora and demonstrate strong abilities in understanding and generating human-like language across a wide range of domains and tasksZhao et al. (2023). In particular, their general-purpose design enables them to produce contextually appropriate, coherent, and informative responses, making them well-suited for dialogue-based applicationsDong et al. (2024). Within this broader framework of response generation, follow-up question generation emerges as a specific and cognitively demanding task that leverages the model's contextual reasoning abilities to extend conversations in a purposeful manner.

Follow-up question generation refers to the task of creating questions that logically and naturally continue a dialogue based on preceding interactions. In contrast to factoid question generation, which usually focuses on generating short, factual questions from isolated text passages, follow-up questions require a deeper understanding of dialogue context and user's intent in order to formulate a relevant queryDong et al. (2024). Factoid QG systems are often trained on reading comprehension datasets, where the goal is to generate questions answerable by specific spans in the text, such as names, dates, or definitions Du et al. (2017). These models emphasize syntactic alignment and semantic matching with the source text, and are commonly used for tasks like quiz generation and machine reading evaluation.

Conversational follow-up questions, on the other hand, go beyond surface-level fact extraction. They must be tightly grounded in the flow of prior conversation, leveraging both cognitive reasoning (e.g., inference, abstraction, or hypothesis generation) and contextual understanding (e.g., tracking discourse goals, maintaining coherence) to formulate inquiries that extend, clarify, or deepen the ongoing exchange. This positions follow-up QG as a more complex and interaction-oriented subtask of question generation, essential for maintaining engagement in multi-turn dialogues. As illustrated in datasets such as CoQA Reddy et al. (2019), effective follow-up questions are not mere repetitions of known details but build on the content of the preceding turn to probe deeper or explore related aspects. This often involves both higher-order reasoning such as inferring implicit information or applying domain knowledge and discourse-level awareness to ensure the question is both relevant and meaningful in the ongoing conversation.

Early research in question generation (QG) primarily focused on generating straightforward factoid questions from text, often within the scope of reading comprehension tasks. These systems generally assume the answer is known beforehand and generate questions targeting this known information. While suitable for assessing basic textual comprehension, this approach diverges significantly from human information-seeking behaviors, as it lacks genuine curiosity or exploratory reasoning. Humans typically ask follow-up questions to bridge knowledge gaps, explore deeper implications, or seek clarification.

Generating context-aware follow-up questions demands more advanced cognitive operations—including inferencing, analogy formation, and probing underlying assumptions—to effectively identify unknown or unclear aspects and craft meaningful inquiries. Recent studies, such as Meng et al. (2023), introduced the task of real-world follow-up question generation, highlighting that effective follow-up questions employ diverse pragmatic strategies (e.g., clarification, causal reasoning, analogy, etc.) and reflect higher-order cognitive skills compared to traditional factoid questions. Similarly, Prior work on clarification question generation for forums and product reviews has addressed specific aspects of this broader challenge, aiming to generate contextually relevant follow-up questions that disambiguate user-generated textMeng et al. (2023). Further studies have extended follow-up QG to contexts such as social media interactions and conversational surveys, emphasizing the dynamic and personalized nature required of follow-up questioningGatto et al. (2025). Collectively, these efforts underscore the critical distinction between simplistic question generation and contextually informed, information-seeking follow-ups that effectively propel conversational engagement.

## 2.2 Follow-up Questions in Medical Dialogue Systems

In clinical dialogue systems, the ability to ask relevant follow-up questions is crucial for effectively gathering complete patient information and clarifying ambiguities. Medical consultations are inherently interactive and multi-turn: Clinicians must sequentially pose targeted questions to clarify patient-reported symptoms, history, and other factors necessary for accurate diagnosis and treatment. Recently, researchers have increasingly adapted follow-up question generation techniques specifically for healthcare settings. For example, Winston et al. (2024) investigated GPT-4's ability to simulate medical history-taking by posing follow-up questions to patients. Their findings indicated that without specialized domain guidance, even highly capable general-purpose large language models (LLMs) frequently generate overly broad or verbose questions, potentially overwhelming patients and omitting critical clinical details. This underscores the need for domain-specific adaptation in leveraging LLMs effectively within healthcare consultations.

To overcome these limitations, researchers have adopted several domain-aware strategies. One prominent approach involves leveraging structured medical knowledge to guide question generation. Gupta et al. (2022) , for instance, employed the standardized PHQ-9 depression screening questionnaire as a template for generating clinically relevant follow-up questions on mental health forums. Their system checks which of the nine PHQ-9 symptom questions have already been addressed in the patient's post and systematically identifies the unanswered items, which are then expanded into targeted follow-up questions guided by clinical interview frameworks such as the SCID. By fine-tuning a model on this structured procedure, they ensured the generated questions were clinically relevant (e.g., asking about symptoms not yet discussed) and aligned with medical assessment standards. This template-driven strategy underscores how medical ontologies or standardized assessments can provide a backbone for question generation. However, beyond specialized contexts like depression screening, broader application of such structured frameworks remains relatively unexplored.

Recent advancements have further progressed toward richer, multi-turn medical dialogues and diagnostic goal-oriented questioning. Wang et al. (2024) proposed HealthQ, a framework for systematically evaluating an LLM's questioning ability over multiple turns in healthcare dialogues. By leveraging advanced prompting techniques (e.g. retrieval-augmented generation and chain-of-thought reasoning) and employing an LLM-based judge to rate question quality (specificity, relevance, usefulness), HealthQ provided the first systematic assessment linking high-quality inquiries to improved patient information.

Complementing this, Gatto et al. (2025) developed a multi-agent system that integrates background patient data with conversational context to generate personalized medical follow-up questions. Their system, FollowupQ, draws on patient electronic health records (EHR) alongside the patient's messages to ask targeted follow-up questions clarifying reported conditions. Tested in asynchronous telehealth scenarios, FollowupQ significantly reduced clinician workload and produced a dataset comprising approximately 2,300 clinician-authored follow-up questions directly linked to patient context.

## 2.3 Data Scarcity and Self-Instructional Data Generation

Despite the above progress, the field faces a persistent data scarcity problem. High-quality, large-scale corpora of contextually grounded follow-up questions, particularly those that capture multi-turn dynamics, diverse pragmatic strategies, and domain-specific constraints, remain limited. This challenge is especially pronounced in healthcare, where concerns over privacy, high annotation costs, and the limited availability of clinician time constrain both dataset size and diversity. Even valuable resources, such as the approximately 2.3k FollowupQ instancesGatto et al. (2025)) are modest compared with the data requirements of robust instruction-following models.

To address this challenge, Wang et al. (2022) proposed Self-Instruct, a general-purpose framework for synthetic instruction data generation. Traditionally, instruction-tuned language models relied on a small set of human-crafted prompts and responses, which constrained their generality due to the limited scale and creativity of human-generated examples. Self-Instruct mitigates this limitation by bootstrapping synthetic instructional data directly from a pretrained model. In this approach, a base model is prompted to generate novel instructions together with corresponding inputs and high-quality outputs, thereby reducing the reliance on large-scale human annotations. Generated instruction-output pairs are subsequently filtered to exclude low-quality or repetitive entries, ensuring only relevant examples are utilized in model fine-tuning. This "almost annotation-free" pipeline leverages the model's intrinsic knowledge, providing an efficient mechanism to expand task coverage while minimizing human effort. Its core assumption is that sufficiently advanced LLMs can generate diverse, high-quality instructional data, thereby enhancing dataset diversity and coverage.

For follow-up question generation, particularly in medical settings, self-instructional pipelines provide a promising avenue to bridge data gaps while respecting privacy and cost constraints. Such pipelines can: (i) seed the generator with domain schemas (e.g., symptom ontologies or screening frameworks) and dialogue contexts; (ii) elicit candidate follow-ups conditioned on missing or ambiguous information; (iii) apply automatic and human-in-the-loop filtering for clinical validity, redundancy, and safety; and (iv) iteratively refine the resulting pool for fine-tuning and evaluation. In sum, Self-Instruct-style data creation complements scarce human-labeled corpora and enables broader coverage of pragmatic strategies and multi-turn dynamics, both of which are essential for high-quality follow-up question generation.

## 2.4 Research Gap

The literature reviewed above demonstrates a clear progression in follow-up question generation, evolving from factoid-style approaches toward context-sensitive, multi-turn strategies. Recent studies have further extended this trajectory into medical dialogue systems, with growing attention to synthetic data generation as a means of addressing data scarcity. Within healthcare, preliminary efforts using prompting strategies, domain-specific datasets, and scripted dialogue frameworks have illustrated the feasibility of computational approaches for patient-centered questioning.

Despite these advances, several critical gaps remain. Personalization and comprehensive coverage are still insufficiently addressed, as current systems tend to focus narrowly on immediate symptom elicitation while neglecting broader aspects of patient well-being, including physical functioning, emotional health, and daily activities. Systematic comparisons between prompting-based methods and fine-tuning approaches in medical dialogue systems are lacking, leaving open questions about their relative strengths, trade-offs, and complementarities. Furthermore, although the integration of structured resources has been attempted in specific contexts (for example, the PHQ-9 questionnaire for depressionGupta et al. (2022)), the systematic use of established healthcare ontologies and functional frameworks such as the International Classification of Functioning, Disability and Health (ICF)World Health Organization (2001) to guide follow-up question generation remains rare. This gap limits both domain coverage and clinical interpretability. Finally, evaluation practices remain fragmented, as most existing studies rely heavily on automatic metrics with limited incorporation of clinically grounded expert assessments or LLM-based judgments.

This research addresses these gaps through an integrated approach. It leverages the ICF framework as a structured ontology to ensure comprehensive coverage across physical, emotional, and functional health domains. By synthetically generating dialogues grounded in the ICF framework, this study constructs a rich training corpus that captures a wide spectrum of patient scenarios and question types, thereby addressing the scarcity of labeled data in this domain. A Qwen3-8B model is subsequently fine-tuned on this synthetic dataset, enabling it to learn both the style and structure of ICF-guided, personalized follow-up questioning and how to formulate questions based on prior conversational context.

In parallel, this work advances evaluation methodologies by incorporating both LLM-based automatic evaluators and expert-informed metrics to assess the contextual relevance and correctness of generated follow-up questions within clinical dialogues. Through this integrated approach, the study situates itself at the intersection of natural language processing and health informatics, with the overarching goal of improving the quality of patient-provider interactions via personalized and logically coherent questioning. Ultimately, the proposed framework not only builds upon prior advances in medical dialogue modeling but also extends them by systematically integrating a health-centered ontology (ICF) and rigorously comparing prompting and fine-tuning strategies for this critical application domain.

# Chapter 3

# Methodology

## 3.1 Data collection

To provide a high-level summary of the overall research process before diving into data collection specifics, Figure 3.1 illustrates the key steps undertaken in this study. These include data collection for both evaluation and training, model adaptation via fine-tuning and prompting, and multi-faceted evaluation of the generated outputs.

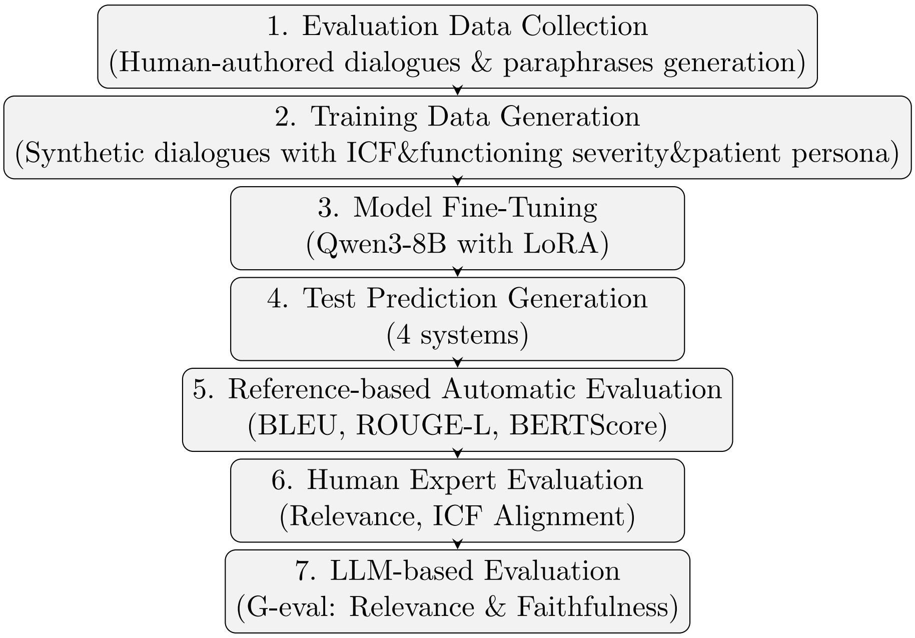

1. Evaluation Data Collection
(Human-authored dialogues & paraphrases generation)
2. Training Data Generation
(Synthetic dialogues with ICF&functioning severity&patient persona)
3. Model Fine-Tuning
(Qwen3-8B with LoRA)
4. Test Prediction Generation
(4 systems)
5. Reference-based Automatic Evaluation
(BLEU, ROUGE-L, BERTScore)
6. Human Expert Evaluation
(Relevance, ICF Alignment)
7. LLM-based Evaluation
(G-eval: Relevance & Faithfulness)

Figure 3.1: Overview of methodological steps

The training dataset employed in this study comprises synthetic, multi-turn medical dialogues, locally generated using the LLaMA3.1-8B-Instruct large language model in combination with prompt engineering techniques. The data generation process was guided by carefully crafted Self-Instruct-style prompts and enriched with structured patient persona attributes, including age, gender, medical condition, and clinician role. This synthetic data approach effectively mitigates the scarcity of publicly available clinical dialogue datasets.

The dialogues specifically focus on four categories from the International Classifica-

| ICF Code | Name |
|---|---|
| D420 | Transferring oneself |
| D445 | Hand and arm use |
| D465 | Moving around using equipment |
| D470 | Using transportation |

Table 3.1: ICF categories selected

tion of Functioning, Disability, and Health (ICF). These categories were selected based on their alignment with the clinical expertise of the annotators who constructed the test dataset. The chosen ICF categories are listed below:

Focusing on these specific ICF categories ensures consistency and domain coherence between the training dialogues and the clinical scenarios represented in the test dataset.

### 3.1.1 Synthetic data generation

Inspired by the Self-Instruct framework proposed by Wang et al. (2022), a customized pipeline was developed for generating synthetic multi-turn clinical dialogues in the medical domain. The goal is to create realistic and clinically meaningful dialogues between a patient and a clinician, suitable for training a medical dialogue system. A locally deployed LLaMA (v3.1) 8B-Instruct model serves as the primary generation engine. Through careful prompt engineering, the model is guided to produce believable clinician–patient conversations that are both diverse in content and grounded in clinical knowledge.

Dialogues are produced using specialized instruction prompts explicitly designed to encode the intended clinical scenario and conversational structure. Each scenario follows a high-level structure comprising five core elements: (1) an introduction phase where the clinician greets the patient and sets the consultation context; (2) a background phase outlining the patient persona, including demographics and medical condition; (3) specification of an ICF category that defines the functional focus of the dialogue; (4) assignment of a severity level (1–5) based on the WHO ICF scale to capture the degree of impairment; and (5) an emotional perspective that highlights the patient's subjective feelings, coping strategies, and support systems. Together, these elements provide systematic and clinically interpretable dialogues, while controlled randomization introduces variability.

To operationalize this structure, each synthetic dialogue prompt incorporates key components that are randomized or predefined, ensuring variability and specificity across dialogue instances:

**Patient Persona:** This component provides a concise profile of the patient, including demographic attributes such as age, gender, medical condition, and the specific clinician involved in the consultation. An example prompt could start with:

*"You are a professional and empathetic occupational therapist conducting a clinical consultation with a 55-year-old female patient diagnosed with muscular dystrophy."*

By randomizing patient attributes and clinical conditions, we achieve substantial diversity across patient profiles, encompassing various ages, genders, and medical conditions relevant to the targeted ICF categories. Additionally, it is noteworthy that all diseases described in the patient personas have been validated by medical experts to ensure clinical relevance and accurate alignment with their corresponding ICF categories.

**ICF Functional Target:** Each dialogue is anchored to a specific category from the International Classification of Functioning, Disability, and Health (ICF), chosen from a predefined set of mobility-related codes. These include categories such as d420 (transferring oneself), d445 (hand and arm use), d465 (moving around using equipment), and d470 (using transportation), each representing a distinct aspect of physical functioning. The selected ICF code determines the core functional ability that the clinician will investigate in detail during the consultation.

To ensure domain specificity and clinical relevance, the ICF category and its brief definition are explicitly embedded within the system prompt. For example, a prompt may state:

"The focus of this session is on assessing his or her ability related to the *ICF category D420 – Transferring Oneself*, which refers to *moving from one surface to another, such as sliding along a bench or moving from a bed to a chair, without changing body position.*"

This grounding mechanism ensures that the generated dialogue remains centered on a concrete, meaningful aspect of the patient's daily functioning, providing clarity and consistency across data instances.

**Severity Level:** A fixed severity level (ranging from 1 to 5) is assigned to indicate the degree of impairment in the selected functional domain, following the WHO ICF qualifier scale. For example, level 1 represents complete dysfunction (i.e., the patient is unable to perform the activity at all), while level 5 indicates no problem (i.e., normal functioning without difficulty). This severity information is explicitly included in the prompt to ensure that the generated dialogue aligns with the appropriate level of clinical concern and detail. The prompt incorporates this information through instructions such as:

*"The patient reports a severity level of 5 (1–complete dysfunction to 5–no problem). Your goal is to explore the functional challenges and emotional experiences associated with this activity.".*

This prompt also lays out the expected conversation flow, divided into three sequential stages:

Basic Consultation (3–6 turns): The dialogue begins with a typical clinical opening. The clinician greets the patient and explores general details about the chief complaint or medical history. The patient responds with their initial concerns or symptoms. This stage establishes rapport and gathers background information in a warm, conversational manner.

Functional Follow-up (ICF-focused) (4–8 turns): Next, the clinician directs the conversation toward the specific ICF functioning category selected for the dialogue. Targeted questions are posed to assess the patient's capabilities and limitations in that domain, with a strict focus on functional aspects. For example, if the ICF target is D445 (hand and arm use), the clinician might inquire about the patient's ability to perform daily activities such as buttoning a shirt or lifting objects. The patient then provides information on their functional status, specific difficulties, and any relevant contextual factors. This stage plays a critical role in eliciting detailed insights into

| **Basic Consultation (Severity Level 2, ICF Category D420)** |
|---|
| C: Hi, it's nice to meet you. What's your name? |
| P: My name is Jim. It's nice to meet you too. |
| C: Alright, um so Jim. How old are you? |
| P: 65. |
| C: OK. Now tell me what's going on. |
| P: My legs. I've been having trouble transferring myself these days. |
| C: Sorry to hear that. What do you mean by transferring yourself? |
| P: Well, I've been having trouble getting out of bed recently. |

Table 3.2: Example of Basic Consultation (Student-authored, Researcher-revised)

the patient's abilities, while the prompt constrains the model to remain focused on the designated functional category.

Emotional Feedback & Coping(2–4 turns): In the final stage, the clinician addresses the patient's emotional well-being and coping strategies related to their condition. After discussing functional issues, the clinician adopts a supportive tone, acknowledging any frustrations or anxieties the patient expressed. The clinician might offer encouragement, coping tips, or brief counseling to help the patient emotionally. This closing phase adds realism, as real clinical consultations often conclude with psychosocial support and an empathetic summary. The patient responds with any final concerns or gratitude.

To ensure both clinical plausibility and diversity in the generated dialogues, the prompt incorporates structural guidance and stylistic constraints designed to simulate realistic clinician–patient interactions. The model is instructed to produce warm, empathetic exchanges with clearly contextualized content, maintaining a consistent focus on the specified ICF category and severity level. Patient responses are expected to reflect functional challenges that correspond to the stated severity, while clinician questions are crafted to guide the discussion through relevant daily-life scenarios, probing for specific difficulties, contextual factors, and emotional impacts. The total number of dialogue turns (each turn being a question-answer exchange) is randomized between approximately 9 and 18 turns per conversation. This variability is implemented by not fixing the conversation length in the prompt; instead, the model is directed to sustain the exchange until the topic has been sufficiently explored, resulting in some dialogues being concise while others are more extended. This stochastic turn length mirrors the natural variability of clinical consultations, where certain cases can be addressed quickly, whereas others demand more comprehensive discussion. It also reduces the risk of producing a dataset with uniform, and therefore less realistic, conversational lengths.

With these constraints, the generated dialogues present a diverse range of scenarios while maintaining clinical plausibility. Randomly assigned personas and conditions help ensure that no two conversations are identical, while the ICF-based focus keeps each dialogue anchored to a medically relevant topic. The prompt template was iteratively tested and refined to mitigate mode collapse (i.e., repetitive styles or content) and to address implausible model outputs, for example, by including explicit instructions to avoid providing direct medical advice or diagnoses beyond the scope of a single consultation.

A notable advantage of this self-instructed generation approach lies in its scalability. Once the prompt framework was established, it enabled the synthesis of a large corpus

| **Functional Follow-up (Severity Level 2, ICF Category D420)** |
|---|
| C: Thank you for clarifying. Is it just getting out of bed, or do you also struggle when moving from a chair to another surface, like from a wheelchair or the toilet?<br>P: Actually, both. The other day, I nearly fell when I tried to transfer to my chair.<br>C: That's important to know. Is there any particular position or surface that makes it more challenging for you? For instance, is it harder when you're sitting on a soft chair or trying to get up from a low bed?<br>P: Yes, it's definitely harder when I'm sitting on soft surfaces or trying to get up from bed. I feel like I have to use my arms to push myself up.<br>C: I see. And what time of day does this usually happen? Is it worse in the morning or evening?<br>P: Mornings are the hardest. My muscles feel stiff and weak when I first wake up.<br>C: Do you have any equipment nearby, like grab bars or a walker, to help with transfers?<br>P: I do have a walker, but I often forget to bring it close to the bed, so I end up pushing off with my hands or calling my daughter.<br>C: Has there been a time when you felt unsafe or thought you might fall during these transfers?<br>P: Yes, last week I nearly slipped trying to move from the toilet to my wheelchair.<br>C: Are there any surfaces you find easier, like firmer chairs or raised seats?<br>P: Raised seats definitely help. I added a cushion to my armchair, and that's made it a bit easier.<br>C: Do you feel like your ability to transfer has changed recently—become more difficult over the past few weeks or months?<br>P: Yes, I've noticed it getting worse over the last month or so. I feel like I'm losing strength and confidence. |

Table 3.3: Example of Functional Follow-up (Student-authored, Researcher-revised)

of domain-specific dialogues with minimal human involvement. Human input was limited to preparing curated lists of persona attributes and target ICF codes, along with selective manual verification to monitor output quality. Compared with manual data collection or annotation—both of which would require clinicians to script dialogues or transcribe real consultations—this approach can produce more realistic dialogue data in a fully automated manner. Such efficiency is particularly valuable in low-resource settings for medical NLP, where obtaining extensive, annotated clinical conversation datasets is often challenging.

### 3.1.2 data validation

After generation, the dialogues were subjected to an automated filtering procedure designed to remove suboptimal samples according to predefined quality criteria. Specifically, dialogues were excluded if they contained fewer than eight conversational turns, if any turn had negligible content, if speaker turns failed to alternate between clinician

| **Emotional Feedback (Severity Level 2, ICF Category D420)** |
|---|
| C: Do you live alone? Or is there someone around to help you? |
| P: Yes, I live alone. My daughter works full-time, but she visits me from time to time. |
| C: I see. And how are you feeling right now? |
| P: Honestly, I feel a bit frustrated. I used to be able to move around easily, but now I worry about getting too tired or losing my balance. |

Table 3.4: Example of Emotional Feedback (Student-authored, Researcher-revised)

and patient, or if utterances were repeated verbatim. To ensure alignment with the targeted ICF category, each category was associated with a customized set of domain-specific keywords derived from the official ICF definitions (e.g., for D445 Hand and Arm Use: “push,” “pull,” “reach,” “grasp,” “lift,” “carry,” “arm,” “hand” ). Dialogues were automatically scanned for the presence of at least one of these keywords; samples lacking such coverage were excluded on the grounds that they did not adequately address the intended functional focus. This keyword-based filtering allowed for systematic, reproducible quality control while minimizing manual intervention, with only limited spot-checks performed to verify plausibility.

Following the automated filtering stage, a manual review was carried out by A-proof’s medical experts on a random sample of 17 dialogues (approximately four per ICF category), comprising 256 utterances in total. Both clinician and patient turns were examined utterance by utterance against three criteria:

- Clinical Appropriateness: Determining whether each dialogue accurately reflects realistic clinical scenarios and employs appropriate medical terminology.
- ICF Alignment: Verifying that the dialogues align with the definitions and severity levels specified for the relevant ICF categories.
- Conversational Realism: Assessing the naturalness, coherence, and empathy conveyed throughout the exchanges.

The outcomes are summarized in Table 3.5. High scores were achieved on surface quality. Clinical Appropriateness reached 100.00% (256/256), and Conversational Realism reached 97.66% (250/256), indicating that the generated dialogues were generally consistent with realistic clinical contexts. In contrast, ICF Alignment was notably lower at 75.78% (194/256), with 24.22% (62/256) of utterances judged as misaligned. A key observation from the manual evaluation concerned limited topical consistency across certain dialogue segments. In particular, reviewers noted that “the conversations tend to change to a different category pretty fast; it would be better if they would have more turns on the given category.” This issue was especially evident in the Functional Follow-up section, where dialogues frequently shifted away from the intended ICF domain prematurely.

To address this, the prompt engineering strategy was refined through two targeted interventions. First, the one-shot example embedded in the prompt was updated to present a longer and more consistent exchange within the Functional Follow-up part. This adjustment encouraged the model to maintain focus on a single ICF category over a longer sequence. All revised one-shot examples, which were originally authored

| Criteria | Validation Rate |
|---|---|
| Clinical Appropriateness | 100.00% |
| ICF Alignment | 75.78% |
| Conversational Realism | 97.66% |

Table 3.5: Expert validation on a random sample of 17 dialogues (256 utterances)

| | |
|---|---|
| Total nr. of conversations | 1646 |
| Total nr. of turns | 28187 |
| Average turns/Conversation | 17.12 |
| Average turn length (words) | 24.63 |
| Avg words/Dialogue | 421.75 |

Table 3.6: Overall training data summary

by medical students and later refined by the researcher based on feedback from A-proof's medical experts, were subsequently revalidated in a second round of expert review. Second, explicit domain constraints were incorporated directly into the prompt to further reinforce topical focus. The final prompt version included the directive: *"Keep the dialogue focused entirely on the stated ICF category. Avoid switching to unrelated domains."*

The final statistics of the training dataset are presented below. The training corpus is well balanced across ICF categories and severity levels. Dialogues per category range from 398 to 422, and per-level counts within categories are comparably distributed (typically 72–104), so no single domain or impairment level dominates. Overall, the dataset comprises 1,646 conversations and 28,187 turns; the mean length is 17.12 turns per conversation (range 8–31), and the mean turn length is 24.63 words ($\approx$ 421 words per dialogue). Distributions are similar across ICF categories. These statistics indicate sufficient diversity coupled with stable structure, which is conducive to reliable model training.

### 3.1.3 Test Data

The test dataset was independently developed by a team of medical students specializing in Physiotherapy at the Amsterdam University of Applied Sciences.

To ensure methodological rigor and maintain comparability with the synthetic training data, each manually authored dialogue adhered strictly to a predefined three-part structure: Basic Consultation, Functional Follow-up, and Emotional Feedback and Coping. The dialogues encompassed a diverse range of patient profiles and clinical scenarios, all aligned with the four selected ICF categories (D420, D445, D465, and D470). Annotators were explicitly instructed to vary contextual details, severity levels, and environmental constraints to produce realistic and challenging cases.

All dialogues produced by the medical students underwent cross-validation by medical experts from the A-proof team. Only those meeting the alignment criteria for the target ICF categories were included in the final test dataset. Importantly, the test set was created entirely independently of the training data, thereby ensuring an unbiased and reliable basis for evaluating model performance.

The final statistics of the test dataset are presented below. The final test set comprises 85 independently authored dialogues totaling 1,607 turns. On average, dialogues

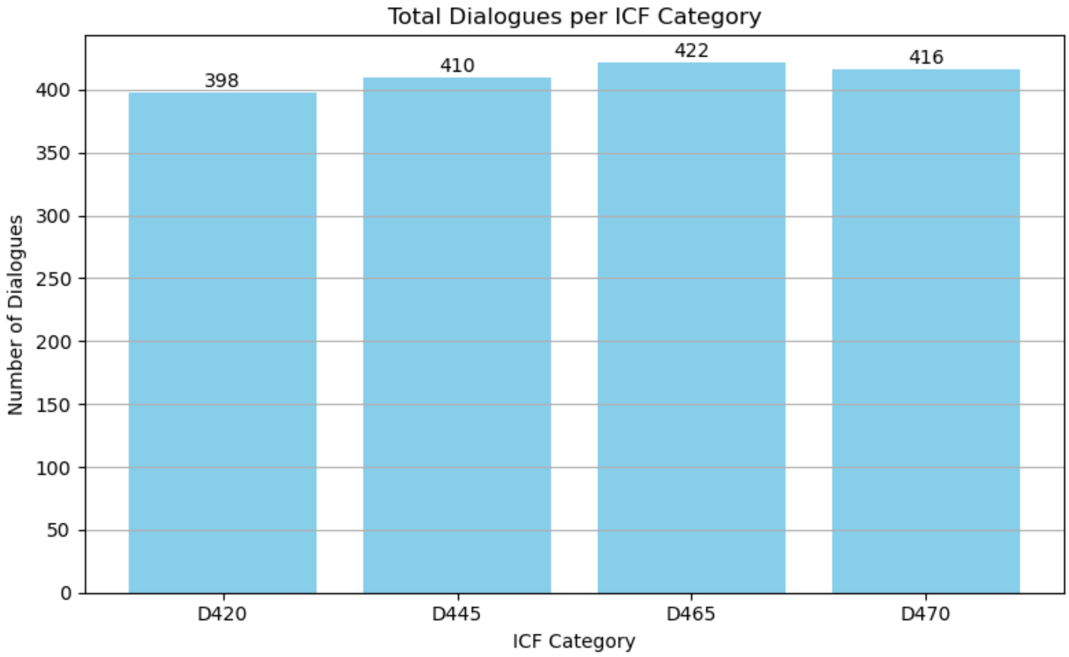


Figure 3.2: Category Distribution of Training Data after validation

| | |
|---|---|
| Total nr. of conversations | 85 |
| Total nr. of turns | 1607 |
| Average turns/Conversation | 18.91 |
| Average turn length (words) | 15.03 |
| Avg words/Dialogue | 284.14 |

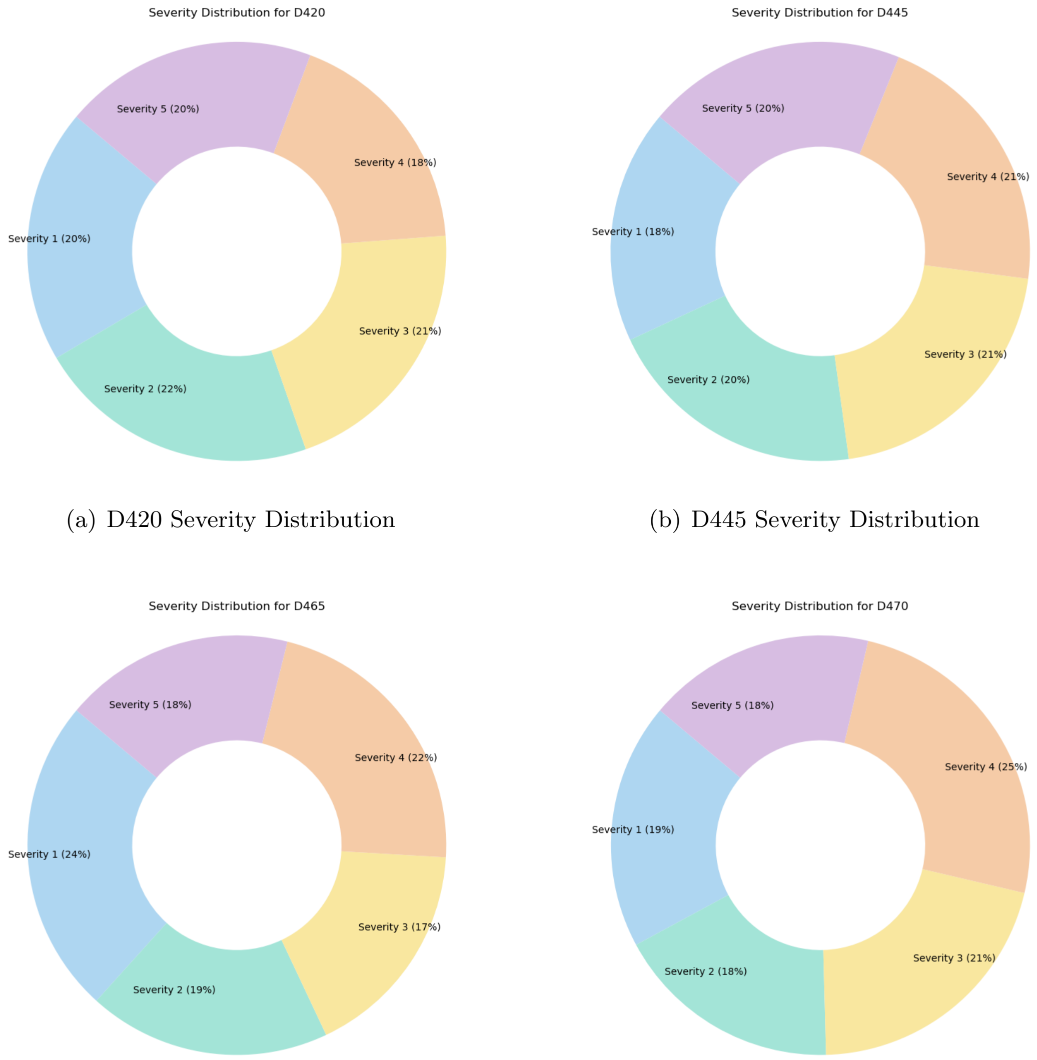
Table 3.7: Overall test data summary

contain 18.91 turns (≈19) and 15.03 words per turn, yielding roughly 284 words per dialogue. Category counts are broadly balanced, supporting comparability across functional domains. Relative to the synthetic training set (17.12 turns per conversation; 24.63 words per turn), the test dialogues have slightly more turns but fewer words per turn, introducing a mild, realistic distributional shift useful for stress-testing generalization.

## 3.2 Model design

### 3.2.1 Benchmark model

The base model employed in this study is Qwen3-8B, a Transformer-based large language model with 8 billion parameters, developed and open-sourced by Alibaba Cloud. As a decoder-only model, Qwen3-8B comprises 36 transformer blocks, each with multi-head self-attention (32 heads) and feed-forward layers. A notable feature of Qwen3-8B is its extended context window, supporting up to 128,000 tokens, which enables it to process and retain information across very long dialogues—an important capability for medical conversation modeling. In addition, Qwen3 introduces a dual-mode architecture with thinking mode for complex, multi-step reasoning and non-thinking mode for



(a) D420 Severity Distribution

(b) D445 Severity Distribution

(c) D465 Severity Distribution

(d) D470 Severity Distribution

Figure 3.3: Severity level distributions in ICF categories of training data

rapid, context-driven responses, allowing dynamic switching between the two based on task requirementsQwen Team (2025).

Qwen3-8B has undergone instruction tuning as part of its multi-stage post-training process Yang et al. (2025); Qwen Team (2025). Specifically, in the later stages, the model was fine-tuned on a mixture of long chain-of-thought (CoT) reasoning data and commonly used instruction-following data—generated by the enhanced reasoning model from the earlier stage—followed by reinforcement learning across more than 20 general-domain tasks. This combination aims at enhancing the model's responsiveness, coherence, and adaptability in interactive contexts, enabling it to generate helpful, safe, and contextually appropriate outputs across a wide range of domains.

These instruction-following capabilities make Qwen3-8B particularly relevant for follow-up question generation in medical dialogues, where the model must interpret the dialogue context, such as a patient's statement or a clinician's remark, and generate a

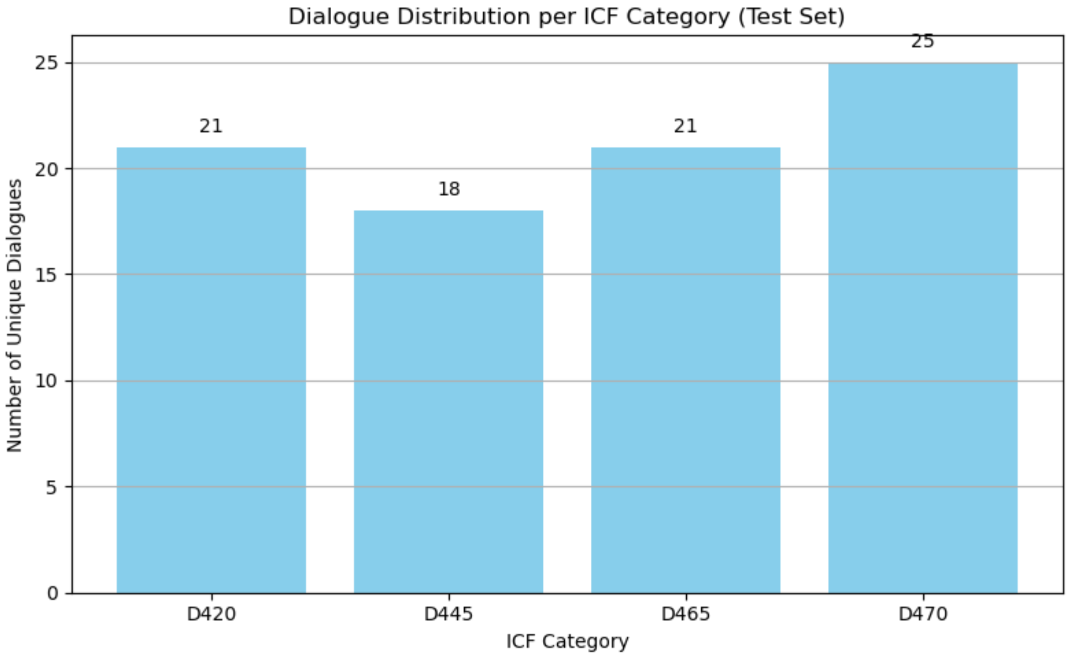


Figure 3.4: Category Distribution of Test Data

coherent, informative, and contextually appropriate follow-up question that aligns with the conversational intent. The open availability of Qwen3-8B's weights, combined with its strong instruction-following foundation, provides an ideal starting point for domain adaptation, enabling the integration of its broad knowledge and conversational skills with fine-tuning tailored to the specific requirements of follow-up question generation in medical settings.

### 3.2.2 Fine-tuning Strategy

To adapt Qwen3-8B to the specialized domain of medical dialogues, a parameter-efficient fine-tuning strategy was employed using the Unsloth library in combination with LoRA (Low-Rank Adaptation) adapters. LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning technique that adds small trainable low-rank matrices to each layer of a frozen pre-trained model, substantially reducing the number of parameters that must be updated for downstream tasksHu et al. (2021). By only training these lightweight matrices, LoRA markedly lowers the memory and computational overhead of fine-tuning large language models while maintaining performance comparable to fully fine-tuned modelZhou et al. (2024).

The implementation was based on the official Qwen3-14B fine-tuning script provided by Unsloth, which was modified to support the 8B version of the model and further customized for the medical dialogue dataset. Instead of updating all 8 billion parameters, LoRA inserts small trainable matrices into specific parts of the model architecture—such as the attention projections and feed-forward networks—while keeping the original weights frozen. As a result, only around 1–10% of the parameters are updated during training, drastically lowering the computational cost. In this work, LoRA adapters are applied to the query, key, value, and output matrices of the self-attention mechanism, as well as to the feed-forward layers within each transformer block.

To further improve memory efficiency, 4-bit quantization was employed, compressing the model parameters into compact 4-bit integers via Unsloth's QLoRA-style backend—a framework that follows the principles of QLoRA Dettmers et al. (2024) by quantizing the base model to 4-bit precision before applying LoRA adapters. This approach reduces memory usage by approximately fourfold compared to full-precision fine-tuningUnsloth Team (2025), enabling training on a single 16GB GPU while preserving competitive performance. Additional GPU-aware optimizations provided by Unsloth, including selective gradient checkpointing, allow the model to handle long input sequences—an essential capability for multi-turn medical dialogues. The model was fine-tuned on a custom synthetic dataset of medical consultations, formatted using ChatML. ChatML provides a structured conversational format using special tokens (e.g., <|im_start|>, <|im_end|>) and clearly separates speaker roles such as "system," "assistant(clinician)," and "user(patient)." This format enables the model to better capture the conversational dynamics of clinician–patient exchanges and maintain speaker consistency.

Each training instance was structured as a multi-turn dialogue with alternating roles, following an instruction-response pattern. Fine-tuning was conducted using Hugging Face's TRL (Transformers Reinforcement Learning) library in supervised fine-tuning (SFT) mode, where the model is trained to predict the next token based on the preceding context. This LoRA + 4-bit setup proved to be highly efficient: according to Unsloth benchmarks, it enables 2× faster training and up to 70% lower GPU memory usage compared to full-model fine-tuningUnsloth Team (2025).

### 3.2.3 Prompt Engineering

#### Zero-shot Prompting

In the zero-shot setting, the model is prompted to perform the task without seeing any explicit example conversation beforehand. Large language models can often be guided by an instructional prompt alone, relying on their pre-trained knowledge and language understanding to generalize to new tasksBrown et al. (2020). In this study, the model receives only a role description and the ongoing dialogue context before being instructed to generate a follow-up question. This setup assesses the model's ability to inherently produce clinically relevant follow-up questions based solely on the instruction and dialogue context, without the aid of additional demonstrations.

For each turn requiring a follow-up question, a prompt is constructed consisting of the following parts in sequence:

**System instruction:** A single system-level message that defines the context and task. In this case, it specifies that *"You are an attentive and empathetic clinician..."* and outlines the relevant ICF categories (e.g., D420 Transferring, D445 Hand and Arm Use, etc.), guiding the model to focus on mobility-related functions. It also instructs the model to ask a concise, clinically appropriate follow-up question based on the patient's last response.

**Conversation history:** All dialogue turns up to the current point, formatted as alternating user (patient) and assistant (clinician) messages. This history provides the model with context about what the patient and clinician have discussed so far. No additional content or examples are added beyond the actual dialogue context.

After these components, the model (as the assistant role) is prompted to produce the next utterance. In this zero-shot prompt, the model must independently decide on a follow-up question, guided solely by the system instruction and the conversation history. No example question–answer pairs are provided; the model's behavior is determined entirely by the given instruction and context.

### Few-shot Prompting

While zero-shot prompting (providing no examples) can work, providing a few examples in the prompt (i.e., few-shot prompting) often helps the model better infer the desired output format and style for more complex tasks. Few-shot prompting is essentially a form of in-context learning, where the model is "primed" with a small number of demonstration examples illustrating the task before it attempts the real task. Brown et al. (2020) first demonstrated the effectiveness of this approach with large language models. Subsequent research has reinforced that including a few well-chosen exemplars in the prompt can significantly improve a model's performance on complex or nuanced tasks (e.g., Min et al. (2022)). In this project, I adopt a few-shot prompting strategy to guide the model in generating questions that are contextually appropriate and empathetic. By including a couple of exemplar clinician–patient conversations in the prompt, the model is primed to follow a similar questioning style and clinical reasoning process during the actual consultation. In this few-shot setup, each prompt is augmented with two example dialogues (a "2-shot" prompt) placed before the real conversation context. The prompt is organized as follows:

**System instruction with examples notice**: The system message still defines the clinician role and task, but it also introduces the forthcoming examples. For instance, the system prompt indicates that *"Below are two example conversations between a clinician (C) and patient (P) for reference, followed by a real consultation"*. This prepares the model to expect and learn from the demonstrations.

**Example 1 – Demonstration dialogue:** An illustrative clinician–patient conversation example (shows the desired questioning style and empathy).

**Example 2 – Demonstration dialogue:** A second example covering a different scenario to reinforce the questioning style and context understanding.

**Transition to actual consultation:** A brief system note such as *"Now here is the real consultation:"* is inserted to clearly separate the example dialogues from the real dialogue. This helps the model reset to the actual patient's context after seeing the examples.

**Conversation history (real consultation)**: The actual dialogue turns between the patient and clinician up to the current point are then provided, as in the zero-shot case. We continue to label the clinician as "C:" and patient as "P:" within these turns for clarity, consistent with the examples. The model thus knows it must now act as the clinician in this specific context.

After setting up the above sequence, the model is prompted to generate the next clinician question for the real consultation. Notably, the two examples were also selected from the medical student–authored test data and subsequently revised by the researcher to reflect typical interactions in this domain (mobility-related ICF scenarios) while demonstrating both functional inquiries and empathy.

### 3.2.4 Experimental Configurations for Model Comparison

To evaluate the effects of model fine-tuning and prompting strategies, this study adopts a single experimental setup with four system configurations. In practice, all combinations are considered between whether the model is fine-tuned on domain-specific data and whether the prompting strategy is zero-shot or few-shot. The four configurations are:

1. Base model + Zero-shot prompting: A base (non-fine-tuned) model with zero-shot prompting (no example dialogue in the prompt).

2. Base model + Few-shot prompting: A base model with few-shot prompting (including a couple of example dialogues in the prompt).

3. Fine-tuned model + Zero-shot prompting: A fine-tuned model (trained on domain-specific clinician–patient data) with zero-shot prompting.

4. Fine-tuned model + Few-shot prompting: A fine-tuned model with few-shot prompting.

This unified design allows for a systematic comparison of model behavior under varying conditions. By evaluating the same consultation scenarios across all four configurations, the impact of each factor—fine-tuning versus prompting style—on the quality of the generated questions can be isolated. For instance, comparing configuration (1) vs. (2) reveals the effect of adding few-shot exemplars to a base model, while (3) vs. (4) shows the effect of exemplars on a fine-tuned model. Likewise, comparing (1) vs. (3) assesses the benefit of fine-tuning in a zero-shot context, and (2) vs. (4) does so in a few-shot context. This factorial approach captures both the main effects of in-context exemplars and model fine-tuning, as well as any interaction between them in terms of the model's question quality. Such a comprehensive evaluation setup is essential, as prior research suggests that fine-tuning and prompting can each significantly influence model performance in complementary waysBrown et al. (2020); Liu et al. (2022). Large language models often exhibit strong zero-shot or few-shot capabilities out-of-the-box, but full fine-tuning on domain-specific data can further boost performance beyond what prompting aloneFatemi and Hu (2023).

## 3.3 Evaluation

### 3.3.1 Automatic metrics

The quality of the generated follow-up questions was assessed using several standard automatic metrics: BLEU, ROUGE-L, and BERTScore. Each metric captures a distinct aspect of similarity between the model-generated question and the reference question(s).

**BLEU (Bilingual Evaluation Understudy):** This metric measures n-gram overlap between the generated and reference questions, producing an aggregate precision score based on matched n-gramsPapineni et al. (2002). Scores range from 0 to 1, with higher values indicating closer matches (a score of 1 represents an identical match to a reference). In this evaluation, sentence-level BLEU was calculated for each generated question and then averaged over the entire test set. Multiple references were allowed for each question, as additional references generally increase the likelihood of n-gram matches.

**ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence):** ROUGE-L was used to assess recall-oriented overlap between the generated and reference questionsLin (2004). ROUGE is a family of metrics originally developed for summary and translation evaluation that compares an automatically generated text to a human-produced reference (or set of references). ROUGE-L specifically measures the length of the Longest Common Subsequence (LCS) between the candidate and reference, which naturally accounts for sequence matching and sentence-level structure. A higher ROUGE-L score indicates that the generated question shares a longer in-sequence overlap with the reference, reflecting stronger content alignment and preservation of the original information flow.

**BERTScore (BERT-based Semantic Score):** BERTScore (BERT-based Semantic Score): BERTScore F1 was used to measure semantic similarity between the model output and reference questions. This neural evaluation metric leverages pretrained Transformer embeddings (e.g., BERT) to compare candidate and reference texts at the token level. Unlike n-gram–based metrics, BERTScore can recognize paraphrasing and meaning equivalence even when the wording differs, making it well-suited for evaluating follow-up questions that may be phrased in multiple ways. The metric computes precision, recall, and F1 by aligning tokens based on the cosine similarity of their contextual embeddings, with the F1 score providing a balanced measure of overall semantic alignment. Higher BERTScore F1 values indicate that the generated question more closely preserves the meaning of the referenceZhang et al. (2020).

For each query in the evaluation set, a human-authored gold follow-up question serves as the primary reference, accompanied by a set of paraphrased variants that preserve the original meaning. These paraphrases account for the fact that multiple valid formulations may exist for the same follow-up question. To generate them, a locally deployed LLaMA-3.1–8B-Instruct model was prompted with the full dialogue context and the corresponding gold question. The prompt explicitly instructed the model to produce ten semantically equivalent variants in a single generation, ensuring that all paraphrases were created simultaneously under consistent contextual conditions. During quality inspection, the first variant in each set was consistently found to be of insufficient quality and was therefore excluded, leaving nine high-quality paraphrase references in addition to the original gold reference for use in automatic metric computation.

The outputs of four model configurations—(1) base model, zero-shot; (2) base model, few-shot; (3) fine-tuned model, zero-shot; and (4) fine-tuned model, few-shot—were evaluated using the selected metrics. For each model-generated follow-up question, BLEU, ROUGE-L, and BERTScore F1 were computed against the reference set as follows:

**Single-reference evaluation:** The model output was first compared to the single gold reference (the original human-written question), providing a strict assessment of how closely the output matched the exact gold phrasing. BLEU, ROUGE-L, and BERTScore were computed for each output–reference pair, and the resulting scores were averaged over the dataset to obtain overall performance for each metric.

**Multi-reference evaluation:** Each output was also evaluated against the set of multiple paraphrase references (the nine remaining paraphrased variants of the gold question). In this setting, a generated question received credit for matching any of the acceptable phrasings of the reference. For BLEU and ROUGE-L, the list of paraphrase references was provided so that overlapping n-grams or subsequences with any reference contributed to the score. This generally yields a higher score when the generated output conveys the correct meaning but uses different wording from the original gold reference, as the additional paraphrase references increase the likelihood of lexical or semantic matches. For BERTScore, the similarity with each reference in the paraphrase set was computed individually, and the highest F1 score among them was taken as the sample's score, reflecting the best semantic alignment the output achieved with the set of valid references. This approach ensures that when the output is highly semantically aligned with at least one valid reference, the score reflects that alignment without being diluted by lower similarity to other variants.

For each model configuration, the average BLEU, ROUGE-L, and BERTScore F1 over all test instances was reported. Together, these automatic metrics provide a quantitative assessment of how closely the generated follow-up questions align with the reference questions, both in wording (BLEU/ROUGE-L) and in meaning (BERTScore). The inclusion of multiple reference variations and a semantic similarity metric ensures that the evaluation fairly credits outputs that are paraphrastically correct, even when they are not word-for-word identical to the gold standard.

### 3.3.2 G-eval (LLM-based Evaluation)

To assess the quality of the generated follow-up questions, a large language model (LLM)-based evaluator was employed within an LLM-as-a-judge frameworkLi et al. (2024). This approach, inspired by G-EvalLiu et al. (2023), uses an instruction-following LLM to read the dialogue context and the proposed question, then rate the question along specific dimensions. A specialized evaluation prompt was designed to encourage step-by-step reasoning, guiding the model through the assessment process before outputting its final verdictChiang and Lee (2023). In particular, the prompt first establishes the model's role as a "medical dialogue evaluator" and explicitly defines the evaluation task and criteria. This ensures that the judgments remain logically grounded in the conversation and aligned with human-like evaluation standards. Two key criteria were applied to each follow-up question:

Relevance (1–5): The extent to which the question is logically and topically related to the prior patient-doctor dialogue. A high relevance score means the question follows naturally from the patient's statements and addresses the context at hand.

Faithfulness (1–5): How well the question stays true to the information in the context, avoiding any hallucinated or unrelated content. A faithful question should not introduce facts or assumptions that cannot be supported by the patient's provided information.

To promote effective reasoning, the prompt breaks down the evaluation process into explicit steps. The LLM is instructed to

- Read the conversation context carefully and identify the main points discussed.
- Examine the follow-up question and consider how it connects to those points.
- Evaluate the question against each criterion (relevance and faithfulness) based on the context.
- And finally assign a score for each dimension (1 = poor, 5 = excellent) and provide a short justification.

By explicitly enumerating these reasoning steps (a form of manual CoT prompting), we aim to guide the model through a thorough analysis before it decides on the scoresChiang and Lee (2023). This strategy mirrors the chain-of-thought style prompting used in G-Eval to improve evaluation qualityLiu et al. (2023). The LLM is instructed to respond only in structured format, with integer scores for each criterion and a concise explanation in the comments field. Importantly, the model is asked to include a brief rationale instead of being restricted to numeric values alone, as recent studies indicate that explanations can enhance the alignment of LLM-based evaluations with human judgmentsChiang and Lee (2023).

An example output would look like: {"relevance": 5, "faithfulness": 5, "comments": "The follow-up question directly builds on the patient's concerns about symptoms and does not introduce any new, unsupported information."}

This G-Eval procedure was implemented using the LLaMA 3.1-8B-Instruct model via the llama.cpp interface as the evaluator. The model was prompted with a low temperature (e.g., 0.2) to minimize randomness and ensure consistent scoring for each question. Through this LLM-based evaluation, all generated follow-up questions in the test dataset were automatically rated on relevance and faithfulness. This method provides a scalable and nuanced assessment of question quality, leveraging the model's ability to interpret context and reason about conversational appropriateness.

### 3.3.3 Human evaluation

A human evaluation of the generated follow-up questions was conducted, focusing on four key criteria: Validity (Clinical Appropriateness), Faithfulness, Relevance, and ICF Alignment. The assessment followed a two-step process carried out by a panel of four medical experts from the A-Proof team.

Validity Check (Clinical Appropriateness): Each candidate question was first reviewed to determine its medical appropriateness and safety given the patient's context. This was a binary decision (valid or invalid), based on the same definition of clinical appropriateness applied in the synthetic data evaluation. Only questions deemed valid progressed to the next stage.

Criteria Ranking: For questions that passed the validity check, the evaluators assessed the remaining three criteria and ranked the candidate questions for each dialogue on each criterion. The criteria were defined as follows:

- Faithfulness: Whether the question is factually accurate and consistent with the information given in the prior dialogue context(this criterion was ultimately excluded).
- Relevance: Whether the question logically follows from the dialogue context, i.e. it addresses or builds upon the patient's last statements or the overall scenario.
- ICF Alignment: How well the question aligns with the specific ICF category relevant to the dialogue (for example, d420 "Transferring oneself" involves moving from one surface to another). A well-aligned question should pertain to the patient's ability/needs in that functional domain.

For each of the three criteria, the outputs for a dialogue were initially intended to be ranked from 1 (best) to 5 (worst) in a comparative manner. In practice, however, the experts adopted a scoring approach, still assigning 1 to the best and 5 to the worst, with ties permitted. This ensured that the scores reflected the relative quality of each question while allowing equally rated outputs when deemed appropriate.

The human evaluation was performed on approximately 40 dialogues (roughly half of the test set, which were written by medical students to simulate realistic patient-provider interactions). Each dialogue was paired with five follow-up question candidates: four produced by different model variations (base model vs. fine-tuned, each in 0-shot or few-shot setting) and one human-written gold-standard question. The order of these five questions was randomized and blinded – the experts were not informed which question came from which model or from the human. They were instructed to evaluate only the content and context appropriateness of the questions, without any bias or knowledge of the source. Notably, evaluators only had access to the visible dialogue context when judging a question and did not see any "hidden" patient replies or additional information beyond the given conversation. This ensured that their assessments of relevance and faithfulness were based strictly on the same context that the models had when generating the questions.

# Chapter 4

# Results

## 4.1 Automatic Metrics

The quality of the generated follow-up questions was evaluated using three complementary automatic metrics: BLEU, ROUGE-L, and BERTScore F1 (see Section 3.3.1 for detailed methodology). Together, these metrics capture both lexical overlap and semantic similarity between model outputs and reference questions. Prior work has shown that BERTScore correlates better with human judgments of generative text quality than BLEU or ROUGEZhang et al. (2020). This combination enables evaluation not only of lexical alignment with references but also of meaning preservation which is an important consideration in open-ended follow-up question generation, where the same intent can be expressed in multiple valid ways.

**Similarity of Gold Answers and Paraphrased Variants:** Before evaluating model outputs, the similarity between the single human gold answers and their paraphrased variants (additional reference answers) was first verified. As shown in Table 4.1, the average BLEU score between gold answers and their variants is about 0.249, and ROUGE-L is 0.278, indicating that roughly 25–28% of n-grams overlap on average. This relatively moderate overlap is expected since the variants were intentionally rephrased. However, the BERTScore F1 for gold vs. variant is much higher (approximately 0.919), confirming that the paraphrases preserve the semantic content of the original answers despite lexical differences. In other words, the variants convey essentially the same information as the gold answers using different wording. This high semantic similarity validates the use of paraphrased references for evaluation. They serve as legitimate alternative answers, reflecting the variety of correct expressions for a given question. Similar observations about leveraging multiple references have been made in prior work on machine translation evaluation: using multiple reference translations or paraphrased references can better capture acceptable variation and improve agreement with human judgmentsKauchak and Barzilay (2006).

**Model vs. Human Gold Performance:** The comparison between model-generated answers and single human gold answers(see Table 4.2) shows that the fine-

| | BLEU (avg) | ROUGE-L (avg) | BERTScore F1 |
|---|---|---|---|
| Gold vs Variants | 0.2494 | 0.2776 | 0.9191 |

Table 4.1: Automatic evaluation results(Human Gold vs. Paraphrased Variants

| Model | BLEU (avg) | ROUGE-L (avg) | BERTScore F1 |
|---|---|---|---|
| Base Model(0 shot) | 0.0334 | 0.1706 | 0.8572 |
| Base Model(few shot) | 0.0404 | 0.1916 | 0.8578 |
| Fine-tuned Model(0 shot) | 0.0553 | 0.1986 | 0.8846 |
| Fine-tuned Model(few shot) | 0.0628 | 0.2124 | 0.8878 |

Table 4.2: Automatic evaluation results (Model Predictions vs. Human Gold)

tuned model produces outputs more closely aligned with the gold references than the base model across all evaluation metrics. In the zero-shot setting (no examples provided in the prompt), the base model achieves a BLEU score of 0.033, whereas the fine-tuned model reaches 0.055. A similar pattern is observed in the few-shot setting (with examples provided), where BLEU increases from 0.040 for the base model to 0.063 for the fine-tuned model. Although these absolute BLEU gains are modest, they represent a relative improvement of approximately 50–60 %, indicating that fine-tuning enables the model to produce wording more similar to that of the gold answers. ROUGE-L scores show a corresponding increase: the base model scores around 0.171 (zero-shot) and 0.192 (few-shot), whereas the fine-tuned model reaches 0.199 and 0.212 respectively. This indicates the fine-tuned model captures more of the key phrases and content fragments found in the references (i.e., longer LCS overlap). The BERTScore F1 shows an even more notable improvement with fine-tuning, which rises from roughly 0.857–0.858 for the base model to 0.885–0.888 for the fine-tuned model. Given that BERTScore measures semantic similarity, this gain indicates that the fine-tuned model produces answers with greater semantic alignment to the gold references, likely as a result of acquiring domain-specific knowledge and answer style during fine-tuning.

The effect of providing few-shot examples versus zero-shot prompting was also examined. As shown in Table 4.2, both the base and fine-tuned models achieve slight performance gains with a few-shot prompt. For the base model, BLEU improves from 0.033 to 0.040 and ROUGE-L from 0.171 to 0.192. For the fine-tuned model, BLEU increases from 0.055 to 0.063 and ROUGE-L from 0.199 to 0.212 when moving to few-shot prompting. These increments suggest that in-context examples can help the models produce answers somewhat closer to the reference wording and content. Nevertheless, the gains from few-shot prompting in this study are relatively small compared with those from fine-tuning. For instance, in the zero-shot setting, the fine-tuned model already outperforms the base model even with few-shot prompting (0.055 vs. 0.040 BLEU). This suggests that while few-shot examples can provide some guidance(perhaps by setting the answer format or reminding the model of relevant facts), the primary improvements in alignment with gold answers mainly come from the model's parameter updates via fine-tuning rather than on-the-fly prompting. A possible reason for the limited gains from prompting is that the model, after fine-tuning, already knows how to approach the questions, so additional examples yield only minor refinements in phrasing. Nonetheless, no strong conclusions can be drawn regarding the overall efficacy of few-shot prompting, as the influence of examples may depend on the specific questions and the relevance of the provided exemplars, factors that require further investigation to fully understand.

It is also worth considering the role of answer length in these results. The base model's answers were significantly longer on average than the gold answers(see Fig-

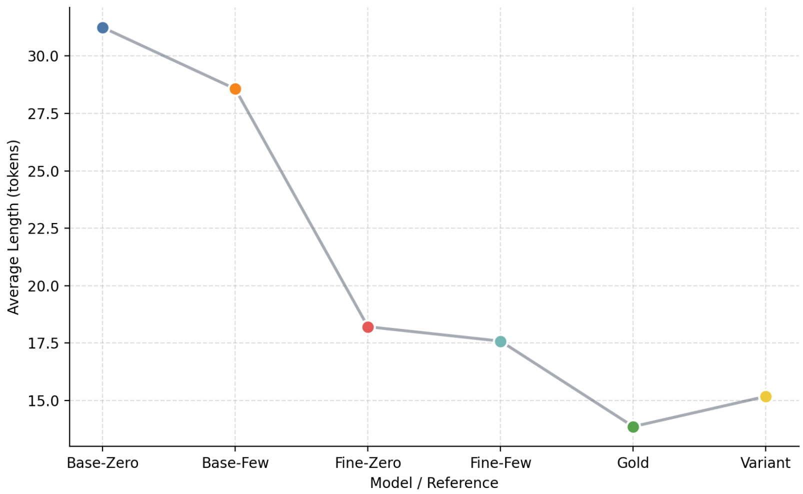


Figure 4.1: Average lengths of all outputs(4 models)&human gold&their variants

ure 4.1 for a detailed comparison), whereas the fine-tuned model's answers were more concise. Specifically, the base model produced answers around 28–31 words on average, roughly double the length of the gold answers ($\tilde{1}$4 words), while the fine-tuned model's answers were much closer in length to the human references. This difference in brevity could partially explain the higher BLEU and ROUGE scores for the fine-tuned model. Since BLEU is a precision-oriented metric that divides the count of overlapping n-grams by the total number of candidate n-gramsPapineni et al. (2002), an excessively long answer with additional unneeded words can dilute the precision (many extra n-grams that do not match the reference) and thus yield a lower BLEU. ROUGE-L, being recall-oriented, is less directly penalized by extra length, but a very long answer might still have a lower fraction of its content aligned with the reference's core information. The fine-tuned model's more concise answers likely stay on-topic and contain fewer irrelevant words, leading to a higher density of n-grams in common with the gold reference. In other words, brevity and focus can improve lexical overlap metrics – a possible reason why the fine-tuned model, which learned to answer more directly, achieves better BLEU/ROUGE scores. However, it is also important to note that this is a correlational observation, the relationship between length and quality is not absolute. A longer answer is not inherently worse (it might contain additional context or detail), but in this case the gold-standard answers are fairly short, and the model that aligned its length closer to the gold tended to score higher on overlap-based metrics. This nuance highlights that automatic metrics should be interpreted carefully: differences in style such as verbosity can influence the scores without strictly indicating correctness.

**Model vs. Paraphrased Reference Performance:** To obtain a more comprehensive evaluation, the model outputs are further compared against paraphrased variants of the gold answers (see Table 4.3). This multi-reference evaluation reveals a pattern consistent with the earlier single-reference results (see Table 4.2). Specifically, the fine-tuned models in both zero-shot and few-shot settings show improvements across all metrics, confirming once again the effectiveness of fine-tuning. However, the impact of few-shot prompt engineering is negligible. In fact, for the fine-tuned model, performance with a few-shot prompt is virtually indistinguishable from that in the

| Model | BLEU (avg) | ROUGE-L (avg) | BERTScore F1 |
|---|---|---|---|
| Base Model(0 shot) | 0.0656 | 0.2469 | 0.8662 |
| Base Model(few shot) | 0.0708 | 0.2578 | 0.8651 |
| Fine-tuned Model(0 shot) | 0.0966 | 0.2690 | 0.8848 |
| Fine-tuned Model(few shot) | 0.0931 | 0.2724 | 0.8865 |

Table 4.3: Automatic evaluation results (Model Predictions vs. Paraphrased Variants)

zero-shot setting, with the latter even yielding a slightly higher BLEU score (0.0966) than the former (0.0931).

When using these additional references, the automatic scores for all models increase across the board compared to the single-reference case. For example, the fine-tuned model (few-shot) achieves a BLEU of 0.093 against the paraphrased variants, higher than the 0.063 BLEU obtained against the single gold reference. A similar rise is seen for the base model (few-shot BLEU 0.071 vs. 0.040 with gold only). The ROUGE-L scores also improve when comparing outputs to the variants (e.g., fine-tuned few-shot ROUGE-L is 0.272, versus 0.212 with only the gold reference). These differences indicate that many model-generated answers, while not lexically matching the original gold answer, do match one of the paraphrased formulations. In other words, the model's wording might differ from the gold standard but still convey the correct idea, aligning better with an alternative phrasing of the answer. The BERTScore F1 for model vs. variants is likewise slightly higher than it was against the original gold. This overall boost in scores with multiple reference variants highlights the benefit of multi-reference evaluation in open-ended tasks: it provides a fairer assessment of the model's output quality by accounting for legitimate variations in phrasing.

The findings here are consistent with the machine translation literature, which has noted that using multiple reference translations can improve evaluation metrics by covering more acceptable wordingsKauchak and Barzilay (2006). In this case, the improved scores in Table 4.3 suggest that the model often produces semantically correct answers that would be undervalued by word-overlap metrics if only one reference were used. By including the paraphrased variants in the evaluation, we capture these correct responses that differ in wording from the original gold answer.

Overall, the automatic evaluation results indicate that the fine-tuned model with few-shot prompting performs best across all metrics, achieving the highest overlap with reference answers. Fine-tuning delivers the most substantial quality gains, while the additional benefit from few-shot prompting is comparatively modest, suggesting that in this task setting, the fine-tuned model generally follows the desired format and content with minimal additional guidance. Incorporating paraphrased variant answers as additional references further reveals that the model frequently produces valid responses that differ in wording from the original gold answers.

While these quantitative metrics are valuable for benchmarking, they also have limitations. The generally low BLEU and ROUGE scores(despite high BERTScore) show that correct answers can score poorly if expressed in an unconventional way. As such, purely quantitative evaluation may not fully capture the model's capabilities or the nuanced differences in answer quality. To address this, the subsequent sections complement the automatic evaluation with large language model self-assessment and human evaluation (see Section 4.2&4.3), examining model outputs from multiple perspectives.

| | avg rel score | avg icf score | top1 rel count | top1 icf count |
|---|---|---|---|---|
| Base Model(0 shot) | 1.45 | 1.96 | 236 | 181 |
| Base Model(few shot) | 1.45 | 2.03 | 235 | 181 |
| Fine-tuned Model(0 shot) | 1.78 | 2.71 | 177 | 117 |
| Fine-tuned Model(few shot) | 1.90 | 2.74 | 170 | 114 |
| Human Gold | 1.97 | 2.82 | 172 | 120 |

Table 4.4: Human Evaluation Results (Relevance & ICF Alignment)

This deeper analysis helps explain why the models behave as the numbers suggest and verifies whether improvements in BLEU, ROUGE, and BERTScore correspond to genuinely better answers in practice.

## 4.2 Human evaluation

### 4.2.1 Combined Results Summary

As mentioned earlier, although the evaluation was originally designed for experts to provide a rank-order comparison of the follow-up questions, in practice each expert assigned a score from 1 to 5 (with 1 indicating the best quality and 5 the lowest quality, hence allowing for ties). Accordingly, the analysis below interprets the results in terms of scores rather than strict ranks.

Table 4.4 presents the aggregated human evaluation results, including average relevance scores, average ICF alignment scores, and counts of top-scored questions across all four experts. On the 1–5 evaluation scale, scores below 2 indicate good quality and scores above 3 indicate poor quality. Overall, the average scores across all systems(including the models and human gold reference) fall between 1.45 and 2.82, suggesting that overall performance ranges from moderate to good.

Within this overall range, the follow-up questions generated by the Base Model achieved the strongest performance, with the lowest (best) average relevance score of 1.45 in both zero-shot and few-shot settings. By contrast, the Fine-tuned Model's questions received higher (worse) average relevance scores of 1.78–1.90, while the human gold-standard questions averaged around 1.97. A similar pattern is seen for ICF alignment: the Base Model scored approximately 1.96–2.03 on average, compared to 2.7 for the Fine-tuned Model and 2.82 for the human questions. In line with these averages, the Base Model's questions were most frequently rated as the best in both relevance and ICF alignment (e.g., 236 top-scores for relevance), outperforming both the Fine-tuned Model and human-authored questions, which achieved only 170–172 top-scores in relevance.

These findings run counter to the initial expectation that fine-tuning would enhance performance in both relevance and ICF alignment, and that human-crafted questions would receive the highest ratings. Rather, the results indicate that experts tended to prefer the base model outputs over the student-authored responses, which were frequently ranked equally low or even lower. Given that the automatic metrics show the fine-tuned model's outputs to be more similar in style to the student responses, its lower scores may reflect this preference bias rather than a genuine quality deficit. Importantly, the score differences between the fine-tuned and base models are rela-

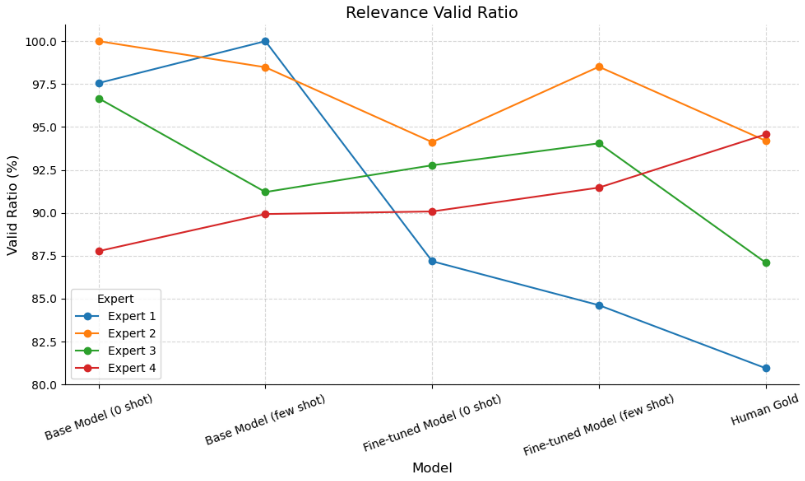


Figure 4.2: Valid Ratio Across Experts and Models

tively small, suggesting a slight but consistent preference rather than a substantive performance gap. To further probe the consistency of these tendencies, each expert's evaluations was examined individually to identify potential biases, revealing notable differences in both the filtering of invalid questions and the scoring of the outputs.

Each expert began by conducting a clinical appropriateness screening, removing any follow-up questions they deemed invalid or inappropriate before scoring the remaining items, as shown in Figure 4.2. For instance, Expert 4 judged approximately 10–12% of the Base Model's questions as invalid (122 out of 139 Base Model 0-shot questions were considered valid) but assessed the Human-written questions as more appropriate, with about 95% deemed valid (122/129). In contrast, Expert 1 found nearly all Base Model questions acceptable (97–100% valid across conditions) yet considered a larger proportion of Human gold questions inappropriate, with only around 81% rated as valid. These discrepancies suggest that certain models, such as the Base Model from Expert 4's perspective or the Human questions from Expert 1's perspective, occasionally produced follow-ups judged clinically inappropriate by individual evaluators. A model that generates a substantial number of invalid questions could appear artificially stronger on other criteria if those invalid cases are simply excluded from scoring. Consequently, these validity differences were taken into account when comparing the models' scores.

Noticeable inter-rater variability in scoring was observed even among questions that passed the validity screening. The four experts did not always agree on which question was best for a given dialogue. For example, Expert 1&2 consistently awarded the Base Model's question the highest relevance score (1) and frequently rated the Human question as the lowest (5) across many dialogues. In contrast, Expert 3 tended to assign the highest relevance score to the Fine-tuned Model's question and produced a more varied ordering for the others. These divergent scoring patterns suggest that each evaluator placed different emphasis on specific aspects of the follow-up questions (such as medical detail, tone, specificity, or alignment with patient needs) when assessing "relevance" or "ICF alignment." Such subjectivity led to variations in scores for the same set of questions. As a result, although the aggregated results indicate clear overall trends, these individual differences highlight the need for cautious interpretation, as well as the importance of inter-annotator agreement (IAA) ratings, which could not be obtained due to time limitations within the scope of this thesis. The unexpected preference for the Base Model was consistent among three of the four experts but not unanimous, underscoring how evaluator bias or perspective can shape outcomes. Taking these factors into account, the following sections provide a more detailed analysis of the scoring results, focusing on the two primary evaluation criteria: relevance and ICF alignment.

### 4.2.2 Relevance Score Performance

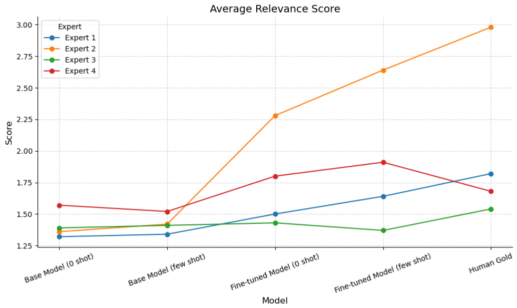


Figure 4.3: Average Relevance Score Given by Experts

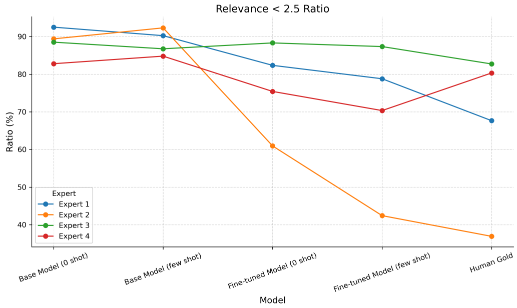


Figure 4.4: Relevance<Median score(2.5) Ratio Across Experts and Models

While the aggregated results show the Base Model outperforming all other systems in relevance, the individual evaluator data reveal marked differences in scoring patterns. Three of the four experts most frequently awarded the best relevance score (1) to the Base Model's questions, while the Human gold questions often received the lowest relevance scores. The Fine-tuned Model's scores and top-1 counts tended to fall between these two, reflecting that fine-tuning appeared to move the model's behavior toward the style of the human-written questions, which paradoxically led to slightly lower perceived relevance according to the experts.

The proportion of questions rated in the top half of relevance (i.e., with a relevance score below the median value of 2.5 on the 1–5 scale) was then examined for each model. Results indicate that including example dialogues in the prompt (few-shot) had minimal influence on relevance. For the Base Model, the proportion of above-median relevance scores remained largely unchanged between the 0-shot and few-shot settings (e.g., Expert 1 rated approximately 82.8% of Base 0-shot questions and 84.8% of Base few-shot questions as highly relevant). The Fine-tuned Model showed a similarly small difference (Expert 1: 75.4% in 0-shot vs. 70.3% in few-shot). This pattern was consistent across evaluators, suggesting that the model's intrinsic capabilities (base vs. fine-tuned) were the dominant factor in relevance performance, rather than prompt format.

Overall, all models produced outputs that were predominantly relevant. Across evaluators, well over half of each model's questions were placed in the top half of relevance scores. Even in the weakest case—the human-authored questions—the majority were judged more relevant than not. For example, Expert 3 considered approximately 82.7% of Human Gold questions to be above the 2.5 threshold. However, substantial variation existed between evaluators. One evaluator (Expert 2) was far more critical, rating only around 36.9% of Human Gold questions in the top relevance half, with an average relevance score of 2.98. In contrast, another evaluator (Expert 4) found roughly 80.3% of the same human questions to be highly relevant. This divergence highlights the considerable variability underlying the aggregate averages, despite the overall trend favoring the Fine-tuned Model's outputs as the most consistently relevant. This result also emphasizes that initial Qwen-3 (base) responses were already very strong in relevance, and fine-tuning altered the style in ways that were not always rewarded by the evaluators.

### 4.2.3 ICF Alignment Score Performance

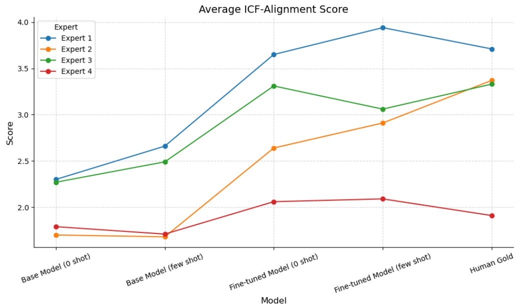


Figure 4.5: Average ICF-alignment Score Given by Experts

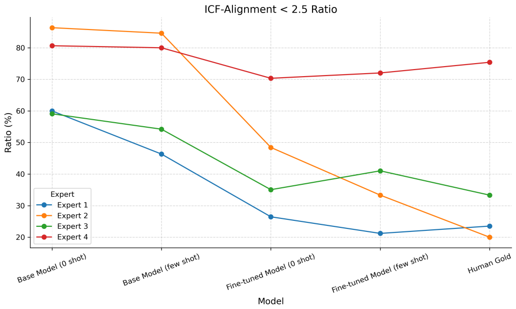


Figure 4.6: ICF-alignment< Median score(2.5) Ratio Across Experts and Models

Across all experts, aligning questions with the target ICF domain was noticeably more difficult than maintaining relevance. The Base Model consistently earned the best (lowest) average alignment scores, generally around 1.7–2.3 depending on the evaluator. In contrast, the Fine-tuned Model and even the Human Gold standard tended to score worse (higher), often exceeding 3.0 under stricter evaluation. No system achieved average scores near the perfect alignment of 1.0, underscoring that even ideal alignment was rarely attained. Comparison with the median score of 2.5 (midpoint of the rating scale) shows that most generated questions fell in the lower half of alignment quality. For example, in the Base Model (0-shot) setting, one expert rated as many as 86% of questions as well-aligned (scores <2.5), while another one rated only 60%, leaving 14–40% above the median (poor alignment). The gap widened under the Fine-tuned Model (0-shot): one evaluator found just 26% of questions well-aligned (<2.5), whereas another found 70%, meaning up to 74% of questions were above the median with the strictest judge. In almost every condition, the majority of outputs had scores $\geq 2.5$, highlighting how rare strong alignment ratings were.

The four experts varied significantly in strictness. Experts 4 and 2 were generally lenient, usually scoring the Base Model around 1.7–1.8 and keeping all other systems below 3.0 on average. In contrast, Expert 1 applied very strict standards: all his scores were the highest (worst alignment), with the Fine-tuned few-shot output averaging 3.94 and even the Human Gold averaging 3.71. Expert 3 fell in between, giving the Base Model moderate scores ( 2.3–2.5) but rating the Fine-tuned and Human outputs above 3.0 on average. These examples illustrate the subjective nature of alignment judgments. One expert might see a follow-up question as clearly on-target while another finds it off-topic (for example, focusing on emotions rather than the specified physical function). This led to large gaps in the percentage of questions deemed well-aligned by different judges. For instance, in the Fine-tuned few-shot scenario, one rater approved about 72% of questions as well-aligned (<2.5) versus only 33% by another. Similarly, for the Base model (few-shot) setting, the most permissive judge gave about 84% of questions scores <2.5, whereas the strictest judge gave only around 46%, a gap of nearly 40 percentage points.

ICF alignment was consistently rated lower than relevance, reflecting both the design of the task and the strictness of the evaluation rubric. First, the training data and questioning policy followed a three-stage structure (small talk → physical function → emotional feedback). Consequently, transitions from physical domains to psychosocial concerns were not only expected but deliberately embedded in the behavior of the Fine-tuned Model and the Human Gold. For example, after sufficient detail had been elicited regarding a physical function, follow-ups such as

*"C: I'm wondering to what extent this incident is affecting you emotionally?"*

*"C: and how do you emotionally feel about the situation?"*

*"C: How are you coping with your current situation?"*

were regarded as clinically appropriate and consistent with the dialogue's final emotional feedback phase. Nevertheless, under the ICF-alignment rubric, evaluators were instructed to judge whether a follow-up directly targeted the designated ICF code, which meant that such emotionally focused transitions were frequently rated as weakly aligned.

Other cases involved clinically relevant physical-function questions that nonetheless received low alignment scores. This typically occurred when the question extended beyond the specific target activity associated with the coded ICF domain and shifted toward another, related functional aspect. For instance, in a rehabilitation scenario, after confirming that the patient's grip was intact, the clinician asked:

*"C: how is your endurance and strength doing?"*

Although clinically meaningful as a core rehabilitation assessment, such a shift was rated as weakly aligned, whereas further elaboration strictly within the scope of grip received higher alignment scores.

A further example comes from the D470 Using transportation scenario. When the patient reported needing her partner's help to board and find a seat (*"P: I can, but no longer alone. I am dependent on my partner. He has to help me board and find a seat."*), The fine-tuned model (few-shot) asked:

*"C: That's a big change. Do you feel like it affects your independence or freedom to go out and do things on your own?"*

while the Human Gold posed:

*"C: I see, what part of the journey do you struggle with most?"*

Both questions were clinically appropriate, but less directly tied to the specific ICF-coded physical impairment than a base model follow-up:

*"C: I see. So you're still able to take the tram, but you need assistance with boarding and finding a seat? Have you noticed any changes in your ability to move around independently, like walking or standing for longer periods?"*

Accordingly, the fine-tuned few-shot model and the Human Gold received alignment scores of 4 and 5, respectively, whereas the base model question received a score of 1 (best). Within this evaluation framework, such cases were often scored as weakly aligned (higher scores) not because the follow-ups were clinically inappropriate, but because they diverged from continued probing of the coded physical function at that point. In contrast, the Base Model, which did not internalize the scripted three-phase structure, tended to remain focused on the physical function domain for longer and thus consistently received better (lower) ICF-alignment scores.

Second, even within the physical phase itself, the granularity of ICF categories made it difficult to formulate questions that consistently targeted a single category (e.g., d420 Transferring oneself) without drifting into neighboring aspects of functioning. In the first-round validation of synthetic data, experts had already noted that "the conversations tend to change to a different category pretty fast." Although the final training data were subsequently refined to minimize such premature shifts, some residual switching across fine-grained categories persisted. Together, these two dynamics (intended late-stage transitions to emotional feedback and occasional early or intra-phase switching among categories) systematically reduced ICF-alignment scores when judged under a strict, category-focused criterion.

Finally, precise alignment requires not only conversational fluency but also specialized knowledge of ICF taxonomy and deliberate category-specific phrasing. While large language models can leverage general medical knowledge to remain contextually relevant, they lack inherent awareness of ICF categories. The Fine-tuned Model, trained on ICF-labeled data, was expected to compensate for this gap. However, the observed gains were modest: many of its outputs were clinically reasonable but not explicitly tied to the target category, and thus received weaker alignment ratings. Even Human-written follow-up questions in the test set frequently strayed beyond the prescribed category, underscoring the inherent difficulty of the task. In practice, both models and humans struggled to consistently formulate questions that adhered tightly to the narrow boundaries of a given ICF category.

### 4.2.4 Base vs. Fine-Tuned vs. Human – Explaining the Surprising Outcome

The human evaluation results revealed an unexpected ranking: the Base Qwen3-8b model (without task-specific fine-tuning) was frequently favored by the experts over both the Fine-tuned Model and the Human-authored gold questions. This outcome contrasts with the automated evaluation, which had indicated clear gains from fine-tuning, and highlights the complex dynamics underlying expert judgment. Several factors may explain this discrepancy.

Fine-tuning shifted the model's behavior toward the style of the synthetic training data and the human-authored gold questions. The synthetic dialogues often contained fixed patterns or a strong tendency toward structured, concise phrasing, which the Fine-tuned Model subsequently inherited. As a result, its outputs tended to be shorter, more targeted, and more structured, sometimes referencing earlier turns or shifting toward emotional concerns in line with the scripted three-stage design of the training set. While such nuanced behaviors are clinically valuable, they were not always rewarded under the evaluation rubric, which prioritized immediate contextual relevance and strict adherence to a single ICF category. In contrast, the Base Model often generated longer and more polite follow-ups that explicitly expanded on the patient's most recent utterance, sometimes sustaining the dialogue for additional turns—albeit with increasing repetition over time. Under the rubric, this longer and more open-ended style was often rewarded with higher scores for both relevance and alignment. In other words, fine-tuning aligned the model with the style of the training data (and the human-authored gold questions), but this alignment did not necessarily match evaluator preferences, leading to the Base Model outperforming the Fine-tuned version under the given criteria.

The unexpected outcome may partly reflect the limited sample size (∼40 dialogues) and the particular scenarios included in the evaluation set. It should also be noted that, not every follow-up question in these dialogues was evaluated: due to time constraints, the experts agreed to exclude the small talk stage and focused only on follow-ups within the physical function and emotional feedback stages. If the dialogues had been more varied or the dataset larger, the advantages of the Fine-tuned Model might have become more evident. Furthermore, individual evaluator preferences (as discussed earlier) exerted a strong influence in a small-scale study. For example, an evaluator who consistently favored longer, broader phrasing would disproportionately elevate the Base Model's scores. With only four evaluators and without crosscheck, such preferences had a pronounced effect on the average scores. A larger evaluation involving more judges would likely smooth out these individual tendencies and yield different outcomes. These observations suggest that the results are sensitive both to the evaluator pool and to the particular dialogues selected for assessment.

The Human-authored gold questions were produced by medical students who did not have access to actual patient cases and received limited training in dialogue structuring and ICF categorization. Moreover, the examples they followed were designed by the project's researcher , who has no medical background. These factors likely introduced systematic limitations in the quality of the gold data. In practice, some gold questions referenced information that was not explicitly visible in the dialogue excerpts presented to evaluators, leading to lower scores when experts judged them as off-topic or redundant. In other cases, the gold questions were a bit complex or oriented toward nuances that evaluators did not prioritize. This mismatch between the intended structure of the dataset and the evaluators' visible context meant that the "gold" questions were not consistently superior. Furthermore, the average ICF-alignment scores confirmed that even human-authored follow-ups frequently drifted from the designated ICF category, reflecting both the difficulty of the task and the inexperience of the annotators. Taken together, these observations suggest that the gold set cannot be assumed to represent an indisputable upper bound. In future work, gold-standard questions crafted or validated by experienced clinicians would provide a more reliable benchmark, ensuring that model outputs are evaluated against a consistently high-quality reference.

A further factor lies in the limited size and quality of the fine-tuning dataset. The Fine-tuned Model was trained on approximately 1,600 dialogues, all of which were synthetic and generated by a smaller 8B-parameter LLaMA model. Such a dataset may not have fully captured the richness and diversity of real clinical interactions. The restricted scale limited the model's exposure to varied patient cases, dialogue structures, and pragmatic nuances, reducing its ability to generalize beyond the stylistic patterns present in the synthetic data. Moreover, because the synthetic data itself carried inherent stylistic biases, the small dataset size amplified these tendencies rather than balancing them out. In effect, the Fine-tuned Model became highly specialized in reproducing the patterns it had seen, but less adaptable in contexts that required broader coverage or more flexible phrasing. This contrasts with the Base Model, whose broad pretraining endowed it with stronger general-purpose dialogue abilities by default.

In sum, these results reinforce that improvements on automated metrics or alignment with reference style do not guarantee better expert-rated performance – such discrepancies between metrics and human judgment are well-documented. Human evaluation remains the gold standard for assessing clinical dialogue quality, and these findings highlight the need to align fine-tuning objectives more closely with expert criteria (and to expand evaluations across more dialogues and judges) to ensure that model enhancements truly translate into human-perceived gains.

## 4.3 G-eval

| **Model** | **Relevance** | **Faithfulness** |
|---|---|---|
| Base Model(0 shot) | 4.967 | 4.937 |
| Base Model(few shot) | 4.943 | 4.906 |
| Fine-tuned Model(0 shot) | 4.770 | 4.903 |
| Fine-tuned Model(few shot) | 4.740 | 4.852 |
| Human Gold | 4.420 | 4.695 |

Table 4.5: Mean G-Eval results for relevance & faithfulness

As a complementary evaluation to the automatic and human judgments, G-Eval was introduced as a new attempt to assess model outputs automatically, using the LLM-as-judge framework described in the methodology. In this setup, a local LLaMA-3.1-8B model served as the evaluator, scoring follow-up questions on Relevance and Faithfulness. Table 4.5 reports the mean G-Eval scores (1–5 scale, with 1 indicating low quality and 5 indicating high quality) across zero-shot and few-shot prompting, as well as for the human-authored gold questions.

All approaches achieved relatively high scores on both criteria, but several noteworthy differences emerged. The base model consistently scored highest, with near-perfect relevance (4.967 in zero-shot, 4.943 in few-shot) and faithfulness (4.937 and 4.906). The fine-tuned model performed slightly worse, particularly on relevance (≈4.74–4.77), while its faithfulness remained comparable (≈4.85–4.90) . By contrast, the human-authored gold questions received the lowest scores (4.420 for relevance, 4.695 for faithfulness).

These outcomes highlight a clear similarity with the human evaluation. Experts had often favored the base model as well, and G-Eval results confirm some of their reasoning: the base model's longer and more polite style adhered more directly to the immediate dialogue context, which the LLaMA-3.1-8B evaluator rewarded with very high marks. Conversely, the fine-tuned model's alignment with the scripted three-stage training data, particularly its tendency to shift questions toward emotional feedback in later turns, was frequently judged "less relevant" under G-Eval. This mirrors the human evaluation, where similar stylistic shifts also received lower scores. Importantly, however, the fine-tuned model's faithfulness remained stable, indicating that it rarely introduced unsupported information and generally stayed true to the patient's statements.

Another notable finding was the relatively low scores assigned to human-authored follow-up questions. Consistent with the human evaluation, this pattern does not imply that models are intrinsically better at follow-up questioning. Rather, it reflects a structural mismatch between authentic clinical practice and the LLM judge's rubric, which prioritizes narrow, local relevance and explicit textual support. When composing the evaluation dialogues, medical students often drew on broader clinical knowledge, engaged in anticipatory reasoning, or introduced background details not explicitly present in the immediate exchange. Although clinically appropriate, such moves were routinely penalized as “off-topic” (lower relevance) or “unsupported” (lower faithfulness).

Consider a case from D445 (Hand and Arm Use): after the patient reported no difficulty with grip and described daily activities as manageable ( *“P: No problem at all, my grip was strong, and the hose moved easily.”*), the clinician drew on the patient’s longitudinal history, citing difficulties documented at a prior visit:

*“C: That’s excellent. What about those stubborn pickle jars you’ve been having trouble with? Are they easier to open now?”*

Despite its clinical plausibility and grounding in longitudinal history, the LLM judge rated it only weakly relevant: while the query stayed within the domain of daily activities, it introduced “pickle jars”—a detail absent from the current exchange focused on gardening and manual dexterity. In other words, questions that extended beyond the immediate textual context were systematically downgraded.

A similar pattern appeared in D470 (Using transportation). During small talk, the patient described a holiday and denied transportation problems (“C: That sounds wonderful. Did you have much trouble with the plane or ferry? P: It all went swimmingly!”). The clinician then asked:

*“C: How did your new ACL hold up during the trip?”*

Here, the introduction of additional background information—clinically reasonable in practice—received the lowest score (1). The judge’s rationale was that the question lacked logical connection to the preceding turns, as no ACL injury had been mentioned in the ongoing conversation. Taken together, these cases illustrate a consistent bias: the judge rewards questions tightly confined to the most recent patient utterance and penalizes clinically meaningful, free-form queries that integrate external knowledge or anticipatory reasoning. This dynamic helps explain why the base model whose follow-ups tended to remain rigidly tied to the last patient response often achieved higher scores than the more contextually expansive, human-authored questions.

Finally, the results show that providing few-shot exemplars had minimal impact on the G-Eval scores. The differences between zero-shot and few-shot performance for both the base and fine-tuned models are negligible (on the order of only 0.02–0.03 points). This indicates that the quality of questions (as measured by relevance and faithfulness) was robust to the prompting strategy. In practical terms, even without explicit examples, both models were able to generate highly relevant and faithful follow-up questions, and additional prompt guidance did not substantially change the evaluated outcome. The consistency of these scores across prompting conditions further underscores the reliability of the models’ question-generation ability under the G-Eval metrics.

# Chapter 5

# Discussion

## 5.1 Limitation

This study investigated the generation of clinically meaningful follow-up questions in health-related dialogues by fine-tuning a large language model (Qwen3-8b) on synthetic data. The findings point to promising directions; however, several key considerations remain, particularly concerning the scope of evaluation, methodological limitations, and the balance between synthetic and real-world data.

ICF-based severity levels were incorporated during the generation of synthetic dialogues to represent the degree of impairment in the selected functional domain. However, this dimension was not explicitly reflected in the evaluation metrics. Consequently, the model's performance across "mild" versus "severe" cases was not assessed, leaving a potentially important factor unexplored. In the absence of severity-aware evaluation, systematic performance differences tied to severity may remain undetected. Future research should therefore adopt severity-specific analyses or employ multi-dimensional metrics to ensure alignment between fine-tuning objectives and evaluation criteria across all key data attributes.

Across the three evaluation approaches adopted in this study(Automatic reference metrics, LLM-based evaluators, and human expert ratings), clear discrepancies emerged, highlighting the distinctive strengths and limitations of each. Standard automatic reference metrics such as BLEU, ROUGE-L, and BERTScore primarily measure surface-level lexical or embedding overlap. Although useful as baseline benchmarks, these measures are insufficient for reliably assessing clinical relevance, pragmatic appropriateness, or the nuanced faithfulness required in health-oriented dialogues. Prior research has likewise demonstrated that optimizing for lexical overlap alone rarely translates into improvements in human-perceived quality, particularly in open-ended or knowledge-intensive tasks Callison-Burch et al. (2006); Liu et al. (2016).

In contrast, human evaluation remains the gold standard, as domain experts are able to assess not only linguistic alignment but also conceptual correctness, contextual appropriateness, and clinical utility. Nevertheless, expert annotation is both resource-intensive and subject to variability in judgment Hämäläinen and Alnajjar (2021). To mitigate scalability constraints, recent work has investigated the use of LLM-based evaluators such as G-Eval Liu et al. (2023), which generate rubric-based assessments at substantially lower cost. While these models show promise in approximating human preferences, they may also inherit biases from their training data and exhibit limitations in domain-specific reasoning Calderon et al. (2025).

The present results of this study confirm these trade-offs: models achieving the highest BLEU or ROUGE did not necessarily receive the strongest human scores for relevance or ICF alignment. Similarly, G-Eval sometimes favored stylistically consistent but clinically less informative questions. This divergence highlights the broader issue that single evaluation modalities are insufficient to capture quality in specialized dialogue systemsReiter and Belz (2009); Fabbri et al. (2021). A hybrid evaluation strategy is therefore essential(eg, one that combines automatic metrics for reproducibility, LLM-based evaluators for scalable approximation of qualitative judgments, and human expert ratings for clinical validation). Such multi-faceted evaluation frameworks have been increasingly advocated in recent NLP surveysGehrmann et al. (2021), reflecting the need for diverse, task-tailored assessment protocols in knowledge-intensive and safety-critical domains. At the same time, aligning the outcomes of different evaluation systems remains particularly difficult, as their underlying criteria often diverge, making it challenging to ensure consistency and interpretability across modalitiesDeriu et al. (2021)

Beyond the above two things, it is also necessary to address the limitations of synthetic data. In the medical NLP context, synthetic and real dialogue data serve complementary purposes. Ontology-guided synthetic dialogues (e.g., based on ICF categories) can systematically cover a broad range of clinical scenarios and rare conditions without raising privacy concerns. As Pezoulas et al. (2024) note, synthetic generation is a promising solution to data scarcity and privacy constraints, enabling large sample sizes and inclusion of underrepresented cases (such as rare diseases) that would be difficult to obtain from real patients.

Despite these advantages, synthetic dialogues often lack the full authenticity of human speech. Machine-generated text tends to be more formulaic and may exhibit limited linguistic diversity compared to real patient language. For instance, empirical evaluations of LLM-generated clinical dialogues have found demographic homogeneity and a relatively uniform style in synthetic patientsHaider et al. (2025). Even high-quality models such as GPT-4 can achieve strong medical accuracy and relevance, yet their outputs frequently remain somewhat mechanical in tone. Furthermore, synthetic datasets struggle to capture pragmatic cues (e.g., sarcasm, hesitations, overlapping speech) and rare linguistic constructions present in real interactions Giuffrè and Shung (2023). By contrast, real patient conversations provide linguistic richness and unexpected contextual details. Authentic transcripts capture colloquialisms, interruptions, emotional inflections, and other subtle features that are difficult to simulate.

In practice, hybrid datasets best exploit this complementarity: structured synthetic data ensures systematic coverage and clinical completeness, while a smaller portion of real dialogues injects natural language variability and genuinely novel contexts into model training. These shortcomings underline that synthetic corpora alone may under-represent the true variability of patient–physician dialogue. Put differently, while synthetic data can enforce comprehensive coverage of clinical concepts, it may underperform on dimensions of conversational nuance and diversity unless carefully augmented with authentic examples.

Taken together, these observations point toward concrete avenues for improvement: incorporating severity-specific evaluation, combining quantitative and human assessments to capture different dimensions of quality and aligning the outcomes of different evaluation systems, and maintaining a balanced mix of synthetic and real data.

# Chapter 6

# Conclusion

This thesis investigated the generation of clinically meaningful follow-up questions in health dialogues, with a focus on adapting large language models through synthetic, ICF-guided training data. Several key findings emerged. First, fine-tuning proved effective: the fine-tuned Qwen3-8B model produced questions that more closely aligned with training data and human-authored gold references. compared to fine-tuning, the benefits of prompt engineering were limited. Although few-shot prompting yielded small gains in automatic metrics (BLEU, ROUGE-L, BERTScore), it had little to no effect under human and LLM-based G-eval assessments. Third, achieving ICF-alignment was substantially more challenging than achieving general relevance. Even with explicit encoding of ICF categories in the training data, the model struggled to learn category boundaries and apply them consistently. Fourth, each evaluation method revealed intrinsic shortcomings: automatic metrics prioritized surface-level similarity, human experts emphasized conversational naturalness but introduced subjectivity, and LLM-based evaluators offered scalability while remaining sensitive to prompt design.

Building on these findings, the contributions of this work can be summarized as follows. It demonstrates the feasibility of ontology-driven synthetic data, based on the ICF framework, as a scalable resource for clinical dialogue modeling. It provides a systematic comparison between prompt-based and fine-tuned approaches, highlighting their complementary strengths and limitations. Finally, it develops a multi-perspective evaluation design, combining automatic, human, and LLM-based assessments, thereby advancing both methodological rigor and critical reflection on evaluation practices in medical NLP.

Despite these contributions, several limitations should be acknowledged. The training data, though diverse in content, were synthetic and may not fully capture the variability of real patient interactions. The absence of large-scale validation sets during data generation limited opportunities for iterative refinement. The evaluation framework, while multi-faceted, exposed inherent shortcomings and discrepancies across methods without offering a unified measure of quality. In addition, the test data used for evaluation were human-authored but produced under constrained conditions, with limited medical expertise and simplified dialogue structures. As a result, the quality of these references was uneven, restricting their reliability as a gold standard.

Future research should therefore focus on enhancing personalization and question quality by refining fine-tuning procedures and incorporating external resources such as retrieval-augmented generation or episodic knowledge graphs(eg., Báez Santamaría et al. (2022) to enable patient-specific reasoning. At the same time, evaluation practices need to evolve toward severity-aware, cross-system, and interpretable frameworks that minimize subjectivity and enable fairer comparisons across models. In addition, emerging new approaches such as knowledge-aware fine-tuning (KaFT)Zhong et al. (2025), which weight training examples by importance or novelty, deserve further exploration as a means of improving robustness and reliability, particularly in rare or high-stakes cases. Advancing along these lines will allow future dialogue systems to move beyond technical adequacy toward becoming clinically meaningful, trustworthy tools for patient monitoring and healthcare support.

# Appendix

In addition to the evaluation methods presented in the main text, I experimented with the RAGAS framework and explored several metrics, including response_relevancy, faithfulness, context_precision, and response_groundedness (see the official documentation for definitions). Note that RAGAS metrics do not share a single input schema: input requirements vary by metric. For example, response_relevancy is computed over a triplet—(user_input, response, retrieved_contexts); faithfulness is computed over (response, retrieved_contexts); and context_precision operates on (retrieved_contexts) together with user_input and optionally reference materials depending on the variant. For example, in the case of response_relevancy, the required inputs would be:

user_input = "When was the first Super Bowl?"

response = "The first Super Bowl was held on Jan 15, 1967."

retrieved_contexts = [ "The First AFL–NFL World Championship Game was played on January 15, 1967, at the Los Angeles Memorial Coliseum."]Ragas (2023)

By contrast, my task concerns follow-up question generation in multi-turn clinical dialogues. There is no retrieval step, and the model output is an interrogative (a question), not a factual answer. To adapt RAGAS, I treated the immediate dialogue window (e.g., the last K turns) as "retrieved_contexts," the patient's most recent utterance as "user_input," and the model's follow-up question as "response." In practice, this mapping proved ill-posed:

Metric–task mismatch: Metrics like faithfulness and response_groundedness assume verifiable propositions; interrogative outputs contain few assertive claims and are often penalized for introducing clinically appropriate hypotheses or longitudinal context not verbatim in the window.

Over-local relevance bias: response_relevancy/context_precision favor narrow lexical overlap with the most recent turns, down-weighting anticipatory reasoning that is common and desirable in clinical practice (e.g., probing previously documented issues or safety risks).

Instability to windowing: Scores varied markedly with the size and composition of the context window (chunking and speaker-turn selection), reducing face validity and comparability with expert judgments.

Given these limitations and the qualitative divergence from expert ratings, I concluded that the current RAGAS setup is not suitable for this project's objective—namely, evaluating the quality of follow-up questions (relevance to the evolving clinical goal, ICF alignment, and conversational realism) rather than the groundedness of answers to retrieved passages. Accordingly, RAGAS results are not included in the main analysis.

# References

M. Abbasian, I. Azimi, A. M. Rahmani, and R. Jain. Conversational health agents: A personalized llm-powered agent framework, 2023.

S. Báez Santamaría, P. Vossen, and T. Baier. Evaluating agent interactions through episodic knowledge graphs. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–28. International Committee on Computational Linguistics, Oct 2022.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL `https://arxiv.org/abs/2005.14165`.

N. Calderon, R. Reichart, and R. Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.782. URL `https://aclanthology.org/2025.acl-long.782/`.

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of Bleu in machine translation research. In D. McCarthy and S. Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, Apr. 2006. Association for Computational Linguistics. URL `https://aclanthology.org/E06-1032/`.

C.-H. Chiang and H.-y. Lee. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*, 2023. URL `https://arxiv.org/abs/2310.05657`.

J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810, Jan. 2021. ISSN 0269-2821. doi: 10.1007/s10462-020-09866-x. URL `https://doi.org/10.1007/s10462-020-09866-x`.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. URL `https://arxiv.org/abs/2305.14314`.

S. Dong, Y. Ling, S. Luo, S. Wang, Y. Feng, Z. Liu, H. Li, A. Goyal, and B. Ferry. Context-aware and user intent-aware follow-up question generation (ca-uia-qg): Mimicking user behavior in multi-turn setting. Amazon Science Publication, 2024. Accessed: 2025-08-17.

X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, 2017.

A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

S. Fatemi and Y. Hu. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*, 2023. URL https://arxiv.org/abs/2312.08725.

J. Gatto, P. Seegmiller, T. Burdick, I. S. Khayal, S. DeLozier, and S. M. Preum. Follow-up question generation for enhanced patient-provider conversations, 2025. URL https://arxiv.org/abs/2503.17509.

S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Anuoluwapo, A. Bosselut, K. R. Chandu, M. Clinciu, D. Das, K. D. Dhole, W. Du, E. Durmus, O. Dušek, C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, R. A. Niyongabo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. S. Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou. The gem benchmark: Natural language generation, its evaluation and metrics, 2021.

M. Giuffrè and D. L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6:186, 2023. doi: 10.1038/s41746-023-00927-3. URL https://www.nature.com/articles/s41746-023-00927-3.

S. Gupta, A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, and A. Sheth. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, 2022. doi: 10.18653/v1/2022.clpsych-1.12. URL https://aclanthology.org/2022.clpsych-1.12/.

S. A. Haider, S. Prabha, C. A. Gomez-Cabello, S. Borna, A. Genovese, M. Trabilsy, B. G. Collaco, N. G. Wood, S. Bagaria, C. Tao, and A. J. Forte. Synthetic patient–physician conversations simulated by large language models: A multidimensional evaluation. *Sensors*, 25(14), 2025. ISSN 1424-8220. doi: 10.3390/s25144305. URL https://www.mdpi.com/1424-8220/25/14/4305.

M. Hämäläinen and K. Alnajjar. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, editors, *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.9. URL `https://aclanthology.org/2021.gem-1.9/`.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Wang. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL `https://arxiv.org/abs/2106.09685`.

D. Kauchak and R. Barzilay. Paraphrasing for automatic evaluation. In R. C. Moore, J. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA, June 2006. Association for Computational Linguistics. URL `https://aclanthology.org/N06-1058/`.

D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, and H. Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024. doi: 10.48550/arXiv.2411.16594. URL `https://arxiv.org/abs/2411.16594`.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL `https://aclanthology.org/D16-1230/`.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022. URL `https://arxiv.org/abs/2205.05638`.

Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. URL `https://arxiv.org/abs/2303.16634`.

Y. Meng, L. Pan, Y. Cao, and M.-Y. Kan. Followupqg: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.17. URL `https://aclanthology.org/2023.ijcnlp-main.17/`.

S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*,

2022. URL https://arxiv.org/abs/2202.12837. Version 2, last revised 2 June 2022.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23:2892–2910, 2024. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2024.07.005. URL https://www.sciencedirect.com/science/article/pii/S2001037024002393.

Qwen Team. Qwen3 official documentation, 2025. URL https://qwenlm.github.io/blog/qwen3/#:~:text=parameters,0.6B%2C%20under%20Apache%202.0%20license. Accessed: 2025-08-10.

E. . Ragas. Answer relevancy, 2023. URL https://docs.ragas.io/en/v0.1.21/concepts/metrics/answer_relevance.html. Accessed: 2025-08-20.

S. Reddy, D. Chen, and C. D. Manning. Coqa: A conversational question answering challenge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2477–2487, 2019.

E. Reiter and A. Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, Dec. 2009. doi: 10.1162/coli.2009.35.4.35405. URL https://aclanthology.org/J09-4008/.

L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun. Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research*, 22(8):e17158, 2020. doi: 10.2196/17158.

Unsloth Team. Qwen3: How to run and fine-tune, 2025. URL https://docs.unsloth.ai/basics/qwen3-how-to-run-and-fine-tune. Accessed: 2025-08-10.

P. Vossen, S. B. Santamaría, and T. Baier. A conversational agent for structured diary construction enabling monitoring of functioning & well-being. In F. Lorig, J. Tucker, A. D. Lindström, F. Dignum, P. K. Murukannaiah, A. Theodorou, and P. Yolum, editors, *Proceedings of the 2024 International Conference on Hybrid Human-Artificial Intelligence (HHAI)*, volume 386 of *Frontiers in Artificial Intelligence and Applications*, pages 315–324. IOS Press, 2024. ISBN 978-1-64368-522-9. doi: 10.3233/FAIA240204. URL https://doi.org/10.3233/FAIA240204.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions.

*arXiv preprint arXiv:2212.10560*, 2022. URL https://arxiv.org/abs/2212.10560. Version 2, last revised 25 May 2023.

Z. Wang, H. Li, D. Huang, and A. M. Rahmani. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations, 2024. URL https://arxiv.org/abs/2409.19487.

C. Winston, C. Winston, C. Winston, and C. Winston. Medical question-generation for pre-consultation with llm in-context learning. In *Proceedings of the NeurIPS 2024 Workshop on Generative AI for Health (GenAI4Health)*, 2024. URL https://openreview.net/forum?id=du26Irf5kE.

World Health Organization. *International Classification of Functioning, Disability and Health (ICF).* World Health Organization, Geneva, Switzerland, 2001. URL https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2020. URL https://arxiv.org/abs/1904.09675.

W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. doi: 10.48550/arXiv.2303.18223. URL https://arxiv.org/abs/2303.18223.

Q. Zhong, L. Ding, X. Cai, J. Liu, B. Du, and D. Tao. KaFT: Knowledge-aware fine-tuning for boosting LLMs' domain-specific question-answering performance. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24085–24100, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1235. URL https://aclanthology.org/2025.findings-acl.1235/.

X. Zhou, J. He, Y. Ke, G. Zhu, V. Gutiérrez-Basulto, and J. Z. Pan. An empirical study on parameter-efficient fine-tuning for multimodal large language models. *arXiv preprint arXiv:2406.05130*, 2024. URL https://arxiv.org/abs/2406.05130.