Careful Generation

An Exploration of Open-Source Large Language Model Support for Advance Care Planning in Paediatric Palliative Care

Master thesis | MSc Technical Medicine

Ellemijn Vernhout

[this page is intentionally left blank]

Careful Generation

An Exploration of Open-Source Large Language Model Support for Advance Care Planning in Paediatric Palliative Care

Ellemijn Vernhout

Student number : 4835530

2 October 2025

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Public Health, Erasmus MC

Dept. of General Paediatrics, Erasmus MC Sophia Children's hospital

26 March 2025 – 17 October 2025

Supervisors:

dr. Carine van Capelle

dr. ing. Megha Khosla

drs. Liselotte Mahieu

Thesis committee members:

dr. ir. Frank Gijsen (chair), TU Delft, Erasmus MC

dr. Carine van Capelle, Erasmus MC

dr. ing. Megha Khosla, TU Delft

prof. dr. Agnes van der Heide, Erasmus MC

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Preface

This thesis project concludes my time as a student in Delft. I have thoroughly enjoyed the student life and have grown in many, many ways. I've felt at home at KT. The uncertainty of what we would become has now been replaced with solid trust that clinical physicians have a meaningful place in healthcare.
This project has been a ride. Migrating from gamification to large language models. Thankfully, under the same supervision of Carine and Liselotte and with the invaluable support and encouragement of Eris.
Carine, thank you for not just supervising me, but mostly mentoring me. You took the time to get to know me and gave words to processes I couldn't see through yet. With those words, you gave me room to grow. Thank you as well for your enthusiasm and positivity.
Liselotte, thank you for being my daily workplace buddy and for the coffee/tea listening ear moments. I've enjoyed sharing this work experience with you. Thank you as well for introducing me to the MBL team that I've really felt a part of.
Megha, thank you for stepping in without hesitation, even without any prior knowledge of this project. Thank you for helping me map out the new project direction and breaking it down into manageable steps.
Thanks, fellow TM graduate students, for the walks, countless coffees and the safe place for sharing our graduation project emotions. And thanks to my housemates, family and friends for even more coffee, motivational speeches, distraction, and the endless *gezelligheid*.
Thank you Michiel, for your love, patience and availability for sparring sessions as well as the unconditional support in moments where I couldn't oversee this project.
Above all, I'm grateful to God for His nearness, guidance and peace throughout this project, but most of all in my life;

"Lord, I don't want to rush on ahead
in my own strength,
when You're right here.
I'm not in a hurry,
when it comes to Your spirit,
when it comes to Your presence,
when it comes to Your voice.
I'm learning to listen,
just to rest in Your nearness.
I'm starting to notice,
You are speaking."
– Not in a Hurry, United Pursuit

Ellemijn Vernhout
Delft, October 2025

Summary

Introduction

Paediatric palliative care (PPC) aims to optimise the quality of life of children with life-limiting or life-threatening conditions by addressing physical, psychosocial, emotional and spiritual needs of children as well as their family members. Advance care planning (ACP) is a central element of PPC, as it helps children and family members formulate values, needs, and goals for future care. However, ACP documentation is time-consuming and burdensome for healthcare professionals (HCPs). Large Language Models (LLMs) may support this process by automatically extracting and structuring ACP outcomes. This study explored the support of open-source LLMs summarising ACP outcomes from Individual Care Plans (ICPs) in a Dutch PPC setting.

Methods

We constructed a pseudonymised dataset of 38 ICPs, with reference ACP summaries structured around three guiding questions: (1) Who are you?, (2) What is important to you?, and (3) What are your goals and wishes for future care and treatment? Two open-source decoder-only LLMs were selected: Llama-3.1-8B-instruct (Llama-3.1) and Fietje-2-instruct (Fietje-2). We evaluated their performance under zero-shot prompting, in-context learning (ICL) with up to eight examples, and QLoRA fine-tuning on 30 training samples. Outputs were assessed with automatic metrics (BLEU, ROUGE-L, BERTScore, MEDCON), complemented by textual analysis and a human reader study.

Results

Automatic metrics indicated comparable overall performance of both models across conditions, with semantic similarity exceeding syntactic similarity.

For in-context learning (ICL), Llama-3.1 showed a slight increase in BLEU and BERT scores at ICL-1, whereas Fietje-2 performed best under zero-shot prompting. However, textual analysis revealed that both models frequently copied content from the example reference text rather than the source ICPs, resulting in limited structural improvements. Hallucinations and interpretation shifts were observed in all generated texts, with Fietje-2 producing fewer structured answers and omitting key details such as age and condition.

For QLoRA fine-tuning, automatic metrics showed only minor improvements, mainly in recall, without meaningful gains in precision or overall quality. Textual analysis indicated that Llama-3.1 produced summaries in the correct structure more consistently than Fietje-2, though it often misplaced information or repeated segments. Fietje-2 generated shorter, less repetitive texts, but with more hallucinations and nonsensical statements. In the human reader evaluation, experts unanimously preferred the reference summaries over both models. When comparing model outputs, most readers favoured Llama-3.1 for completeness, while Fietje-2 was preferred for conciseness. No clear preference was expressed for correctness.

Conclusion

This study demonstrates that while open-source LLMs can extract some relevant ACP outcomes from ICPs, their outputs remain incomplete and unreliable for clinical use. Neither ICL nor QLoRA substantially improved summarisation quality under current data and computational constraints. The limited dataset size, single-reference summaries, and complex prompts likely constrained model performance. Future research should focus on larger, clinically representative datasets derived from transcribed ACP conversations, the inclusion of multiple expert-written references, and systematic prompt optimisation in collaboration with ACP experts, PPC patients and their family members. With careful dataset construction, iterative fine-tuning, and human evaluation, LLMs may in the future contribute to reducing administrative workload and supporting ACP implementation in PPC.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ACP	Advance Care Planning
AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
EMC	Erasmus Medical Centre
EMC-Sophia	Erasmus Medical Centre Sophia Children's Hospital
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HCP	Health Care Professional
ICL	In-Context Learning
ICP	Individual Care Plan
IMPACT	Implementing Paediatric Advance Care Planning Toolkit
LoRA	Low-Rank Adaptation
JSON	JavaScript Object Notation
LLM	Large Language Model
PPC	Paediatric Palliative Care
PPCT	Paediatric Palliative Care Team
QLoRA	Quantised Low-Rank Adaptation
QuickUMLS	Quick Unified Medical Language System
RAM	Random-Access Memory
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
UMLS	Unified Medical Language System

AI disclosure

Generative AI was used in this thesis to improve the structure and clarity of text written by the author, and to support code generation in Python. All analyses and interpretations were determined and written by the author.

1

Introduction

Paediatric patients with life-threatening or life-limiting conditions require specialised care due to the complexity and intensity of their needs. This specialised care is called paediatric palliative care (PPC) [1]. PPC begins from the moment a child's condition is recognised and continues until after their passing, ensuring aftercare for surviving parents and siblings. This care process can last for several months or years [2]. The group of children receiving PPC is diverse. Some have potentially curable conditions for which treatment may fail, such as cancer, while others live with incurable life-threatening conditions that require long periods of intensive treatment, such as cystic fibrosis or severe immune deficiencies. Children with progressive diseases, for which treatment is aimed at symptom relief, such as metabolic disorders, as well as children with conditions that cause severe abnormalities and a high risk of complications, such as cerebral palsy or extreme prematurity, receive PPC [1].

PPC patients may experience a poor quality of life due to high levels of suffering, often caused by physical symptoms such as pain and dyspnoea [3, 4]. PPC aims to prevent and relieve this suffering to achieve the best possible quality of life. The concept of suffering in PPC is not limited to the physical domain, but extends to the psychosocial, emotional, pedagogical and spiritual domains. Therefore, PPC takes a holistic approach by focusing on the child, their family members and their cultural background from a multidisciplinary perspective [1].

The intensity and complexity of the PPC process affect both the child and their family caregivers. Family caregivers often experience emotional, physical and financial burdens that can accumulate over time, increasing the risk of emotional exhaustion, anxiety, depression and burnout [5]. Besides their parental role, parents often have to take on medical caregiving responsibilities, such as managing medical devices and administering medication. They also play a central coordination role in the care process, which involves multiple healthcare professionals (HCPs) and healthcare sites [5, 6]. Moreover, during the PPC process, several difficult decisions need to be made regarding the types of care and treatment to offer, the goals of care and the settings in which care should take place [2]. For these reasons, it is important in the PPC process to pay attention to the child as well as their family members and to support them in their different roles. Aligning this support with the unique needs, values and preferences of both the child and their family is essential for setting meaningful future goals of care that reflect what matters most to all involved [7].

The process of talking about the needs, goals and preferences of PPC patients and their family members is called 'advance care planning' (ACP) [7]. ACP in a PPC setting focuses on the fundamental question of what is truly important to the child and to each family member. ACP conversations provide an open and equal exchange of experiences, knowledge and perspectives between the child, their family and HCPs. These formulated needs, goals, and preferences then form a helping hand in shared decision-making for future care and treatment. As the outcomes of ACP conversations can change over time, it is important to repeat these conversations regularly [7].

The rising number of children with chronic diseases and multimorbidity [8] and the increasing life expectancy of PPC patients due to medical advancements [9], make PPC, and thereby ACP, relevant for a growing number of children. Although ACP has been proven to positively contribute to the care process and HCPs

generally acknowledge its importance, several barriers seem to limit its wider implementation and upscaling [7, 10]. Besides insufficient communication skills and preparedness for ACP, health care teams experience a lack of time and resources as the foremost barriers for implementing ACP conversations in routine care [10]. Yet, due to the nature of the conversation that gradually progresses from exploring the care situation towards discussing personal topics, it is important to take sufficient time for ACP conversations. It is recommended to reserve at least 60 minutes for the conversation itself, which is up to four times longer than a regular clinical consultation [4]. Furthermore, the documentation of patient interactions can take up double the amount of time than the interaction itself [11]. Along with the documentation of other patient interactions, this can account for more than half of an HCP's workday, which is at the expense of direct patient contact and care. Moreover, this workload can increase stress and dissatisfaction among HCPs [11]. The deployment of large language models (LLMs) in the ACP process may help overcome some of these barriers.

An LLM is a large deep learning model that can understand human-like text after being pre-trained on enormous amounts of text data [12]. Pre-trained LLMs can be fine-tuned to perform several language processing tasks, such as text generation, document summarisation and language translation. They are built using the Transformer architecture, a model structure consisting of an encoder, decoder, or both [13]. These model elements use different so-called attention layers to obtain their language understanding. Several studies have explored the potential of LLMs to assist with clinical tasks. Promising clinical tasks could be, for example, automatically generating treatment plans, predicting disease risks and educating patients [14]. Another important potential for LLMs is to alleviate the administrative burden in healthcare [14].

Recently, Van Veen et al. [11] compared the performance of a selection of LLMs after adaptation on a range of clinical summarisation tasks. The selection consisted of both open-source and proprietary models with either encoder-decoder or decoder-only architecture. The authors found that the proprietary, decoder-only models (ChatGPT-3.5 and ChatGPT-4) outperformed all open-source models (FLAN-T5, FLAN-UL2, Llama-2, Vicuna). ChatGPT-4 executed the clinical summarisation tasks the best, showing comparable or even higher quality summaries than medical experts. Whereas the open-source models, the encoder-decoder models (FLAN-T5, FLAN-UL2) produced better summaries of radiology reports than the open-source, decoder-only models (Llama-2, Vicuna), the decoder-only models were better at summarising patient questions and progress notes. Similarly, Chao et al. [15] fine-tuned encoder-decoder and decoder-only LLMs to perform echocardiography report summarisation. Unlike Van Veen et al. [11], they selected open-source models only, with a maximum size of seven billion parameters (Llama-2, MedAlpaca, Zephyr, FLAN-T5). Llama-2 achieved the best overall performance. The final qualitative evaluation concluded that the summaries generated by this model had a comparable quality to those written by cardiologists.

In the palliative care setting, few studies have researched the documentation potential of LLMs. Chen et al. [16] conducted a comparative study between decoder-only LLMs (ChatGPT-3.5, ChatGPT-4, Llama-2). In the study, the models were instructed to summarise a benchmark palliative care doctor-patient conversation, using zero-shot prompting- a method where the model receives no prior examples to guide its response. The results showed that the Llama-2 model was most precise in wording, while ChatGPT-4 showed the most nuanced understanding of text and context. Overall, the authors concluded that the models performed comparably, expressing a slight favour for ChatGPT-4.

To date, no studies have investigated the potential use of LLMs in paediatric palliative advance care planning. Given the promising results in related fields, this study aims to address this gap by taking a first step towards understanding how LLMs may contribute to ACP conversations in PPC. Unlike straightforward clinical documents, such as echocardiogram reports, ACP conversations are sensitive and nuanced, involving personal opinions, emotions, and family dynamics. Capturing these subtleties is therefore critical for any meaningful application of LLMs in this context.

1.1. Study setting

In the Netherlands, each academic medical centre has specialised paediatric palliative care teams (PPCTs) with medical, nursing, pedagogical, psychosocial and spiritual expertise. These teams support children receiving PPC and their families, offer advice to involved HCPs and can play a coordinative role in the care process. At the Erasmus Medical Centre Sophia Children's Hospital (EMC-Sophia), the PPCT also conducts ACP conversations with PPC patients and their families, using the 'Implementing Paediatric Advance Care Planning Toolkit' (IMPACT) [17]. This evidence-based toolkit was developed to facilitate a holistic approach

to ACP in paediatrics and structures the conversation into four stages: introduction, exploration, decision-making and rounding off. First, the introduction step discusses the goal of the conversation: to discover what is important for this particular child concerning future care and treatment. The second step forms the body of the conversation. It is a broad exploration of the child's identity and the family's belief system, responsibilities, fears and worries, expectations, goals and preferences for care and treatment. Then, the third step discusses who will make decisions about this care and treatment and helps formulate and document concrete care goals for the near future. Finally, in the last step, the care expert summarises the ACP conversation and verifies the information with the child and their parents [17].

The outcomes of the ACP conversations are documented in the patient's individual care plan (ICP; see Appendix A) [18]. The ICP contains all wishes and agreements with the child and their family regarding care, treatment, quality of life and quality of dying. The first two sections of the ICP contain general information, such as contact details, family composition, housing situation, and the current care team involved. In the third section, the outcomes of the exploration step of the ACP conversation are documented. Additionally, in the fourth section, the future goals of care and treatment and the decision-making roles are documented. The fifth and last section of the ICP describes the wishes and agreements about end-of-life care. To ensure everyone involved is up to date with these wishes and agreements, the ICP is shared with all caregivers.

The PPCT of EMC-Sophia recognises several barriers to ACP implementation within their institution and is exploring methods to better support the ACP process. A substantial step forward would be tools that assist with the preparation and documentation of ACP conversations. In particular, an artificial intelligence (AI)-driven solution could support HCPs by automatically transcribing conversations and generating structured documentation in the desired format. Before moving towards real-time transcription, this master's thesis provides a first step in this direction by analysing ACP conversation outcomes as documented in ICPs.

1.2. Research objectives

This study addresses the following research question:

How can Large Language Models be applied to summarise Advance Care Planning outcomes from Individual Care Plans?

The primary objective of this study is to evaluate and compare the performance of different LLMs on a dataset of ACP conversations documented in ICPs. To achieve this aim, the study addresses the following sub-questions:

1. Which ACP outcomes should be summarised from the ICPs?
2. Which LLMs are suitable for the summarisation task?
3. Which fine-tuning strategies can optimise the performance of these LLMs?
4. Which evaluation metrics and assessment methods can be applied to measure LLM performance?

Ultimately, this study aims to provide first insights into the potential of LLMs to support ACP documentation in PPC.

2

Methods

This chapter outlines the methodology of the study. First, we introduce the methodological approach in which we discuss the choices and considerations made when determining the research design. Next, we describe the application of this approach in an experiment with our research data.

2.1. Methodological approach

Given the novelty of the research area and the absence of existing ACP datasets, our study has an exploratory character with no established methodology at hand. First, we defined and validated the relevant ACP outcomes through consultation with an expert panel. This formed the basis for constructing the dataset used in this study. For the application and comparison of LLMs, we then adopted common practices seen in clinical LLM research. Specifically, we adopted components of the research approaches of Van Veen et al. [11] and Chao et al. [15].

2.1.1. ACP outcomes selection

We set up an expert meeting with members of EMC-Sophia's PPCT to determine 1) what sections of the ICP report contained relevant ACP outcomes and 2) which information the ACP outcome summary should contain. During the meeting, we first broadly explored the team's view on ACP. We focused on their experience with conducting ACP conversations and discussed the factors that contribute to making these conversations meaningful and valuable. Following this exploration phase, we focused on IMPACT as the methodology for ACP conversations in PPC at the EMC-Sophia. For each conversation step (i.e., introduction, assessment, decision-making, conclusion), we asked the team what they typically discuss and what they consider valuable outcomes. Finally, we introduced the ICP standard form (see Appendix A) and discussed which sections contained ACP conversation outcomes. Based on the results of the expert meeting, we defined a structure for the ACP outcome summaries. The experts then reviewed a sample of ACP outcome summaries written in this structure.

2.1.2. Model selection

To identify which LLMs are most suitable for the ACP summarisation task, we first considered different Transformer architectures and the distinction between open-source and proprietary models. Based on these outcomes, we applied specific selection criteria to choose a small set of candidate models from the Hugging Face model library [19], and compared their performance on inference tasks to make the final selection.

Model architectures

As mentioned in the introduction, LLMs are built on the Transformer architecture. This architecture was first published in 2017, presenting an encoder-decoder model trained for language translation [13]. In the following years, the capacity of Transformer models expanded to additional tasks such as summarisation, question-answering and creative text generation. Alongside these developments, two variants of the architecture emerged, containing a single Transformer component: the encoder-only model and the decoder-only model.

Encoder-decoder models
Encoder-decoder models utilise both components in the Transformers architecture. The encoder processes the entire input text at once and transforms each word into a numerical representation, whilst capturing both word meaning and context information. This representation is then passed to the decoder, which uses it to generate the output text (e.g. a Dutch translation of the English input text). During this generation process, the decoder has full access to the encoded input text and therefore to all word meanings and context information. Due to this access to the encoded text, encoder-decoder models can be applied to tasks that require an understanding of the complete input text, such as language translation, summarisation and question answering [20].

Encoder-only models
Encoder-only models utilise the encoder component of the Transformer architecture. Similar to encoder-decoders, the encoder transforms the input text into a numerical representation that captures meaning and context. Unlike encoder-decoders, they do not include a decoder and can therefore not generate new text. Instead, encoders are effective for tasks that focus on analysing the input text, such as text classification, named entity recognition –identifying and labelling entities such as persons, organisations or locations in a text–, or extractive question answering –locating the exact text that answers the given question [20].

Decoder-only models
Decoder-only models utilise the decoder component of the Transformer architecture. Unlike encoder-decoders, where the encoder constructs a bidirectional representation of the entire input in a single step by attending to both preceding and following words, decoder-only models create their numerical representation within the decoder itself. For each word in the input, the decoder captures meaning and context only from the current and preceding words, resulting in a unidirectional representation. As new words are added –either as part of the input text or as generated output– they are incorporated into the representation, which is then recalculated based on the new information. As a result, decoder-only models are particularly suitable for fluent text generation [20].

The main goal of our study was to summarise ACP outcomes from ICP documents. Since this requires models with generative capacity, encoder-only models were not considered. The combination of contextual understanding with text generation in encoder-decoders, as well as the general text generation capacity of decoders, align with our research goal. Therefore, both subtypes were considered for further evaluation.

Open-source versus proprietary LLMs

Another consideration for model selection is the choice between open-source and proprietary LLMs. Proprietary models such as ChatGPT or Gemini are controlled by the organisations that developed them. While these models often excel in performance, they require a paid license and offer no access to their architecture or pre-training datasets. This limited access prevents fine-tuning for specific tasks, which leaves users dependent on the models' general capabilities and makes it impossible to guarantee explainability and transparency [11, 21]. In contrast, open-source LLMs provide full access to the model architecture and training processes. They can be downloaded freely and run in local environments. This allows users to fine-tune open-source models to specific tasks on training data of choice [21].

In this study, we aimed to fine-tune the selected models on our training data. As these data contain sensitive, personal information, data security and privacy were key requirements. In addition, explainability and transparency were essential to provide insight into what information the models use for their output and whether they can capture the nuances inherent to ACP outcomes. For these reasons, as well as the financial and hardware-related constraints of this study, only open-source models were considered.

Selection criteria

Based on the above considerations, we selected two encoder-decoder and two decoder-only, pre-trained models from the HuggingFace model library [19]. The models had to be open-source and pre-trained on text corpora that included the Dutch language. Additionally, we chose models that, beyond their initial pre-training, had already been fine-tuned on instruction-response tasks. Given the hardware constraints, we selected models with a maximum size of eight billion parameters.

Inference tasks

We based the selection of the two final models on multiple inference tasks. Inference is the process of instructing a model to perform a task on unseen data. During inference, we consecutively prompted each model to summarise seven subsections of seven different ICPs from the dataset, without providing additional context. For this, we selected the texts of the following subsections from the ICP form (see Appendix A):

- 3A Who is the child?
- 3B Perception of illness and life vision
- 3C Current situation
- 3D Future prospects
- 4A Goals for care and treatment
- 4B Medical situation
- 5A End of life

We used the following set-up for the encoder-decoder prompt:

```
Instruction: Summarise the following text in Dutch:
{subsection text}
```

For the decoder-only models, we added an input/output structure to trigger text generation:

```
Instruction: Summarise the following text in Dutch:
Input: {subsection text}
Output:
```

We inferred the models using their default parameters. For the decoder-only models, we limited the length of the generated text by setting the maximum number of newly generated tokens 512 tokens and the repetition penalty to 1.2.

We evaluated each model's seven generated summaries using four criteria, sorted from the minimal required performance level to the ideal performance level. Each criterion was assigned a weight that increased with performance level:

Criterion	**Weight**
Is the generated text in Dutch?	1
Does the generated text consist of one or more sentences?	2
Is the generated text coherent?	3
Is the generated text a summary of the input text?	4

For each model, we determined which criteria were fulfilled by which of the seven generated texts. Thereafter, we assigned a score to each criterion based on the number of texts that fulfilled this criterion. The total score was calculated as follows:

$$\text{criterion score} = \text{weight} \times \frac{\text{number of texts that fulfilled criterion}}{\text{total number of texts}} \tag{2.1}$$

For example, if five out of the seven generated texts from model A were coherent, the score for this criterion would be:

$$3 \times \frac{5}{7} \approx 2.14 \tag{2.2}$$

Finally, for each model, a final score was calculated by summing the four separate criterion scores. The two models with the highest final score were selected for model adaptation.

2.1.3. Model fine-tuning strategy

LLMs acquire their general language understanding during pre-training. During this stage, a model with no prior knowledge is trained using a self-supervised approach, whereby models learn directly from raw data without the need for labelled examples [22]. Because pre-training requires enormous datasets and computational resources, it is typically performed only once by the model developers. As a result, for research projects such as this, it is neither feasible nor necessary to pre-train a new model. Instead, most studies build on pre-trained models by applying fine-tuning, a subsequent training step in which the model

is further optimised for a specific task or domain, using a dedicated dataset [22]. This process is far less resource-intensive than pre-training and is therefore feasible within the scope of this study.

Different strategies exist to fine-tune models on specific datasets. In our study, we used the approaches reported by Van Veen et al. [11] and Chao et al. [15]. In both studies, LLMs were fine-tuned for medical summarisation, using in-context learning (ICL) as well as quantised low-rank adaptation (QLoRA).

ICL

ICL is a model adaptation method that allows LLMs to learn tasks by providing examples of the desired output to the prompt (see Figure 2.1). This method does not alter the model parameters, which makes it a resource-efficient fine-tuning method [23].

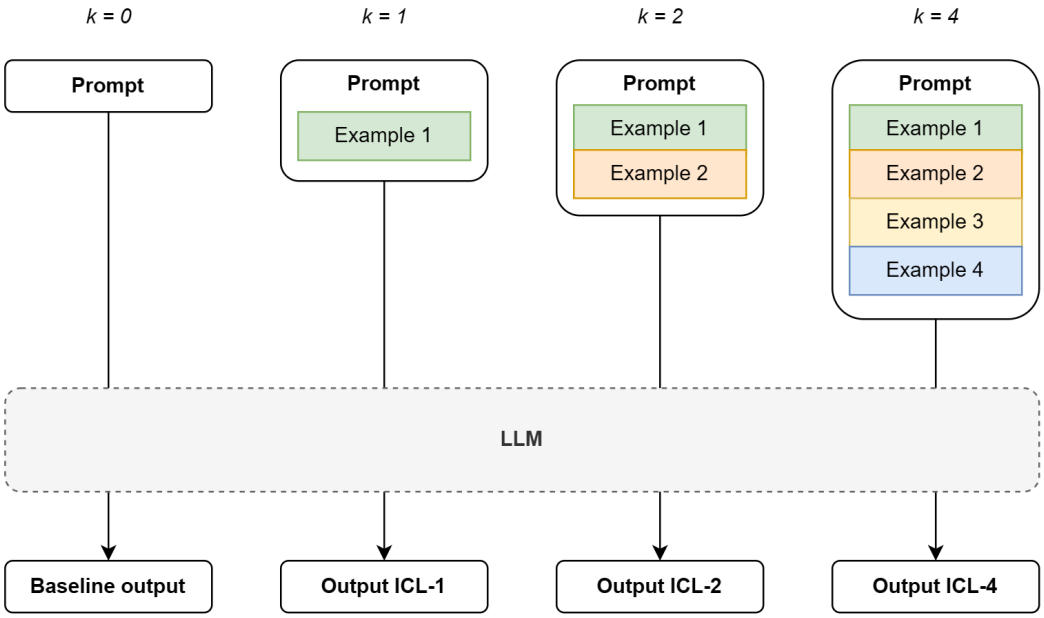


Figure 2.1: Schematic overview of In-Context Learning (ICL) with the addition of up to four examples to the prompt. First, the model is prompted without the addition of examples, resulting in the baseline output text. Then, the model is evaluated with an increasing number of examples k added to the prompt, each example containing a source text and its corresponding reference text.

QLoRA

QLoRa is a variant of Low-Rank Adaptation (LoRA), a supervised fine-tuning method in which a model is trained on a labelled dataset that contains source texts, paired with the desired task outputs, for example, a summary of the source text. LoRA fine-tunes a model by adding additional layers to the model weights, called adapters, while keeping the original model weights frozen (see Figure 2.2). These adapters are matrices with lower dimensions (lower rank) than the original weight matrices. During the fine-tuning, only the smaller adapter matrices are updated, which makes LoRA more efficient and faster than methods that update all model weights. At the end of the fine-tuning process, the adapter weights are merged with the original weights, resulting in a fine-tuned model [24, 25].

QLoRA works similarly, but in addition to adding adapters, it also compresses the original model weights to a lower-precision format, for example, 4-bit instead of 16-bit (see Figure 2.2). This further reduces memory requirements and training time, which makes it possible to fine-tune LLMs on a single graphics processing unit (GPU) [24].

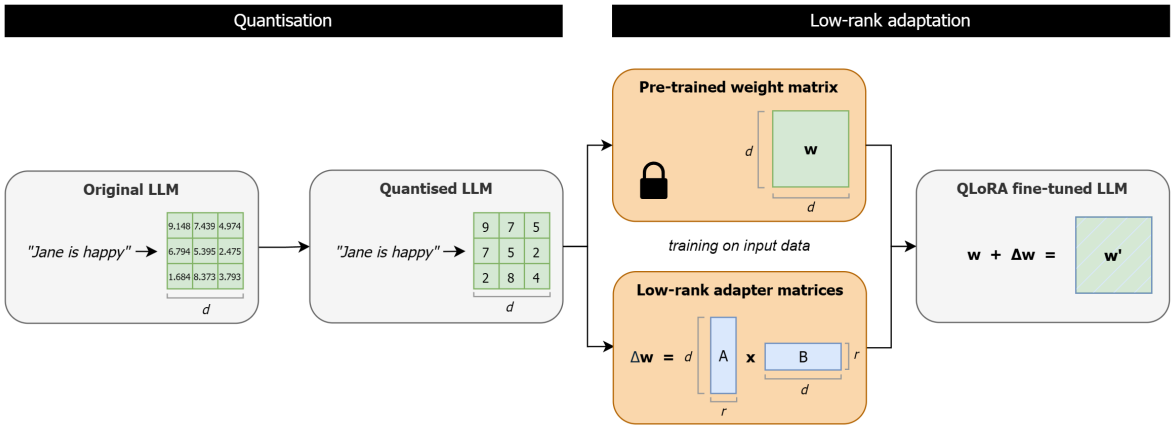


Figure 2.2: Schematic outline of the Quantised Low-Rank Adaptation (QLoRA) process. First, the precision of the weight matrices of dimension d is reduced through quantisation. Then, the adapter matrices of dimension d and rank r are updated during the training, based on the input data from the dataset, while the pre-trained weights remain frozen. After training, the adapter matrices A & B are multiplied, creating a weight matrix Δw with the same dimensions as the original pre-trained weight matrix w. Finally, all updated adapter matrices are merged with the pre-trained weights, resulting in the QLoRA fine-tuned model with weight matrix w'.

2.1.4. Model evaluation strategy

To assess model performance, we applied automatic evaluation metrics, as well as complementary analyses consisting of manual textual analysis and a human reader study.

Automatic evaluation metrics

For the automatic evaluation of texts generated by LLMs, several quantitative metrics exist, each of which evaluates different aspects of the output texts. Where some metrics measure syntactic similarity through the overlap of words, or sequences of words (n-grams), others assess the semantic similarity by comparing whether the meaning of words or sentences align between generated and reference texts. No single metric exists that can provide an all-encompassing evaluation, which makes it important to select a complementary set of metrics. A full review of all available metrics fell outside the scope of this study. Instead, we adopted the approaches of Van Veen et al.[11] and Chao et al.[15], who combined syntactic and semantic metrics by using the BLEU, ROUGE-L and BERTScore. In line with Van Veen et al. [11], we additionally included the MEDCON score that evaluates the medical similarity between the reference and generated texts.

BLEU score

The Bilingual Evaluation Understudy (BLEU) score is a syntactic similarity metric that evaluates the number of matching n-grams between the generated text and the reference text [26]. BLEU computes the modified precision of these matching n-grams, which means that a word in the generated text is only counted as a match for the number of times it occurs in the reference text.

The modified precision is computed for n-grams of up to four words in length, after which the BLEU score is calculated as the geometric mean of the four modified precisions. Finally, a brevity penalty is applied if the generated text is shorter than the reference text. The BLEU score is defined as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log p_n\right) \tag{2.3}$$

with:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{2.4}$$

where:

BLEU	=	BLEU score (0-1)
BP	=	brevity penalty
N	=	maximum n-gram order
p_n	=	modified precision for n-grams of order n
r	=	number of words in reference text
c	=	number of words in generated text

ROUGE-L score

Similar to the BLEU score, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores are syntactic similarity metrics that compare the generated text against its reference by evaluating the number of matching n-grams between the texts [27]. Unlike the BLEU score, the ROUGE scores also consider recall; how much of the information in the reference text is captured in the generated text. The ROUGE scores contain different variants: ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-Lsum, measuring unigrams (single words), bigrams (two consecutive words) and the longest common subsequence (LCS), respectively. Although all ROUGE variants were considered, we decided only to include the ROUGE-L to avoid redundancy and to maintain interpretability. For each variant, the precision, recall and harmonic mean of the precision and recall (F1 score) is calculated. The recall (R), precision (P) and F1 scores for the ROUGE-L score are defined below. The definitions of the other scores are shown in appendix B.

$$\mathsf{R}_{\text{ROUGE-L}} = \frac{\textit{longest common subsequence}}{\textit{total number of unigrams in the complete reference text}} \tag{2.5}$$

$$\mathsf{P}_{\text{ROUGE-L}} = \frac{\textit{longest common subsequence}}{\textit{total number of unigrams in the complete generated text}} \tag{2.6}$$

where the LCS is defined as the longest sequence of words that appears in both texts, in the same order, but not necessarily contiguously.

$$\mathsf{F1}_{\text{ROUGE-L}} = 2 * \frac{\mathsf{R}_{\text{ROUGE-L}} \cdot \mathsf{P}_{\text{ROUGE-L}}}{\mathsf{R}_{\text{ROUGE-L}} + \mathsf{P}_{\text{ROUGE-L}}} \tag{2.7}$$

BERTScore

The BERTScore evaluates the semantic similarity between the generated and reference texts by converting each word into a numerical representation (embedding) using a pre-trained encoder-only model [28]. The numerical representations of words with similar meanings, such as synonyms, are closely related. Determining the relationship between the different embeddings, therefore, allows BERTScore to evaluate overlap in meanings rather than in exact words. The relationship between the embeddings of the generated text and the reference text is calculated as the cosine similarity, where higher values indicate greater semantic overlap. To calculate the BERTScore, all embeddings in the reference text are paired with their most similar embeddings in the generated text, and vice versa. By taking the sum of the similarity values of these pairs, the recall, precision and F1 score can be computed:

$$\mathsf{R}_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{2.8}$$

$$\mathsf{P}_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \tag{2.9}$$

$$\mathsf{F1}_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{2.10}$$

where:

$\lvert x \rvert$	=	number of tokens in the reference text
$\lvert \hat{x} \rvert$	=	number of tokens in the generated text
x_i	=	i-th embedding in the reference text
$\hat{x}_j$	=	j-th embedding in the generated text
$x_i^\top \hat{x}_j$	=	cosine similarity between embeddings x_i and $\hat{x}_j$

MEDCON score
The MEDCON score is a medical concept-based metric [29]. It employs the Quick Unified Medical Language System (QuickUMLS) tool to identify predefined medical concepts from the Unified Medical Language System (UMLS) database in the generated and reference texts. The MEDCON score is then computed by evaluating the overlap of the matched concepts between both texts. The recall, precision and F1 are calculated as follows:

$$\mathrm{R}_{\mathrm{MEDCON}} = \frac{\textit{number of overlapping medical concepts between generated and reference texts}}{\textit{number of medical concept matches in the reference text}} \tag{2.11}$$

$$\mathrm{P}_{\mathrm{MEDCON}} = \frac{\textit{number of overlapping medical concept matches between generated and reference texts}}{\textit{number of medical concept matches in the generated text}} \tag{2.12}$$

$$\mathrm{F1}_{\mathrm{MEDCON}} = \frac{2 \cdot P_{\mathrm{MEDCON}} \cdot R_{\mathrm{MEDCON}}}{P_{\mathrm{MEDCON}} + R_{\mathrm{MEDCON}}} \tag{2.13}$$

Considerations
Syntactic similarity metrics indicate the degree of overlap in word choice, sentence structure, and grammar between the reference and generated texts. They are valuable for verifying whether key elements, such as the child's age or condition, are reproduced in the generated text. However, the main limitation of syntactic similarity metrics is that they penalise synonym use or differences in word orders, even when the meaning remains unchanged. For example, the reference text *"It's important for both parents to be involved in all care decisions"* and the generated text *"The parents want to be included in every decision regarding care"* have the same meaning, but would yield low n-gram scores. Therefore, the syntactic similarity metrics need to be interpreted with care in ACP context.
The BERTScore, on the other hand, assesses similarity in meaning. In the context of ACP conversations, where many different formulations can correctly express the same idea, the BERTScore can provide a more reliable indication of the generated text quality. MEDCON complements the other metrics by specifically measuring the overlap of standardised medical concepts, which is directly relevant to the clinical usefulness of the generated texts.

In this study, we therefore decided to consider the BERTScore as most indicative of overall text output quality, while the BLEU and ROUGE-L scores provide supporting insight into the syntactic accuracy of the texts, and the MEDCON score adds a medical domain-specific perspective.

Complementary analyses
Besides the automatic evaluation metrics, we performed complementary analyses to capture text quality aspects that cannot be fully reflected in similarity scores. First, a manual textual analysis was conducted on a sample of generated texts that focused on text structure, the presence of model hallucinations, segments of text repetition, and segments where the model interpreted text differently than in the reference text. This qualitative analysis allowed us to verify whether the indications suggested by the automatic metrics were indeed reflected in the generated texts.
Additionally, for the texts generated by the QLoRA fine-tuned models, we conducted a human reader study, in which HCPs compared the generated texts of the selected models with each other and with the reference text. To keep the evaluation feasible for participating HCPs within the available time, the reader study was limited to QLoRA outputs. Following the approach by Van Veen et al. [11], the HCPs assessed three aspects of the texts:

- Completeness: which text contains more relevant information?
- Correctness: which text contains less false information?
- Conciseness: which text contains less non-important information?

2.2. Experimental evaluation

To evaluate and compare LLM performance on the ACP summarisation task, we carried out an experimental evaluation. The following subsections describe the dataset construction, model fine-tuning, and the evaluation process.

2.2.1. Dataset construction

For model fine-tuning, a labelled dataset is required that contains both the source texts and the corresponding reference texts. In this study, the source texts were the ACP-relevant sections of the ICPs, and the reference texts were the ACP summaries. Together, these formed the dataset that was used for model fine-tuning and evaluation.

Data acquisition

The ICP reports used in this study were the documented outcomes from ACP conversations held with PPC patients, their parents and an HCP of the EMC-Sophia's PPCT, between 2021 and 2024. As the ICPs contained patients' medical and personal information, they were treated as sensitive, personal data, falling under the General Data Protection Regulation (GDPR). We obtained approval for the use of the data from the Medical Ethics Committee of the EMC.

Data pseudonymisation

We performed pseudonymisation of the research data by removing all directly identifying personal data and by minimising indirectly identifying personal data, i.e. by removing contact details and dates from the ICPs and replacing dates of birth with ages. In addition, we replaced names with their corresponding role or function (e.g. child, parent, paediatrician, school, hospital). To identify patients in the pseudonymised research dataset when needed, we assigned a code to each ICP report and stored the corresponding key file in a separate location from the research dataset.

Data extraction

The source texts were derived from ICP reports, which were available as Microsoft Word documents. To automate the process, we developed a Python function that accepted an ICP document as input and extracted the sections and subsections relevant to ACP. The extracted text was then stored in the dataset, ensuring that only ACP-related information, as defined in the expert meeting, was included.

Data labelling

Based on the results of the aforementioned expert meeting, one researcher wrote the reference ACP summaries. To evaluate the quality of the reference texts, three ACP experts reviewed a sample of five ACP summaries. The experts were asked to provide feedback on the texts' completeness, correctness and conciseness. From the feedback, points for improvement were identified, which were then implemented across the complete dataset.

2.2.2. Hardware set-up

We ran the experiments in a virtual environment, on a secure, remote server. Within the virtual environment, we performed model fine-tuning and evaluation on a single Tesla 4 GPU with 8 cores and 56GB random-access memory (RAM). All code was written in Python, using Visual Studio Code.

2.2.3. Model fine-tuning and evaluation

We defined the general prompt structure, selected hyperparameter settings, and applied ICL and QLoRA fine-tuning approaches. Model outputs were evaluated automatically, using BLEU, ROUGE-L, BERTScore and MEDCON, and manually through textual analysis and a human reader study. For BERTScore, given that the ICPs were written in Dutch, we loaded the *bert-base-dutch-cased* encoder-only model [30]. For MEDCON, we implemented both the Dutch and English UMLS concept databases, as the English database is larger, and medical concepts are often similar across both languages. In addition, we selected the same UMLS semantic groups as Van Veen et al. [11]: *Anatomy, Chemicals & Drugs, Devices, Disorders, Genes & Molecular Sequences* and *Phenomena and Physiology*.

Prompt structure

We based the prompt on the outcomes of the expert meeting and the model selection process. Following common practice for prompt engineering, we determined the following structure:

```
Instruction: {the assignment the model has to fulfil}
Examples:    {pair of example ICP source text and reference text} (ICL only)
Context:     {the ICP source text}
Answer:
```

Hyperparameter settings

Each LLM consists of several changeable hyperparameters. In this study, we explored the influence of different values for the temperature hyperparameter and the addition of expertise to the prompt.
Temperature controls the degree of randomness in texts generated by LLMs. At each step of text generation, the model assigns probabilities to all possible next tokens by forming a probability distribution. The temperature value adjusts this distribution. A lower temperature increases the differences between the token probabilities, which makes the model favour the most likely option and results in more deterministic and consistent output. A higher temperature decreases the probability differences, allowing less probable tokens to be selected by the model, which makes the generated text more diverse and less predictable [31].

We evaluated model performance for temperature values of 0.1, 0.5, 0.7 and 0.9. For each temperature value, we ran inference with ten different prompts, based on ten randomly picked ICPs from the dataset. We evaluated the similarity of the generated texts with the reference texts by calculating the mean BLEU, ROUGE-L, BERTScore and MEDCON score. We ultimately selected the temperature value with the highest scores, where the BERTScore was decisive.

Expertise:
After determining the optimal temperature value, we evaluated the additional value of adding an expertise to the prompt to assist the model in determining the correct context:

```
Expertise: You are an expert medical professional.
```

We applied the same inference and evaluation method as during the temperature value exploration.

ICL

We conducted ICL runs with the addition of $k = 0, 1, 2, 4$, and 8 examples to the prompt. Each example consisted of an ICP source text and its reference text. The condition $k = 0$ that corresponds to zero-shot prompting was considered baseline model performance. For each run, ten random samples from the dataset were selected. For each sample, a prompt was constructed with the addition of k examples, after which the model generation started. Besides adjusting the temperature setting, we set the maximum output length to 1,024 tokens, while keeping all other model parameters at their default values. After text generation, the reference and generated texts were stored in a separate JavaScript Object Notation (JSON) file, together with the corresponding evaluation metrics results.

In the textual analysis, the generated texts were systematically read and compared to the corresponding source texts by one researcher. For each sample, it was assessed whether the text followed the predefined ACP summary structure, whether the child's age and condition were mentioned, and whether information was extracted correctly but presented in the wrong context. In addition, the texts were examined for model hallucinations, defined here as statements not present in or contradicting the source text, as well as for segments of repetitive wording. We also recorded the average word count per condition to capture differences in text length between models and prompting strategies. All observations were documented per model and per ICL-k condition. We selected a sample of texts for the analysis by first determining the k-value with the highest automatic metric scores and then choosing the two samples with the highest BERTScore from the ten ICL-k runs. If the models performed best at different k-values, we included samples of each k-value from both models. The findings were summarised per model and ICL-k condition and illustrated with anonymised citations from the generated texts.

QLoRA

We quantised the models from 16-bit to 4-bit using the $bitsandbytes$ function. We added the LoRA adapters to the model's attention layers, using the get_peft_model function from the PEFT library. We set the LoRA configuration hyperparameters to default values ($adapter\_rank = 8$, $lora\_\alpha = 16$, $lora\_dropout = 0.05$, $bias = None$).

Before training, we split the dataset into a training, a validation and a test set ($80\%/10\%/10\%$ respectively). The model training was performed with the $Trainer$ function. Due to the limited dataset size and limited GPU memory, we set the training arguments to a small training and evaluation batch size of one sample, gradient checkpointing, a high learning rate ($2e^{-4}$), an evaluation strategy after every epoch and an early stopping patience of three epochs with constant training and validation losses. During the training process, the training and validation losses were computed per training round (epoch). In one epoch, all training set samples were presented to the model. The training loss indicates how often and how strongly the model's prediction of the next word deviated from the correct next word in the reference text, averaged across all predictions in that epoch. The validation loss is calculated in the same way, but on samples from the separate validation set that the model has not seen during training. Therefore, the validation loss indicates how well the model performed on unseen data.
Together, the training and validation losses provide insight into the model's learning process. Decreasing losses show that the model is improving. A growing gap between the training and validation loss, on the other hand, indicates that the model is adapting too specifically to the training data, leading to worse performance on unseen data (overfitting).

We evaluated the fine-tuned models by inferring them with the samples in the test set. The reference and generated texts, as well as the results of the evaluation metrics, were stored in a separate JSON file. The QLoRA results of both models were then analysed and compared. Additionally, we compared the results of zero-shot prompting, as well as the best-performing ICL setting, defined as the k-value that yielded the highest metric scores.

We performed the same textual analysis as described in the above ICL section (see 2.2.3). We evaluated all generated texts in the test set. For the human reader study, each HCP was provided the ICP source text as context, together with blinded versions of the generated and reference texts. They were then asked to assess which texts complied better with the predefined criteria. Figure 2.3 shows the outline of the assessment form.

Assessment text A versus text B

Which text...	**Text A**	**No difference**	**Text B**
... contains more relevant information?	O	O	O
... contains less false information?	O	O	O
... contains less non-important information?	O	O	O

Figure 2.3: Outline of the assessment form of the human reader study.

3

Results

This chapter outlines the results section of the study. First, we discuss the results of the ACP outcomes selection and model selection process. Next, the results of the experimental evaluation are discussed per fine-tuning strategy.

3.1. Methodological approach

3.1.1. ACP outcomes selection

In the expert meeting with members of the EMC-Sophia's PPCT, we concluded that the following sections of the ICP form (see Appendix A) contained relevant ACP outcomes:

- Section 3: Values, goals, preferences (*"Waarden, doelen, voorkeuren"*)
- Section 4: Care and treatment (*"Zorg en behandeling"*)
- Section 5: End of life (*"Levenseinde"*)

Additionally, we concluded that to obtain a relevant ACP summary, three guiding questions should be answered:

1. Who are you?
   Introduces the child, child's age, condition, current points of attention, character, care dependence, organisation of care, family.
2. What is important to you?
   Discusses what is important in life for both child and parents.
3. What are your goals and wishes for future care and treatment?
   Mentions any goals, wishes, agreements regarding the care and/or treatment process.

Figure 3.1 shows an illustrative, fictive ACP reference text.

3.1.2. Model selection

Based on the defined selection criteria in our approach, we chose two encoder-decoder and two decoder-only transformer models from the HuggingFace model library [19]:

Encoder-decoder:

- Flan-T5-large [32]
- yhavinga/t5-v1.1-base-dutch-cased [33]

Decoder-only:

- Llama-3.1-8B-instruct [34]
- Fietje-2-instruct [35]

Table 3.1 shows additional characteristics of the selected LLMs.

Who are you?
Thijmen is a 3-year-old boy with a progressive neurodegenerative disorder. He enjoys playing with his toys, reading books and being outside. Thijmen can communicate through sounds and gestures. A current point of attention is his ongoing limb pain. Thijmen is fully care-dependent. Both parents are involved in his care, as well as two 'personal budget' caregivers. The family is used to Thijmen's extra needs; his older brother (7) and sister (5) understand this, but it is sometimes hard for them that Thijmen needs more attention.
What is important to you?
Parents want Thijmen to be as comfortable as possible. It's important for them that he can still play and communicate with them and with his siblings. Currently, Thijmen is often uncomfortable due to his limb pains. Parents find this difficult to see.
Parents have slightly different views regarding the care for Thijmen. It's important for father to have tried every treatment, whereas for mother it's more important that Thijmen doesn't unnecessarily suffer.
What are your goals and wishes for future care and treatment?
Parents want to update Thijmen's pain medication plan to reduce his pain. Father would like to discuss an experimental treatment with Thijmen's neurologist. Mother is unsure whether she'd want this for him, but is open to the conversation.
Given Thijmen's unknown prognosis, parents have started thinking about their wishes in the end-of-life care period, but find it difficult to make concrete plans for this.

Figure 3.1: An illustrative, fictive ACP summary.

Table 3.1: Characteristics of the selected LLMs

Model	Size	Base model	Context Length	Instruction-tuned
Flan T5-large	780M parameters	Google/T5	512	Yes
T5-v1.1-large-dutch-cased	247M parameters	Google/T5	1,024	Yes, translation, summarisation
Fietje-2-instruct	2.7B parameters	Microsoft/phi-2	2,048	Yes
Llama-3.1-8B-instruct	8B parameters	Llama-3.1-8B	128,000	Yes

B = Billion, M = Million, Context Length = the maximum number of input tokens processed by the model

Inference tasks

For the model selection, we prompted each model to consecutively summarise seven subsections of the ICPs. We evaluated the generated texts using four criteria, as described in the methods section. Table 3.2 shows the number of texts that passed each criterion and the total scores as the sum of the corresponding criterion scores per model.

Table 3.2: Results of the inference tasks for each model.

Model	Dutch [1]	Sentences [2]	Coherent [3]	Summary [4]	Total score
Flan T5-large	6/7	0/7	0/7	0/7	0.86+0+0+0 = 0.86
T5-v1.1-large-dutch-cased	7/7	4/7	0/7	0/7	1+1.1+0+0 = 2.1
Fietje-2-instruct	7/7	7/7	4/7	2/7	1+2+1.7+1.1 = 5.8
Llama-3.1-8B-instruct	7/7	7/7	1/7	1/7	1+2+0.43+0.57 = 4.0

[weight], 6/7 = number of tasks that met criterion / total number of tasks

Based on these results, we selected Fietje-2-instruct, Llama-3.1-8B-instruct for ICL and QLoRA fine-tuning. Hereafter, we refer to these models as Fietje-2 and Llama-3.1.

3.2. Experimental evaluation

3.2.1. Dataset construction

We included the relevant sections of 38 pseudonymised ICPs in the dataset. For each ICP, one researcher wrote an ACP summary. To validate the quality of these summaries, three ACP experts assessed five out of a total sample of ten reference texts. Their feedback indicated that the summaries should, where applicable, also include current points of attention in the child's care, the family's religious beliefs, and the child's level of care dependence as ACP outcomes.

3.2.2. Model fine-tuning and evaluation

Prompt structure

Based on the ACP summary structure and the selected models, we determined the following prompt structure.

```
Instruction: Answer the following three questions based on the context:
1. Who are you?
2. What is important to you?
3. What are your goals and wishes for future care and treatment?

Use the following examples to guide your word choice.   (ICL-only)
Examples: {Examples}                                    (ICL-only)

Context: {Context}

Answers:
1.
```

Hyperparameter settings

Table 3.3 and Table 3.4 show the results of the parameter tuning. For both models, a temperature of 0.7 yielded the highest evaluation metric scores. Adding expertise did not further increase model performance and was therefore left out of the final prompt.

Table 3.3: Temperature exploration results Fietje-2

Temperature	BLEU	ROUGE-L	BERT	MEDCON
0.1	0.019	0.18	0.60	0.49
0.5	0.025	0.19	0.61	0.51
0.7	**0.038**	**0.21**	**0.62**	**0.63**
0.7, expertise	0.026	0.20	0.61	0.56
0.9	0.031	0.17	0.60	0.58

BLEU score displays modified precision (0-1). All other scores display F1 scores (0-1).

Table 3.4: Temperature exploration results Llama-3.1

Temperature	BLEU	ROUGE-L	BERT	MEDCON
0.1	0.050	0.22	0.62	0.66
0.5	0.054	0.21	0.62	0.65
0.7	**0.047**	**0.22**	**0.63**	**0.67**
0.7, expertise	0.037	0.21	0.61	0.61
0.9	0.031	0.19	0.60	0.60

BLEU score displays modified precision (0-1). All other scores display F1 scores (0-1)

ICL

We performed ICL on ten of the 38 samples in the dataset. For each run, we ensured that the evaluated sample itself was not used as an example in the prompt. The ICL process with Fietje-2 proceeded without issues across all values of k. In contrast, running the larger Llama-3.1 model repeatedly resulted in GPU out-of-memory errors, which made ICL with the original model infeasible. Therefore, we decided to quantise the model from 16-bit to 4-bit precision. For $k = 4$ and $k = 8$ examples, we additionally had to reduce the maximum context length from 128,000 to 4,000 tokens.

We included the average word count of the generated texts per condition in Appendix C. For both models, word count increased with higher k-values. At zero-shot prompting, the generated texts were on average shorter than the reference texts, whereas at ICL-4 they were longer. Fietje-2 produced longer texts at ICL-1 and ICL-2, while Llama-3.1 did so at zero-shot prompting and ICL-4. At ICL-8, both models generated shorter texts again, with Llama-3.1 still producing longer texts than Fietje-2 and the reference texts. Across all conditions, we observed considerable variation in word counts, particularly for Llama-3.1.

Automatic evaluation metrics
Figures 3.2 and 3.3 display the average metric scores and corresponding standard deviations of the ICL process for each model and k-value.

Syntactic similarity
The ROUGE-L F1 and BLEU scores of both models follow similar trends during the ICL process. For Llama-3.1, performance improved slightly when up to two in-context examples were provided, reaching maximum scores of BLEU $= 0.058$ and ROUGE-L F1 $= 0.22$ at ICL-1. The rise in ROUGE-L F1 score is caused by an increase in recall, which suggests that examples help the model capture a larger share of the wording that corresponds with the reference text and align more closely with its structure. Meanwhile, the stable precision trend for $k \leq 2$ indicates that the model does not simultaneously introduce non-matching words. In contrast, for Fietje-2, the BLEU and ROUGE-L F1 scores did not improve when examples were provided. Maximum BLEU and ROUGE-L F1 scores were reached at zero-shot prompting (0.045 and 0.20, respectively). The large decrease in ROUGE-L F1 score from the baseline to $k = 1$ is caused by a drop in ROUGE-L precision. This indicates that the addition of examples causes the model to generate more words that do not match the words in the reference text. Although the recall increases slightly from the baseline to $k = 1$, this is outweighed by the decline in precision.

Semantic similarity
The BERT F1 scores of both models are higher than their ROUGE-L F1 scores, suggesting that the generated texts show more semantic similarity than syntactic similarity with the reference texts. Llama-3.1 shows a similar trend to the syntactic metrics, reaching its highest BERT F1 score of 0.64 at ICL-1. This increase is again driven by a recall incline with stable precision, which indicates that adding an example improves the model's ability to generate texts with meanings closer to the reference text, without introducing unrelated information. When more than one example is added, however, both recall and precision decrease, suggesting that the additional information makes it more difficult for the model to distinguish relevant information in the provided texts.
Fietje-2 also shows a similar trend to the syntactic metrics. The model reached a maximum BERT F1 score of 0.62 at zero-shot prompting. As with the syntactic metrics, the addition of examples led to a decline in precision, while recall remains stable, which indicates that the model introduces unrelated information to the generated text without capturing more relevant content from the source text.

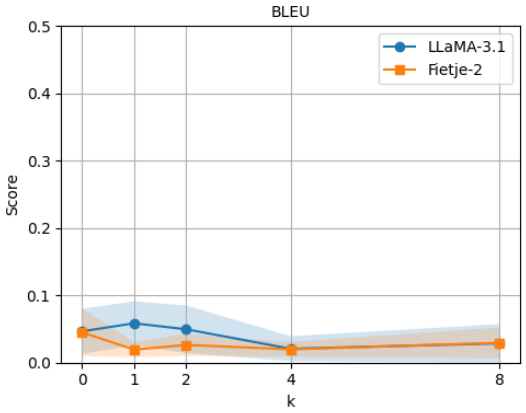


Figure 3.2: Average BLEU scores of ICL, plotted per k-value and per model. Shaded areas indicate standard deviations.
(Note that the y-axis limit is set to [0,05])

For both syntactic and semantic similarity metrics, model performance stabilises at higher k-values: at $k \geq 2$ for Fietje-2 and at $k \geq 4$ for Llama-3.1. This effect can be explained by the limited context length each model can handle: 2,048 tokens for Fietje-2 and 128,000 for Llama-3.1. Once this limit is reached, the model ignores all additional text, which can result in a stabilised performance. Given its lower context length, Fietje-2 reached this point for lower values of k. For Llama-3.1, the reduction of its context length to 4,000 tokens, due to repeated GPU out-of-memory errors, could explain why its performance stabilised at $k = 4$ already, i.e. at approximately twice the prompt length compared to Fietje-2.

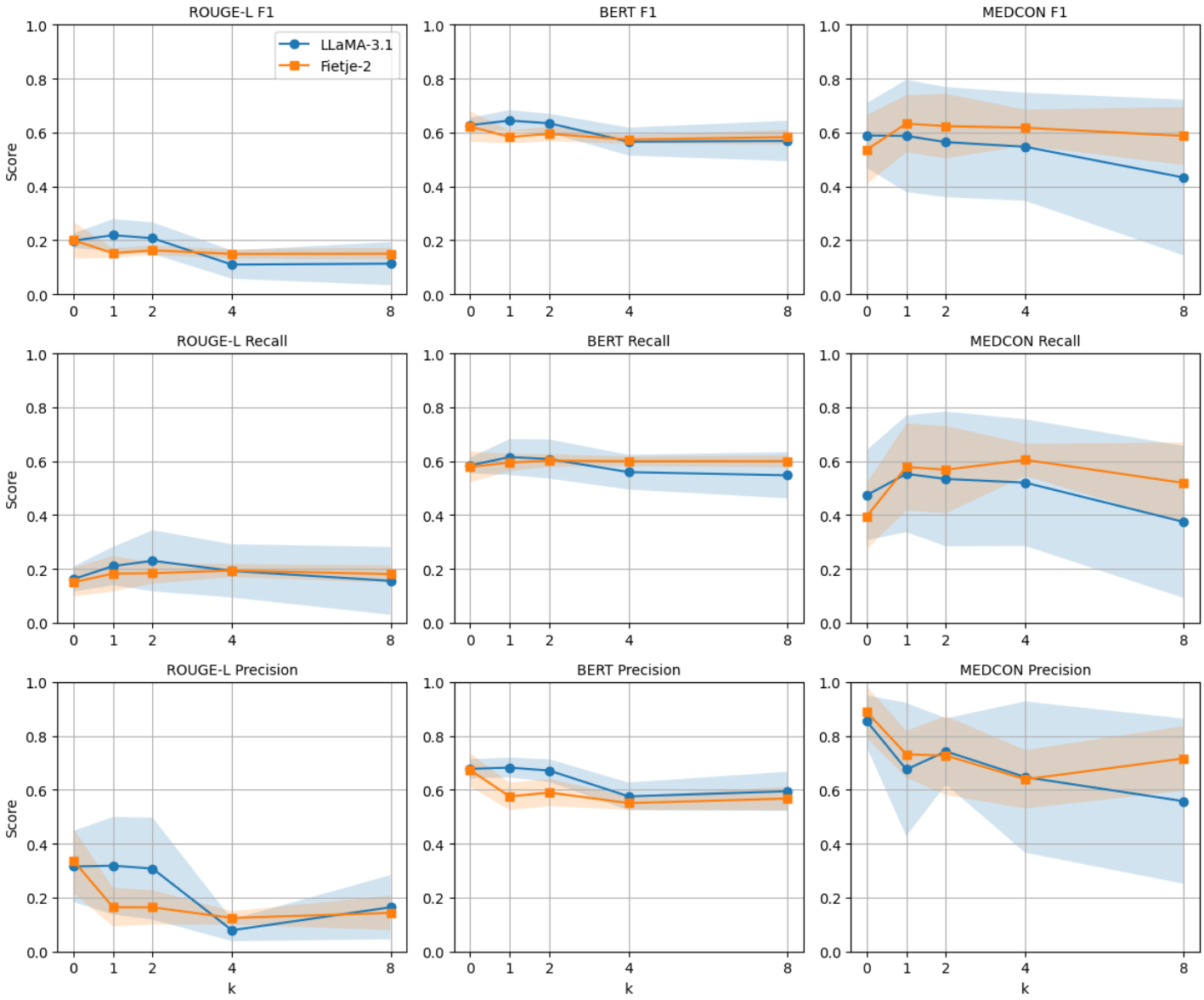


Figure 3.3: Average F1, recall and precision scores of ROUGE-L, BERT and MEDCON across zero-shot prompting and ICL, plotted per k-value and per model. Shaded areas indicate standard deviations.

Medical concept similarity

The two models show different trends in MEDCON scores compared to the other metrics, with larger standard deviations of the averaged scores. For both models, recall increased from zero-shot prompting to ICL-1, while precision decreased. For Fietje-2, this led to a higher F1 score, whereas for Llama-3.1, the increase in recall was smaller, leading to a practically unchanged F1 score. These results suggest that while both models benefited from an in-context example in terms of matching more medical concepts with the reference texts, the gain was larger for Fietje-2 than for Llama-3.1.

When analysing the overall distribution of the medical concept matches, without considering overlap between the texts, we found that the generated texts of both models and the reference texts showed a highly similar distribution (see Figure D.1 and Figure D.2 in Appendix D). Across all conditions, within the five selected UMLs semantic groups, most matches were assigned to the categories *Gene or Genome*, *Amino Acid, Peptide, or Protein* and *Biologically Active Substance*; categories not directly relevant to ACP content. Further examination of the medical concept matches showed that common Dutch function words (e.g.,*het, dan, een, van*) were frequently misclassified into these categories. This indicates that the UMLS-based

matching introduced a substantial number of false positives, which limits the interpretability of the observed distributions.
In addition, a comparison of the total word count (see Appendix C) and the total number of medical concept matches in the generated texts (see Appendix D) showed that longer texts corresponded to more medical concept matches. For both models, this was observed from zero-shot prompting to ICL-1. This finding suggests that the number of matched concepts is driven more by text length than by the presence of relevant medical content. Furthermore, this pattern could also explain the MEDCON recall increase observed here: longer generated texts yield more medical concept matches, which increases the likelihood of overlap with the matches in the reference texts. Since recall is defined as the proportion of overlapping medical concept matches between the generated and the reference text, relative to the total number of matches in the reference text (which remains stable in length), more overlaps directly translate into higher recall.

Textual analysis
As discussed in the previous section, Llama-3.1 achieved the highest metric scores for ICL-k, whereas Fietje-2 performed best during zero-shot prompting. Therefore, for each model, we selected the two baseline texts and the two ICL-1 texts with the highest BERT F1 scores.
Tables 3.5 and 3.6 show the results of the textual analysis of the generated texts at zero-shot prompting ICL-1 for Llama-3.1 and Fietje-2, respectively. If a model mentioned an incorrect age or condition, this was not included in the tables but discussed in the text. No episodes of repetitive texts were observed in any of the texts.

Table 3.5: Textual analysis of the texts generated by Llama-3.1 across zero-shot prompting ($k = 0$ examples) and ICL-1 ($k = 1$ example)

k	**Structure**			**Hallucinations**	**Interpretation shift**s
	QA	*Age*	*Condition*		
0	2/2	1/2	1/2	▪ Incorrect information: full answer on ACP Q2 is hallucinated: *"quiet and safe environment, without worrying for child's condition or development."*	▪ Answered ACP Q1 with: *"you (plural) are [name child] and her parents*." ▪ Answered ACP Q1 with: "you (plural) are a [correct age] old patient with [correct condition]."
1	2/2	1/2	0/2	▪ Incorrect information: copied answers to Q2 and Q3 from example. ▪ Incorrect information: copied all answers from example.	▪ Copied answers from example.

ACP = Advance Care Planning, GoC = Goals Of Care, k = number of examples, QA = Questions Answered, Q = Question,

Structure
All generated texts by Llama-3.1 followed the ACP summary structure. The child's age was mentioned by three out of four texts, of which one ICL-1 text mentioned the wrong age. This same text additionally mentioned the wrong condition. For Fietje-2, only the $k = 0$ texts followed the answer structure. One ICL-1 text only answered the second ACP question. None of the four texts mentioned children's ages or conditions.

Hallucinations
Both $k = 0$ and ICL-1 texts contained model hallucinations. For Llama-3.1, these mainly involved adding incorrect information to the texts that was not present in the source text. In one zero-shot text, the answer to the second question was entirely fabricated, while in both ICL-1 texts, most information was copied from the example instead of from the source text. Fietje-2 showed the same behaviour in the ICL-1 texts. In the zero-shot texts of Fietje-2, a variety of hallucinations was observed, including meaning shifts, a false emphasis and goal misattribution.

Interpretation shifts
We noticed that at zero-shot prompting, Llama-3.1 responded to the first ACP question (*"who are you"*) with a short literal answer, such as *"you are a patient"* or *"you are [name child] and her parents*. This literal interpretation suggests that the abstract meaning of the question was not captured by the model. Fietje-2, on the other hand, extracted the full answer to the ICP question *"who is the child"* as the answer to the first ACP question, at zero-shot prompting. Since this ICP answer contains much of the desired information

for the ACP answer, it could indicate that Fietje-2 might better understand the abstract meaning of the question.

Table 3.6: Results of the textual analysis of the texts generated by Fietje-2 across zero-shot prompting and ICL-1

k	Structure			Hallucinations	Interpretation shifts
	QA	*Age*	*Condition*		
0	2/2	0/2	0/2	▪ Meaning shift: rephrased relationship with parents. ▪ Meaning shift: rephrased information about child's name. ▪ False emphasis: incorrectly states sentence as most important. ▪ Goal misattribution: concludes a GoC not mentioned in the source text.	▪ Extracted full ICP answer to "who is the child" as response to ACP Q1. ▪ Rephrased ICP answer to "parental role" as response to ACP Q2.
1	0/2	0/2	0/2	▪ Incorrect information: included nothing from source text in both samples. ▪ Nonsensical text	▪ Copied information from example in both samples.

ACP = Advance Care Planning, ICP = Individual Care Plan, k = number of examples, QA = Questions Answered, Q = Question

Overall, the $k = 0$ texts of both models showed the best structural quality. For Llama-3.1, the number of hallucinations and interpretation shifts in the zero-shot prompting and ICL-1 texts was comparable, whereas for Fietje-2, fewer hallucinations and interpretation shifts were generated in the ICL-1 texts. However, we noticed that both models tended to copy most information from the examples in their ICL-1 texts. This indicates that both models experienced difficulty distinguishing relevant from less relevant information in the prompt. For Fietje-2, this observation was reflected in a decrease in BLEU and F1 scores (except MEDCON), from zero-shot to ICL-1. Interestingly, for Llama-3.1, the lower content quality during ICL-1 did not correspond to the BLEU and F1 scores, which showed a slight increase compared to zero-shot prompting. This may be explained by the fact that the copied segments from the example reference texts were already written in the desired ACP summary structure, which may have boosted the metric scores despite referring to a different ICP. Additionally, possible overlaps in ACP outcomes between ICPs might have further increased the automatic evaluation metrics.

QLoRA

We split the dataset into training, validation, and test sets of 30, 4 and 4 samples, respectively. For Fietje-2, QLoRA fine-tuning proceeded without issues. For Llama-3.1, however, training with a context length of 128,000 tokens again resulted in GPU out-of-memory errors. To enable training, we had to reduce the input context length to 1,024 tokens. We included the validation and training losses during the fine-tuning process in appendix E. For Fietje-2, the training process lasted for all 30 epochs, whereas for Llama-3.1, the early stopping condition was reached after 13 epochs.

Appendix C displays the average word count of the QLoRA-generated texts. On average, Llama-3.1 produced almost twice as many words as Fietje-2 and as the reference texts. In contrast, the texts generated by Fietje-2 remained within the average word count range of the reference texts. Unlike under zero-shot prompting and ICL conditions, Fietje-2 showed greater variability in word count than Llama-3.1

Evaluation metrics

Figure 3.5 and 3.5 show the results of the test set inference on the QLoRA fine-tuned models. Overall, we observe no great differences between the model performances. The BERT and MEDCON F1 scores are higher than the BLEU and ROUGE-L scores, which indicates a higher semantic and medical concept similarity than syntactic similarity between the generated and reference texts. The mean BLEU and ROUGE-L scores of Llama-3.1 display larger standard deviations, suggesting less consistent syntactic similarity performance compared to Fietje-2. In contrast, the mean MEDCON scores of Fietje-2 show larger standard deviations than Llama-3.1, here suggesting a less consistent capture of relevant medical concepts by Fietje-2. Both models show smaller standard deviations for the BERT scores, which could indicate that the models are more consistent in capturing relevant meaning from the source text.

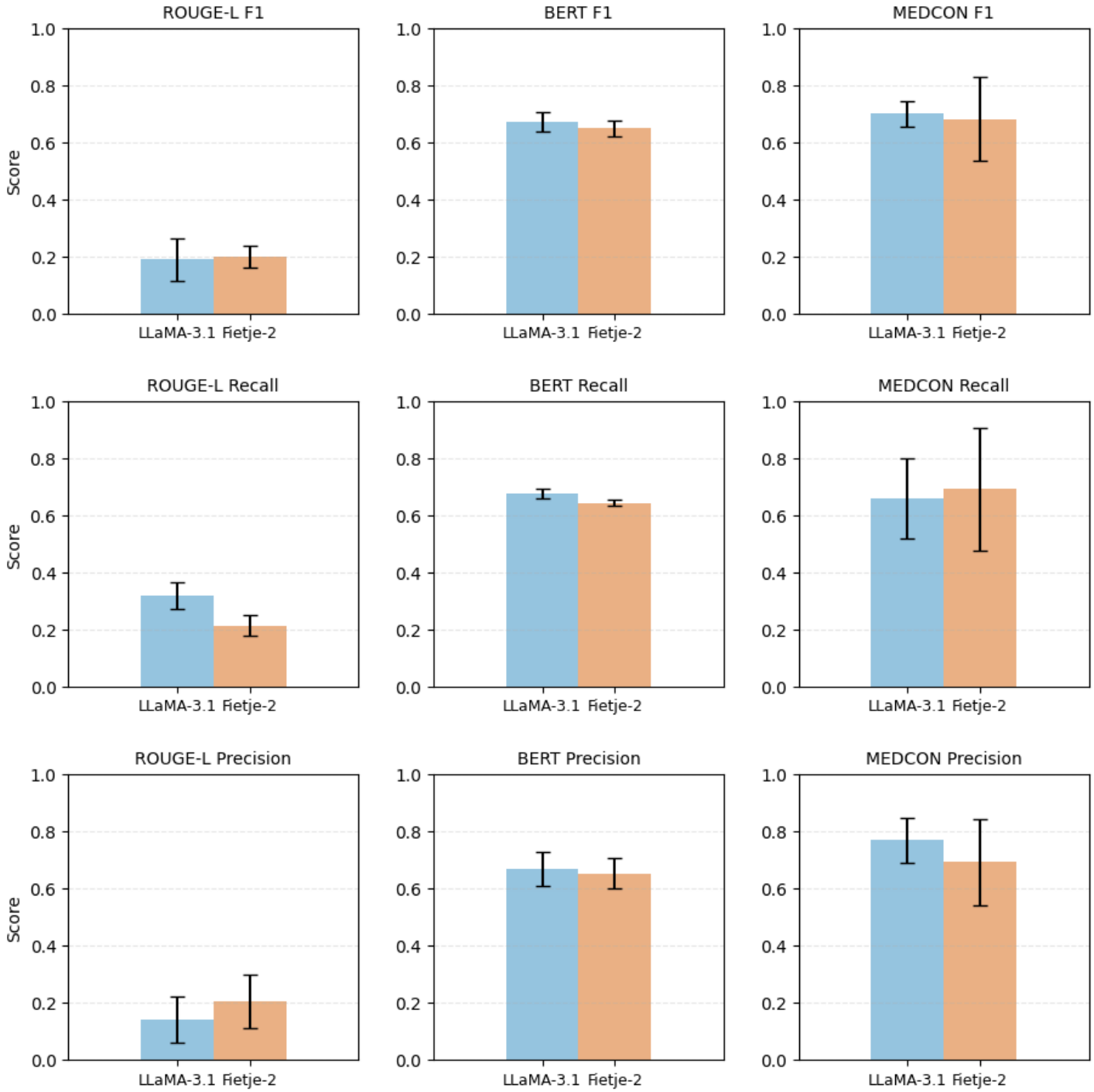


Figure 3.4: Average F1, recall and precision scores of ROUGE-L, BERT and MEDCON of the QLoRA fine-tuned versions of Llama-3.1 and Fietje-2. Error bars indicate standard deviations.

The distributions of medical concept matches in the QLoRA-generated texts (see Figure D.1 and Figure D.2 in Appendix D) were comparable to those observed under zero-shot prompting and ICL conditions, with the majority of matches falling into the categories *Gene or Genome*, *Amino Acid, Peptide, or Protein*, and *Biologically Active Substance*. As before, these categories are not directly relevant to ACP, which limits the interpretability of the observed distributions.

When comparing word counts (see Appendix C) and the total number of medical concept matches (see Appendix D), Llama-3.1 consistently produced longer texts with more concept matches than Fietje-2. Based on the patterns observed under zero-shot and ICL prompting, this would be expected to result in higher recall for Llama-3.1. However, Fietje-2 achieved higher recall despite its shorter outputs. This indicates that Fietje-2 performed better at reproducing the same medical concept matches as the reference texts, while the additional length of Llama-3.1 did not translate into more relevant overlap in matches.

Textual analysis

Table 3.7 shows the results of the structure analysis of the generated texts. Tables 3.8 and Table 3.9 show the results of the content analysis of Llama-3.1 and Fietje-2, respectively.

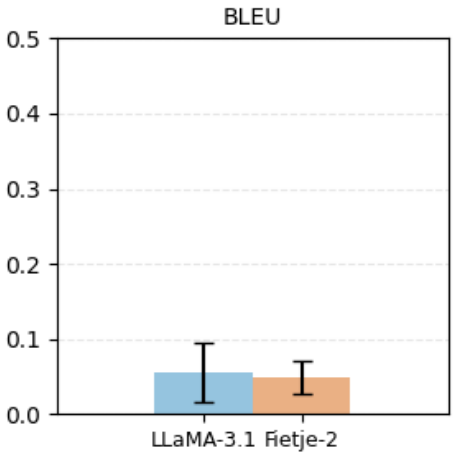


Figure 3.5: Average BLEU scores of the QLoRA fine-tuned versions of Llama-3.1 and Fietje-2. Error bars indicate standard deviations.
(*Note that the BLEU y-axis limit is set to [0,0.5]*)

Structure

All texts generated by Llama-3.1 followed the reference answer structure, compared to three texts of Fietje-2. In the fourth text, Fietje-2 extracted a set of questions from the source text and answered these instead of the ACP questions. For Llama-3.1, one text mentioned the child's condition in different wording than the source text, but still correctly. Another text described the condition as 'severe intellectual disability', which was not the primary condition of the child, but rather a consequence of it, and was therefore not counted as correct.

Table 3.7: Structural analysis of the texts generated by the QLoRa fine-tuned models.

Model	**Structure**		
	QA	*Age*	*Condition*
Fietje-2	3/4	1/4	0/4
Llama-3.1	4/4	3/4	1/4

QA = questions answered

Hallucinations

For Llama-3.1, most model hallucinations were coded as *incorrect information*, followed by *importance misattributions*, where the model stated incorrect statements about what parents or the child considered important. In a few cases, the model produced nonsensical text, either grammatically incorrect or unrelated to the source text. Fietje-2 hallucinated more frequently than Llama-3.1, also mostly producing *incorrect information* hallucinations. These varied from obvious errors, such as *"parents have no children"*, to more subtle ones, such as *"enjoys activity at school"*. The model also generated nonsensical texts, often together with periods of repetition.

Periods of repetition

Llama-3.1 repeated several hallucinations, as well as parts of its ACP answers. In two cases, the periods of repetition occurred at the end of the text and continued until the prior set maximum generation length (1,024 tokens) was reached. Fietje-2 showed fewer repetitions than Llama-3.1, with only two hallucinations being repeated.

Interpretation shifts

Llama-3.1 mainly placed information from the source text in the wrong context. In two cases, the model rephrased sentences from the source text, which changed their meaning and resulted in incorrect statements. Fietje-2 showed fewer interpretation shifts than Llama-3.1. In one case, the model responded to the first ACP question (*"who are you"*) with a short literal answer: *"mother is mama, father is papa"*, similar to the baseline texts of Llama-3.1. In another case, Fietje-2 extracted part of an ICP sentence about a parent describing their hobbies and placed it in a different context, leading to the incorrect statement: *"worrying gives her energy"*.

Overall, Llama-3.1 generated texts in the correct ACP summary structure more consistently than Fietje-2. Both models show similar content quality, where Fietje-2 hallucinated more frequently, but Llama-3.1 showed more repetitive episodes and interpretation shifts. These results correspond to the comparable BLEU and F1 scores of both models.

Table 3.8: Textual analysis of the texts generated by the QLoRA fine-tuned Llama-3.1.

Hallucinations		**Episodes of repetition**	**Interpretation shift**
8x Incorrect information	▪ *"also that he won't weaken in the future and that he won't pass away then"* ▪ *"he suffers from breathing problems"* ▪ *"he suffers from epilepsy"* ▪ *"likes to sit on mother's lap"likes to sit on child'sown lap"* ▪ *"he has two days of care during holidays"* ▪ Used wrong number of involved parents in text ▪ *"parents expect that [name child] will keep growing andthat his quality of life will stay as much as possible"* ▪ *"they expect that [name child] will pass way"*	▪ 3x Repetition of hallucinations ▪ Repetition of answer to ACP Q2 at Q3. ▪ Multiple repetitions of sentence about extramural care. ▪ Ongoing repetition of ACP answers. ▪ Ongoing repetition of nonsensical text at end of text. ▪ Repetition of wrong number of involved parents	▪ 2x Placement of information about mobility, care and treatment at ACP Q2 instead of at ACP Q1 and Q3 answer. ▪ Placed information that belonged to ACP Q2 at Q3. ▪ Placed the word *not* at the wrong position in the sentence, which made the meaning ambiguous ▪ Placed the excitation of stimuli right behind a sentence about pain which implied that stimuli came from pain, but that is not stated in the source text ▪ Added 8 rephrased sections of the source text after ACP Q3 ▪ Rephrased sentences about the end-of-life period of the child wrongfully.
3x Nonsensical text	▪ *"he has a tired body sensation"* ▪ *"parents have had pain in several phases and couldn't do it"* ▪ *"this will depend on the conditions"*		
5x Importance misattribution	▪ *"parents are looking for a life vision and what choices they should make"* ▪ *"parents want to know when there is a point and when the end is there"* ▪ *"important that the child can stay in faith"* ▪ *"want that [name child] can live his best life"* (not stated in source text) ▪ "...also that he doesn't feel restless"		
1x False emphasis	▪ Added that child reacts *well* to other people, while the source text only stated that the child is responsive to other people.		

ACP = Advance Care Planning, Q = Question

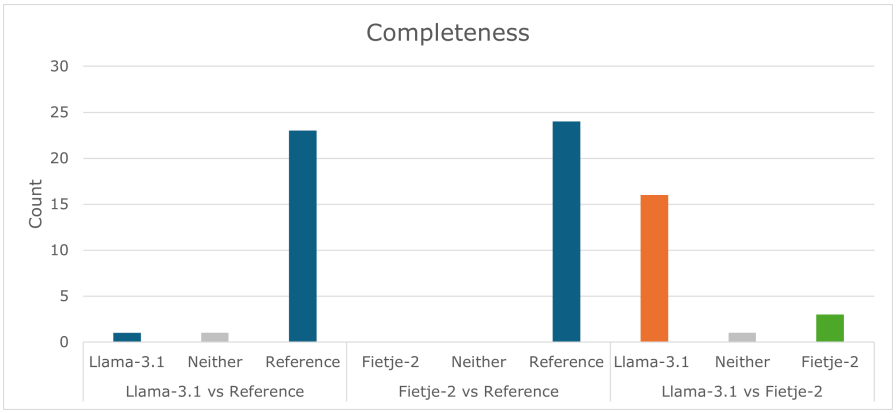
Table 3.9: Textual analysis of the texts generated by the QLoRA fine-tuned Fietje-2.

Hallucinations		**Episodes of repetition**	**Interpretation shift**
12x Incorrect information	▪ *"parents are relieved about the diagnosis"* ▪ wrote *"they are sad that child doesn't respond to world around him anymore"*, while child still can ▪ *"they are thankful that [name child] finally has a name and that he gets a chance to live for a few years"* ▪ *"Parents feel supported by presence of siblings"* ▪ *"... and has discussed this with her sisters"* ▪ *"parents have no children"* ▪ *"don't know what they should do"* ▪ *"life expectancy of less than a year"* ▪ *"... and because of this can listen very well and enjoysrelaxation"* ▪ *"enjoys activity at school"* ▪ *"mother seeks support in church"* ▪ *"parent is in process of writing down her experiences"*	▪ Repetition of hallucination ACP Q1. ▪ Ongoing repetition of hallucination at end of text.	▪ Responded to ACP Q1 with *"mother is mama, father is papa."* ▪ Extracted sentence about parent sharing that her hobby distracts her and gives her energy, but misplaced this in ACP text: *"worrying gives her energy"*
6x Nonsensical text	▪ *"they can't do everything"* ▪ *"parents like it"* ▪ *belt from the injection or an epidural"* ▪ *"he has for the condition of [name child]"* ▪ *"name of [child's name], Relevant context, Relevant contexts, Care of keeping, Context, Cases, Context, Illness, name of [child's name]..."* ▪ *"[name child] doesn't take answer pills. [name child] takesanswer pills and has a prikneus and doesn't get answer pills anymore"*		
3x Importance misattribution	▪ *"doesn't want child to end up in wheelchair"* ▪ *"they want to know when the care becomes too much to handle"* ▪ *"parents want to know what's next in the care for [name child]'"*		

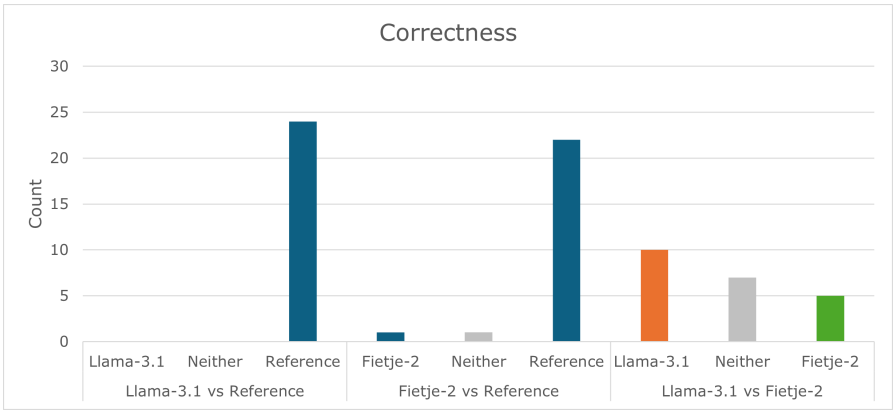
ACP = Advance Care Planning, Q = Question

Human reader study

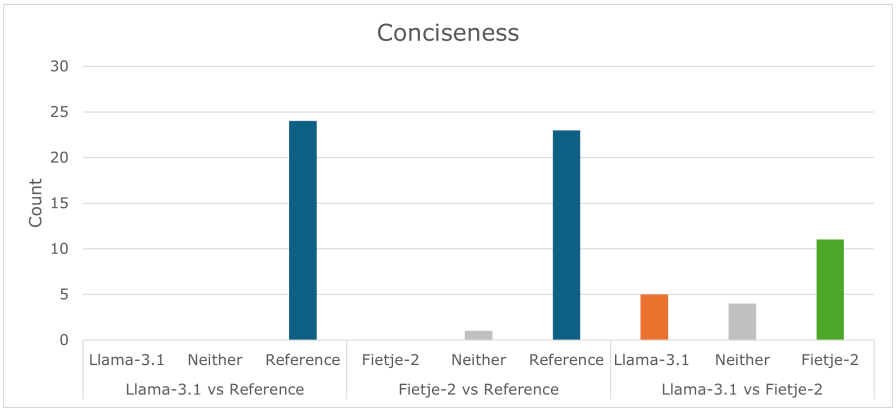
Figure 3.6 shows the results of the human reader study. Six HCPs assessed the four generated texts in the QLoRA test set of both models. Across all three criteria, the reference texts were consistently preferred over the model-generated texts, with only occasional single votes in favour of a model output or "neither." When comparing texts of Llama-3.1 and Fietje-2, readers showed a preference for Llama-3.1 regarding completeness, whereas Fietje-2 was preferred for conciseness. No clear preference was expressed in terms of correctness. These results show that human readers identified differences between the models that were not captured by the automatic metrics, which indicated comparable performance.



(a) Completeness



(b) Correctness



(c) Conciseness

Figure 3.6: Results of the human reader study. Each subfigure shows the number of human readers who preferred a certain model or the reference text for the corresponding category.

Comparison of zero-shot prompting, ICL-1 and QLoRA results

Figure 3.7 shows a comparison of the evaluation metrics of the zero-shot prompting, ICL-1 and QLoRA evaluations of the models. Overall, we notice no clear performance improvement after QLoRA finetuning. Llama-3.1 shows a slight increase in BERT F1 and MEDCON F1 scores, while BLEU and ROUGE-L F1 remain largely unchanged. For Fietje-2, only an increase in MEDCON F1 score is noticed. As seen in the ICL results section, Fietje-2 performs better when zero-shot prompted, whereas Llama-3.1 seems to benefit from a single example in the prompt.

Analysing the underlying recall and precision scores provides further details on these results. For both models, recall scores increase from zero-shot prompting to ICL-1 and QLoRA, suggesting that the models learn from an example or fine-tuning. However, for ROUGE-L and MEDCON, this gain is accompanied by lower precision scores, indicating that the higher recall is accompanied by the simultaneous addition of more irrelevant wording. For BERT, precision remains more stable, showing that model improvement does not come at the cost of precision.

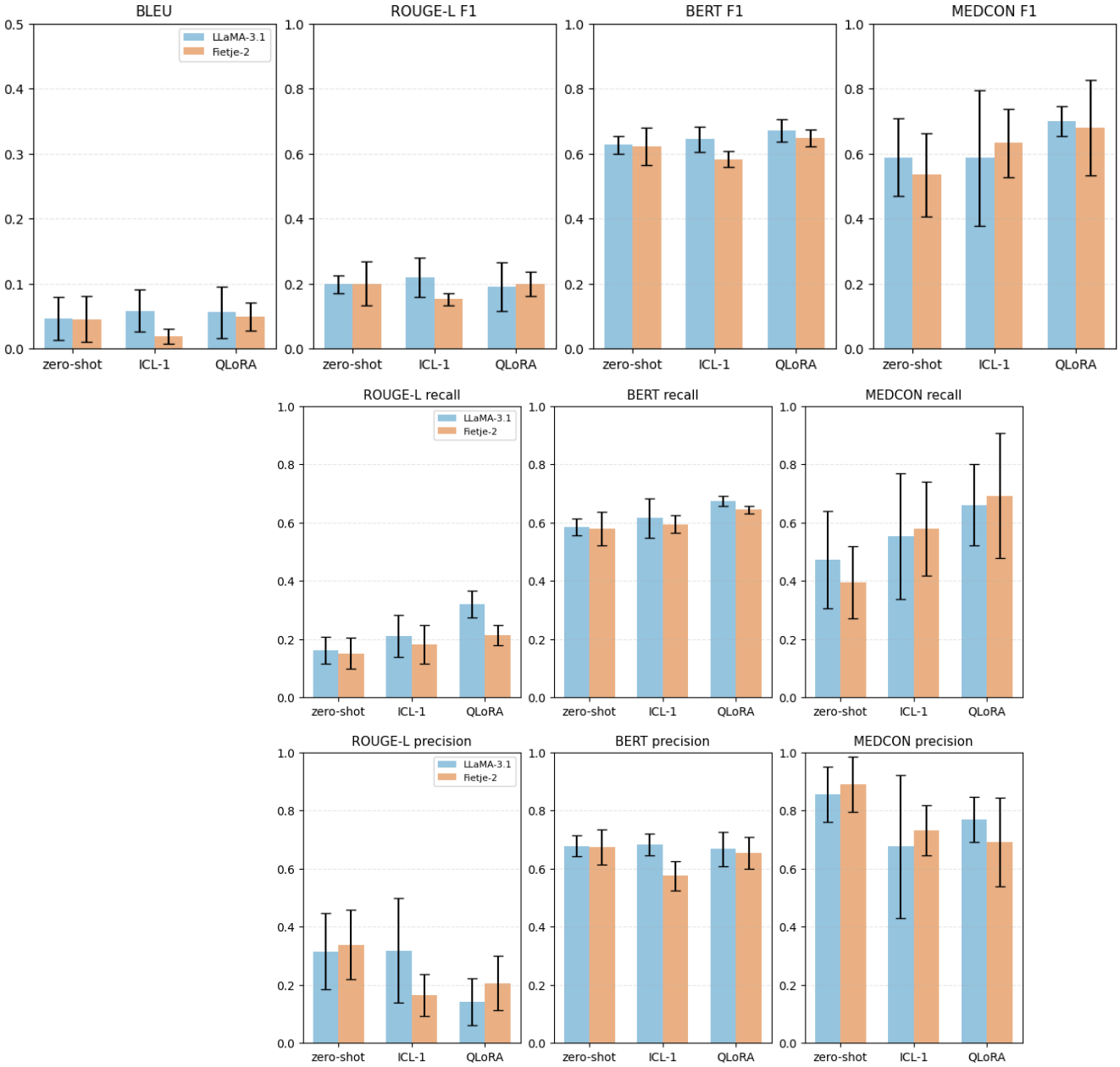


Figure 3.7: Comparison of the average automatic evaluation metrics scores across zero-shot prompting, ICL-1 and QLoRA fine-tuning, plotted per model. Error bars indicate standard deviations. (*Note that the BLEU y-axis limit is set to [0,0.5]*)

4

Discussion

In this study, we aimed to summarise the outcomes of documented ACP conversations using open-source LLMs. We developed a methodological approach that defined an ACP summary structure, compared different LLM architectures, explored ICL and QLoRA fine-tuning methods, and evaluated model performance using automatic metrics, textual analyses and human reader assessment. We applied this approach to two open-source models: Llama-3.1-8B-instruct and Fietje-2-instruct, and compared model performance at zero-shot prompting, ICL and QLoRA conditions.

In-context learning did not substantially improve model performance for either model. Llama-3.1 achieved its highest BLEU and F1 scores at ICL-1. However, textual analysis showed that structural quality was better at zero-shot prompting and that the higher metric scores at ICL-1 seemed to have been boosted by Llama-3.1 directly copying content from the example reference text. Fietje-2 achieved its highest BLEU and F1 scores at zero-shot prompting. Textual analysis confirmed that the structural quality of the texts was higher at zero-shot prompting than at ICL-1. At ICL-1, Fietje-2 also copied content from the example reference text, but unlike Llama-3.1, this copying behaviour did not translate to higher metric scores. A possible explanation for this difference is the shorter context length that Fietje-2 can handle. Once this length is exceeded, parts of the input text are truncated and not processed by the model. For Fietje-2, truncation occurred between k=0 and k=2 examples, depending on the sample, which likely reduced metric scores under these conditions. In addition, the sharper decrease in precision scores from zero-shot prompting to ICL-1, compared to Llama-3.1, suggests that Fietje-2 introduced more irrelevant wording and information, which further lowered its scores. Further textual analysis revealed hallucinations and interpretation shifts in all generated texts, which showed that both models struggled to select relevant information and tended to interpret the questions literally. While the copying behaviour indicates that the models recognised the desired answer structure in the example reference text, both models failed to translate this into meaningful generation based on the source text. These findings suggest that both models have a limited understanding of Dutch ACP conversations and, under the conditions of this study, lack the capacity to learn effectively from ICL.

The QLoRA fine-tuned versions of Llama-3.1 and Fietje-2 performed comparably. In the automatic metric evaluation, the BLEU and F1 scores showed no considerable differences between models. The BERT F1 scores indicated a semantic overlap of 60-70% with the reference texts for both models. The near-identical BERT recall and precision scores with relatively small standard variations indicate that both models consistently capture relevant information, while including a comparable amount of irrelevant content. The lower ROUGE-L F1 scores of both models indicate less overlap between wording and structure with the reference texts. Together, this suggests that both models succeeded reasonably in extracting relevant information, but often failed to place it in the correct structure. Analysis of text structure partially confirmed these results. While both models addressed all three questions in most outputs, they frequently failed to include age or condition, with Fietje-2 omitting these details more often than Llama-3.1.
Further textual analysis revealed additional differences between the models. Fietje-2 produced more hallucinations, including a higher number of nonsensical segments than Llama-3.1, which indicates a lower understanding of both the prompt questions and the ACP context. Although Llama-3.1 more frequently placed information in the wrong context, it seemed to more consistently extract this information from the

source text, suggesting a better understanding of which information to use when generating its answers. In contrast, Fietje-2 seemed to rely more on its pretrained knowledge than on the provided information during text generation.
In the human reader study, the generated texts were compared to each other and to the reference texts. Most readers found the texts produced by Llama-3.1 to be more complete than those by Fietje-2, which aligns with the results of the textual analysis. Interestingly, most readers preferred the texts by Fietje-2 for conciseness. This suggests that, even if it meant a loss of some detail, readers preferred shorter, more concise text, compared to longer, more elaborate responses. No clear preference was observed between the models in terms of correctness. This could indicate that readers considered all inaccuracies equally, regardless of their type. Further comparison of the generated texts with the reference texts revealed a nearly unanimous reader preference for the reference texts across all categories. This indicates that, despite some differences between the models, the overall quality of the generated texts remains substantially lower than that of the reference texts and is not yet suitable for clinical application.

We compared the results of the QLoRA fine-tuned models to those obtained at zero-shot prompting and ICL-1. Overall, no substantial improvement in model performance was observed after QLoRA fine-tuning. Although the metric evaluation showed a slight increase in BERT and MEDCON F1 scores, textual analysis indicated that the qualitative content of the generated texts remained largely unchanged.
A closer examination of the metric scores shows that the slight increase in BERT F1 scores for both models after QLoRA fine-tuning was primarily due to improved BERT recall scores, while BERT precision scores remained stable. This pattern was clearer for Llama-3.1 than for Fietje-2. For both models, the ROUGE-L and MEDCON metrics also showed increased recall, but this was accompanied by a drop in precision. These findings suggest that QLoRA fine-tuning on a small dataset allows the models to learn from the data to some extent, with greater improvements in semantic understanding than in accurate (medical) wording or text structure. Overall, given the current dataset and experimental conditions, QLoRA fine-tuning provides only minor improvements in metrics and does not result in substantially better textual quality of the ACP summaries.

Several factors might explain the limited model performance improvements we observed across zero-shot prompting, ICL and QLoRA conditions. First, the ACP summarisation task itself was complex. The summary questions were relatively abstract, and the source texts were long and contained multiple sections. Adding examples during ICL further increased prompt length and complexity. To explore whether the abstractness of the prompt itself explained part of the observed difficulty of the models, we conducted a small proof-of-concept with more directive instructions in the prompt (see Appendix F). The automatic metric results of this test were largely comparable to those of the experimental evaluation, suggesting that reformulating the instructions alone is insufficient to improve model performance in this study setting. It is important to note, however, that these findings are based on automatic metrics only. A manual textual analysis would provide a better insight into the effect of a different prompt formulation.
Second, for QLoRA, the small size of the fine-tuning dataset likely constrained the models' ability to learn the ACP summarisation task accurately. This is further confirmed by the comparison of evaluation metric scores between the test and training sets after zero-shot prompting on the QLoRA fine-tuned models (see Appendix G), which shows no substantial differences in metric scores between sets.
Third, although we expected Llama-3.1 to outperform Fietje-2 given its substantially larger model size, the results showed comparable performance. This outcome may be explained by differences in pretraining strategies and the computational constraints of this study. Fietje-2 is pre-trained specifically on Dutch, whereas Llama-3.1 is pre-trained on multilingual datasets. Moreover, Llama-3 was quantised, both of which may have reduced its capacity for this Dutch task. Fietje-2, by contrast, is smaller but specifically optimised for Dutch, which may have given it an advantage. Moreover, Fietje-2 has previously been shown to achieve results comparable to Geitje-7B [36], a Dutch-trained LLM of almost the same size as Llama-3.1 [35]. Altogether, these factors may explain why both models ultimately performed at a comparable level.

When comparing our results to previous studies investigating LLM performance on various clinical summarisation tasks, we observe both similarities and differences in study outcomes. Van Veen et al. [11] and Chao et al. [15] reported that models with shorter context lengths tended to show declining performance after zero-shot prompting, whereas models capable of processing longer inputs improved at ICL-1. This finding aligned with our study, where Fietje-2, with a shorter context length, achieved its best performance at zero-shot prompting, while Llama-3.1 showed slightly increased metric scores at ICL-1.
Regarding QLoRA, both Van Veen et al. [11] and Chao et al. [15] reported that fine-tuned models generally

outperformed zero-shot prompting and ICL conditions. We did not observe this effect in our results; however, these studies used substantially larger datasets (ranging from 1,000 to nearly 200,000 samples), whereas our dataset included only 30 samples. This likely limited the benefits of QLoRA fine-tuning and explains why we only observed none or only minor improvement compared to zero-shot prompting and ICL. Notably, Chao et al. [15] expressed a preference for QLoRA over ICL due to the risk of including information from examples, increased generation time, and higher computational requirements; factors that we also observed during ICL in our study.
As noted in the introduction, there is limited prior literature on clinical summarisation in a palliative care context. In this setting, Chen et al. [16] evaluated zero-shot summarisation of a single doctor–patient conversation, comparing different models. They observed comparable performance across models, with evaluation metric scores in line with our zero-shot prompting results. Nevertheless, direct comparison is limited by differences in language, model architectures and dataset characteristics.

Limitations

This study has several limitations. First, given its exploratory character, various empirical choices were made to obtain results within time and resource constraints. This may have influenced the observed outcomes and their generalisability. For instance, we chose a single prompt design and did not systematically compare different structures or formulations of the summary questions. In addition, collaboration with HCPs was limited, and the perspectives of PPC patients and of family members were not included. Regarding methodology, we relied on components from Van Veen et al. [11] and Chao et al. [15], specifically, the automatic evaluation metrics, fine-tuning strategies and human reader study. Finally, we restricted our model comparison to two open-source models, whereas a broader set of models would have allowed for a more comprehensive evaluation.
Second, the constructed dataset was small and included only a single reference ACP summary per ICP report. Although ACP experts were involved in defining the ACP summary structure and in evaluating a subset of the dataset, the reference texts themselves were written by a single researcher. This likely limited the variability of phrasing and style in the dataset and thereby reduced its representativeness. Moreover, treating these reference texts as the gold standard in the automatic metric evaluation lowered the interpretability of the results. In clinical practice, different HCPs use varied wording and may prioritise different ACP outcomes, meaning that there is no objective "true" ACP summary of an ICP report. Consequently, comparing the generated texts to the reference texts provided limited insight into clinical accuracy and may have evaluated the overlap in the researcher's phrasing more.
Third, as aforementioned, the abstract nature and length of the prompt may have reduced the models' ability to generate precise and relevant outputs. Lastly, the models were evaluated on a single GPU, requiring quantisation and input context length reduction of Llama-3.1.

Recommendations

For this specific study, several improvements could enhance model performance and evaluation. Adding multiple versions of ACP summaries per ICP would increase dataset size and variability. Providing one ACP question to the prompt per run and segmentising the source text could provide more structure and overview for the models, increasing their ability to distinguish relevant from irrelevant text. Furthermore, a systematic prompt study, in collaboration with ACP experts, would allow iterative evaluation and optimisation of the prompt design and the model output quality. Moreover, utilising larger or multicore GPUs would enable evaluation of Llama-3.1 in its original size, as well as larger LLMs. At the same time, it is important to refine the evaluation strategy. For future evaluation strategies, we recommend continuing with human evaluation and expanding both the textual analysis team and the human reader group to increase the reliability. Although automatic metrics provided only limited insight into text quality, they should still be included, as they require few resources and are easy to apply to study data. Specifically, we recommend maintaining BERTScore, comparing different ROUGE variants to better assess the utility of this syntactic similarity score, and narrowing the selection of MEDCON categories to those directly relevant to ACP content.

Looking ahead to clinical application, the main objectives would be to reduce the time and administrative burden of ACP conversations. To achieve these objectives, the model should be able to independently and accurately generate the desired documented outcome of the ACP conversations. This first requires a transition of model training data, from written ICP reports to textual recordings of full ACP conversations, e.g. by live-transcribing or obtained from audio or video recordings. Given the computational cost and limited gains of ICL observed in this study, we recommend prioritising QLoRA fine-tuning. Other than ICL,

this approach results in a customised model that can be continuously developed toward clinical applicability. Like in this study, QLoRA fine-tuning requires the construction of a labelled dataset. To increase dataset size and variability, we recommend combining completely transcribed ACP conversations with smaller fragments. In addition, providing multiple reference texts in varying styles and phrasing, written by ACP experts, will increase both the number of training samples and the clinical representativeness of the dataset.

With a sufficiently large and varied dataset, a base model can be fine-tuned iteratively. We suggest selecting a model that is competitive in size with proprietary models, potentially up to 60B parameters, and can handle a large context length. This requires substantial computational resources in a secure digital environment. A potentially suitable model is GPT-NL, a new, partially open-source LLM developed on Dutch texts only, expected at the end of 2025 [37]. This LLM is being developed as a non-profit initiative, trained on responsibly collected and transparently documented Dutch data. Although the model will be only partially open-source due to copyright restrictions, the weights will be made available to researchers on request, enabling fine-tuning.

Once a model is selected and fine-tuned on the augmented dataset, it can be deployed in a trial setting in clinical practice. With close involvement of HCPs and families in PPC, the model performance can then be evaluated, after which iterations of prompt and outcome structure can be applied. Over time, the dataset can be further augmented with new ACP conversation recordings and outcome texts. Periodic retraining and iterative prompt optimisation, guided by ACP expert evaluation, will be necessary until the model consistently produces summaries of sufficient quality.

5

Conclusion

Our study defined a methodological approach for deploying open-source LLMs to summarise ACP outcomes from ICP reports. We applied this approach to Llama-3.1-8B-instruct and Fietje-2-instruct under zero-shot prompting, ICL, and QLoRA conditions.

Across all conditions, the generated texts were significantly lower in quality than the reference texts, often containing hallucinations and irrelevant information. In-context learning did not improve performance, as both models tended to copy examples rather than learn the task. Similarly, QLoRA fine-tuning on a small dataset yielded only minor improvements, primarily in relevant information capture, rather than overall quality. Consequently, our findings indicate that both models have a limited capacity to accurately and reliably summarise ICP reports.

Several factors contributed to the models' limited performance. The small size and lack of clinical variability in our dataset, written by a single researcher, likely constrained the models' ability to sufficiently learn the ACP summarisation task. Additionally, the lengthy prompts along with computational constraints might have also negatively influenced performance.

To advance toward clinical applicability, future research should focus on fine-tuning an open-source LLM that is competitive with proprietary models, using a clinically representative dataset that consists of transcribed ACP conversations instead of ICP reports. We recommend optimising the prompt design and outcome structure in collaboration with ACP experts as well as with PPC patients and their family members. Furthermore, we recommend including multiple ACP summaries per conversation, written by ACP experts, to increase the dataset size and variability. Then, through continuous retraining, expert evaluation and prompt optimisation, careful generation of ACP summaries in clinical practice could be achieved, reducing the administrative burden and contributing to the broader implementation of ACP.

References

[1] *Kinderpalliatieve zorg*. URL: https://kinderpalliatief.nl/over-kinderpalliatieve-zorg/wat-is-kinderpalliatieve-zorg/kinderpalliatieve-zorg.

[2] Kim C. van Teunenbroek et al. "A Dutch paediatric palliative care guideline: a systematic review and recommendations on advance care planning and shared decision-making". In: *BMC Palliative Care* 23.1 (Dec. 2024), p. 270. ISSN: 1472684X. DOI: 10.1186/S12904-024-01568-3. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11585104/.

[3] Kim C. van Teunenbroek et al. "A Dutch paediatric palliative care guideline: a systematic review and evidence-based recommendations for symptom treatment". In: *BMC Palliative Care* 23.1 (Dec. 2024). ISSN: 1472684X. DOI: 10.1186/S12904-024-01367-W,. URL: https://pubmed.ncbi.nlm.nih.gov/38481215/.

[4] *Richtlijn Palliatieve zorg voor kinderen*. Nov. 2022. URL: https://palliaweb.nl/richtlijnen-palliatieve-zorg/richtlijn/palliatieve-zorg-voor-kinderen.

[5] Sefika Aldas et al. "Assessment of Caregiver Burden and Burnout in Pediatric Palliative Care: A Path Toward Improving Children's Well-Being". In: *Healthcare* 13.13 (July 2025), p. 1583. ISSN: 22279032. DOI: 10.3390/HEALTHCARE13131583. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12250540/.

[6] Angelika Eichholz, Joanne Dudeney, and Tiina Jaaniste. "Caregiver Psychological Burden in Pediatric Chronic Pain: A Systematic Review and Meta-Analysis of Associations with Caregiver Sociodemographic and Biopsychosocial Variables". In: *Journal of Pediatric Psychology* 48.9 (Sept. 2023), pp. 747–758. ISSN: 0146-8693. DOI: 10.1093/JPEPSY/JSAD041. URL: https://dx.doi.org/10.1093/jpepsy/jsad041.

[7] Judith A.C. Rietjens et al. "Definition and recommendations for advance care planning: an international consensus supported by the European Association for Palliative Care". In: *The Lancet Oncology* 18.9 (Sept. 2017), e543–e551. ISSN: 14745488. DOI: 10.1016/S1470-2045(17)30582-X. URL: https://pubmed.ncbi.nlm.nih.gov/28884703/.

[8] Danielle Jansen and Károly Illy. "Paediatric Care in The Netherlands: State of Affairs, Challenges and Prospects". In: *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 1037* 19.3 (Jan. 2022), p. 1037. ISSN: 1660-4601. DOI: 10.3390/IJERPH19031037. URL: https://www.mdpi.com/1660-4601/19/3/1037/htm%20https://www.mdpi.com/1660-4601/19/3/1037.

[9] Lorna K. Fraser et al. "Rising national prevalence of life-limiting conditions in children in England". In: *Pediatrics* 129.4 (Apr. 2012). ISSN: 00314005. DOI: 10.1542/PEDS.2011-2846,. URL: https://pubmed.ncbi.nlm.nih.gov/22412035/.

[10] Geronimo Jimenez et al. "Overview of Systematic Reviews of Advance Care Planning: Summary of Evidence and Global Lessons". In: *Journal of Pain and Symptom Management* 56.3 (Sept. 2018), pp. 436–459. ISSN: 18736513. DOI: 10.1016/j.jpainsymman.2018.05.016. URL: https://pubmed.ncbi.nlm.nih.gov/29807158/.

[11] Dave Van Veen et al. "Adapted large language models can outperform medical experts in clinical text summarization". In: *Nature Medicine* 30 (Apr. 2024), pp. 1134–1142. ISSN: 1546170X. DOI: 10.1038/s41591-024-02855-5.

[12] *Natural Language Processing and Large Language Models - Hugging Face LLM Course*. URL: https://huggingface.co/learn/llm-course/chapter1/2?fw=pt.

[13] Ashish Vaswani et al. "Attention Is All You Need". In: (June 2017), p. 1. URL: https://arxiv.org/pdf/1706.03762.

[14] Kerstin Denecke et al. "Potential of Large Language Models in Health Care: Delphi Study". In: *Journal of Medical Internet Research* 26.1 (2024). ISSN: 14388871. DOI: 10.2196/52399,. URL: https://pubmed.ncbi.nlm.nih.gov/38739445/.

[15] Chieh Ju Chao et al. "Evaluating large language models in echocardiography reporting: opportunities and challenges". In: *European Heart Journal - Digital Health* 6.3 (May 2025), pp. 326–339. ISSN: 26343916. DOI: 10.1093/EHJDH/ZTAE086. URL: https://dx.doi.org/10.1093/ehjdh/ztae086.

[16] Xiao Chen et al. "Exploring the opportunities of large language models for summarizing palliative care consultations: A pilot comparative study". In: *Digital Health* 10 (Jan. 2024), p. 20552076241293932. ISSN: 20552076. DOI: 10.1177/20552076241293932. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11577459/.

[17] Jurrianne Fahner and Marijke Kars. *The development of IMPACT*. URL: https://kinderpalliatief.nl/impact/en/about-impact.

[18] *Individueel Zorgplan of IZP*. URL: https://kinderpalliatief.nl/ondersteuning/belangrijke-methodes-tools/izp.

[19] *Models - Hugging Face*. URL: https://huggingface.co/models?sort=trending&search=pos.

[20] Hugging Face. *The Hugging Face Course, 2022*. https://huggingface.co/course. 2022.

[21] Saif Khairat et al. "Performance Evaluation of Popular Open-Source Large Language Models in Healthcare". In: *Studies in Health Technology and Informatics* 328 (June 2025), pp. 215–219. ISSN: 18798365. DOI: 10.3233/SHTI250705,. URL: https://pubmed.ncbi.nlm.nih.gov/40588913/.

[22] *How do Transformers work? - Hugging Face LLM Course*. URL: https://huggingface.co/learn/llm-course/chapter1/4?fw=pt.

[23] Qingxiu Dong et al. "A Survey on In-context Learning". In: *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (Dec. 2022), pp. 1107–1128. DOI: 10.18653/v1/2024.emnlp-main.64. URL: https://arxiv.org/pdf/2301.00234.

[24] *LoRA (Low-Rank Adaptation) - Hugging Face LLM Course*. URL: https://huggingface.co/learn/llm-course/chapter11/4?fw=pt.

[25] *What is In-Context Learning (ICL)? | IBM*. URL: https://www.ibm.com/think/topics/in-context-learning.

[26] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (2002), pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

[27] Chin-Yew Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. 2004. URL: https://aclanthology.org/W04-1013/.

[28] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *8th International Conference on Learning Representations, ICLR 2020* (Apr. 2019). URL: https://arxiv.org/pdf/1904.09675.

[29] Wen wai Yim et al. "Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation". In: *Scientific Data* 10.1 (Dec. 2023), pp. 1–16. ISSN: 20524463. DOI: 10.1038/S41597-023-02487-3;SUBJMETA. URL: https://www.nature.com/articles/s41597-023-02487-3.

[30] Wietse de Vries et al. "BERTje: A Dutch BERT Model". In: (Dec. 2019). URL: http://arxiv.org/abs/1912.09582.

[31] *What is LLM Temperature? | IBM*. URL: https://www.ibm.com/think/topics/llm-temperature.

[32] *google/flan-t5-large · Hugging Face*. URL: https://huggingface.co/google/flan-t5-large.

[33] *yhavinga/t5-v1.1-base-dutch-cased · Hugging Face*. URL: https://huggingface.co/yhavinga/t5-v1.1-base-dutch-cased.

[34] *meta-llama/Llama-3.1-8B-Instruct · Hugging Face*. URL: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

[35] Bram Vanroy. *Fietje: An open, efficient LLM for Dutch*. 2024. URL: https://arxiv.org/abs/2412.15450.

[36] Bram Vanroy. "GEITje 7B Ultra: A Conversational Model for Dutch". In: (Dec. 2024). URL: http://arxiv.org/abs/2412.04092.

[37] *Een verantwoord alternatief op bestaande LLMs - GPT-NL*. URL: https://gpt-nl.nl/.

A

Individual Care Plan

Individueel Zorgplan Palliatieve Zorg voor Kinderen

1. GEGEVENS ZORGPLAN

- **Datum (gepland) ACP gesprek** [Klik of tik om een datum in te voeren]
- **ACP gesprek gevoerd door** [naam/functie]
- **Opgesteld door** [naam/functie opstellers]
- **Opgesteld op** [Klik of tik om een datum in te voeren]
- **Versienummer** [nr.]
- **Laatste update gehad op** [Klik of tik om een datum in te voeren]
- **Voornaamste wijzigingen** [Open tekstveld]
- **Besproken met kind (wanneer mogelijk)** [Kies een item], [datum]
- **Besproken met ouders* op** [klik of tik om een datum in te voeren]
- **Besproken met hoofdbehandelaar op** [Klik of tik om een datum in te voeren]
- **Wordt herzien op** [klik of tik om een datum in te voeren]
- **Gedeeld met** [Geef aan met wie het IZP gedeeld is]
- **Datum waarop gedeeld** [klik of tik om een datum in te voeren]

**Overal waar ouders staat kan ook voogd/verzorgers worden gelezen.*

2. ALGEMENE INFORMATIE

2A GEGEVENS KIND

- **Naam** [Open tekstveld]
- **Geboortedatum** [Klik of tik om een datum in te voeren.]
- **Geslacht** [Open tekstveld]
- **Adres** [Adres kind]
- **Tweede adres** [Wanneer twee adressen geef aan bij wie)]
- **Telefoonnummers (kind, ouders)**
  - **1** [Maak een keuze] [Telefoonnummer]
  - **2** [Maak een keuze] [Telefoonnummer]
  - **3** [Maak een keuze] [Telefoonnummer]
- **Mailadres(sen)** [Open tekstveld]
- **Spreektaal** [Let op: kan verschillen tussen ouder en kind]
  - **Begrip Nederlandse taal** [Maak een keuze]
  - **Tolk nodig of sprake van communicatieproblemen** [Open tekstveld]

2B PRAKTISCHE ZAKEN

- **Gezinssamenstelling**
  [namen gezinsleden, relatie tot kind (en: wanneer broertjes en zusjes, ook leeftijd/geboortejaar)]

- **Wettelijke leefsituatie**
  [Wettelijke situatie, indien gescheiden partner(s): voogdij en woonregeling]

- **Woonsituatie**

[Praktische aspecten (benedenwoning? Waar bevinden zich de slaapkamer en badkamer?)]

- **Indicatie**
  [Hoe is de zorg financieel geregeld?]

- **Werkomstandigheden ouders**
  [Open tekstveld]

2C ZORGTEAM

Hoofdbehandelaar	*Naam, telefoon, emailadres,*
Andere betrokken kinderarts(en) en/of verpleegkundig specialist	*Naam, telefoon, emailadres,*
Kinder Comfort Team	*Naam, telefoon, emailadres,*
Huisarts	*Naam, telefoon, emailadres,*
Thuiszorg/PGB'er	*Naam, telefoon, emailadres,*
Apotheek	*Naam, telefoon, emailadres,*
Revalidatiearts	*Naam, telefoon, emailadres,*
Fysiotherapeut	*Naam, telefoon, emailadres,*
Ergotherapeut	*Naam, telefoon, emailadres,*
Logopedist	*Naam, telefoon, emailadres,*
Diëtist	*Naam, telefoon, emailadres,*
Maatschappelijk werk	*Naam, telefoon, emailadres,*
Psycholoog	*Naam, telefoon, emailadres,*
Geestelijk verzorger	*Naam, telefoon, emailadres,*
Rouw- en verliesbegeleider	*Naam, telefoon, emailadres,*
Medisch pedagogisch zorgverlener	*Naam, telefoon, emailadres,*
School/dagopvang	*Naam, telefoon, emailadres,*
Logeeropvang	*Naam, telefoon, emailadres,*
Overig, namelijk…	*Naam, telefoon, emailadres,*

2D CONTACTMOGELIJKHEDEN

<u>Contactmogelijkheden voor kind en ouders</u>

Bij acute vragen
[Open tekstveld]

Bij niet acute vragen
[Open tekstveld]

<u>Contactmogelijkheden voor zorgverleners</u>
Bij acute vragen
[Open tekstveld]

Bij niet acute vragen
[Open tekstveld]

3. WAARDEN, DOELEN EN VOORKEUREN

3A WIE IS HET KIND?

- **Wie is [naam kind]?**
  [Beschrijf wie het kind is. Hoe staat het kind in het leven. Creëer een zo volledig mogelijk beeld van het kind. Denk hierbij ook aan de ontwikkeling van het kind.]

- **Zorg ondersteuning/zelfredzaamheid**
  [In welke mate zelfstandig/welke mate hulp nodig?]

- **Dagstructuur**
  [Beschrijving dagstructuur]

- **School/opvang/vrije tijdsbesteding**
  [Beschrijving van school/opvang, afstemming met school/opvang, korte beschrijving eventuele hobby's/sporten, afstemming]

3B BELEVING ZIEKTE EN LEVENSVISIE

- **Hoe zien kind, ouders en broertjes/zusjes de ziekte van het kind en wat betekent het voor hun leven?**
  [Wat betekent de ziekte of aandoening voor kind, ouders en het gezin? Denk ook aan vragen rondom hoop/angsten/zorgen]

- **Levensvisie**
  [Wat doet het kind als hij/zij getroost wil worden? Is een bepaalde levensovertuiging belangrijk voor kind en ouders? Waar halen kind en ouders steun vandaan in het leven? Denk ook aan religie en spiritualiteit.]

- **Ouderrol**
  [Hoe zijn de ouders er voor het kind en wat voor ouder willen zij graag zijn?]

3C HUIDIGE SITUATIE

- **Wat vinden kind en ouders belangrijk in de huidige situatie?**
  [Open tekstveld]

3D TOEKOMSTVERWACHTING

- **Wat verwachten kind en ouders van de toekomst?**
  [Wat streven kind en ouders na? Waar hopen kind en ouders op? Waar droomt het kind van?

- **Wat vinden kind en ouders belangrijk mocht het kind geleidelijk achteruitgaan/acuut achteruitgaan/het levenseinde nabij lijken?**
  [Open tekstveld]

4. ZORG EN BEHANDELING

4A DOELEN VOOR ZORG EN BEHANDELING

- **Wat zijn de waarden en voorkeuren van kind en ouders**?
  [Beschrijf hier de waarden en voorkeuren van kind en ouders die uit de advance care planning gesprekken naar voren zijn gekomen]

- **Gezamenlijke doelen voor zorg en behandeling**
  [Benoem hier de gezamenlijke doelen van kind, ouders en zorgverleners die uit de advance care planning gesprekken naar voren zijn gekomen]

- **Geef aan welke afspraken er zijn gemaakt m.b.t. het nemen van beslissingen**
  [bijv. wie moeten er bij beslissingen altijd betrokken worden. Geef ook aan om welke beslissingen het gaat en wat de rol van het kind zelf is.]

- **Gemaakte afspraken over zorgdoelen n.a.v. de waarden en voorkeuren**
  [Open tekstveld]

- **Eventuele verschillen van inzicht/dilemma's**
  [Knelpunten noteren, bijv. waar ouders het over oneens zijn met elkaar of met de zorgverlener of wanneer zorgverleners het oneens zijn met elkaar.]

  Denk ook aan herinneringen maken, zie bijlage 2.

4B MEDISCHE SITUATIE

- **Relevante voorgeschiedenis**
  [Open tekstveld]
- **Diagnose**
  [Open tekstveld]
- **Actuele problematiek**
  [Open tekstveld]

- **Verwachting voor de toekomst**
  [Open tekstveld]
- **Beschrijf wat er is besproken m.b.t. diagnose en verwachtingen voor de toekomst met kind en ouders en hoe kind en ouders hier zelf naar kijken**
  [Open tekstveld]
- **Allergieën**
  [Open tekstveld]
- **Bijwerkingen op (eerdere) medicatie**
  [Open tekstveld]

4C AFSPRAKEN RONDOM DE BEHANDELING

Welke afspraken zijn er gemaakt rondom de behandeling die van invloed kunnen zijn op de levensduur van het kind?

Met wie is onderstaande besproken en wanneer? [Open tekstveld]

- **Reanimatiebeleid**
  - **Hartmassage** Kies een item. [Open tekstveld]
  - **Circulatie; medicamenteuze ondersteuning** Kies een item. [Open tekstveld]
  - **Ventilatie; uitzuigen** Kies een item. [Open tekstveld]
  - **Ventilatie; zuurstoftoediening** Kies een item. [Open tekstveld]
  - **Ventilatie; masker- en ballon** Kies een item. [Open tekstveld]
  - **Ventilatie; intubatie/mechanische ventilatie** Kies een item. [Open tekstveld]
- **Verrichten diagnostiek** Kies een item. [Open tekstveld]
- **Medicatie bij infectie (AB)** Kies een item. [Open tekstveld]
- **Transfusiebeleid** Kies een item. [Open tekstveld]
- **Opname** Kies een item. [Open tekstveld]
- **Intensive Care opname** Kies een item. [Open tekstveld]
- **Vocht en voeding** Kies een item. [Open tekstveld]
- **Overig:** [Open tekstveld]

4D VOEDING

- **Beschrijving voedingspatroon**
  [Omschrijving dagelijkse voeding/vochtinname/sondevoeding (systeem, schema)]

- **Wensen van kind/ouders en doel m.b.t. (sonde)voeding**
  [Open tekstveld]

- **Afspraken omtrent vocht, voeding en supplementen**
  [Open tekstveld]

Voor vocht en voeding in de laatste levensfase, zie levenseindezorg.

4E HUIDIGE SYMPTOMEN EN TE VERWACHTEN SYMPTOMEN PASSEND BIJ DE AANDOENING

Huidige symptomen	**Symptomen verwacht in de toekomst**
☐ Angst en depressie	☐ Angst en depressie
☐ Delier	☐ Delier
☐ Dyspnoe	☐ Dyspnoe
☐ Hematologische verschijnselen	☐ Hematologische verschijnselen
☐ Hoesten	☐ Hoesten
☐ Huidklachten	☐ Huidklachten
☐ Misselijkheid en braken	☐ Misselijkheid en braken
☐ Neurologische symptomen	☐ Neurologische symptomen
☐ Obstipatie	☐ Obstipatie
☐ Pijn	☐ Pijn
☐ Reutelen	☐ Reutelen
☐ Vermoeidheid	☐ Vermoeidheid

Geef indien relevant per symptoom aan:
- Is er diagnostiek verricht (inclusief datum) of gepland (inclusief datum)?
- Welke adviezen voor niet medicamenteuze behandeling zijn gegeven (inclusief datum)?
- Welke adviezen voor medicamenteuze behandeling zijn gegeven (inclusief datum)?

Zie Richtlijn palliatieve zorg voor kinderen voor advies m.b.t. symptomen.

Gewicht in kg waarop onderstaande behandeladviezen zijn gebaseerd [gewicht in kg] op [Klik of tik om een datum in te voeren.]

Onderstaande is besproken door [Open tekstveld] op [Klik of tik om een datum in te voeren.]
Voor starten behandeling afstemmen met [Open tekstveld]

[vul naam in symptoom 1]:
- Is er diagnostiek verricht (inclusief datum) of gepland (inclusief datum)?
- Welke adviezen voor niet medicamenteuze behandeling zijn gegeven (inclusief datum)?
- Welke adviezen voor medicamenteuze behandeling zijn gegeven (inclusief datum)?

[Open tekstveld]

[vul naam in symptoom 2]:
- Is er diagnostiek verricht (inclusief datum) of gepland (inclusief datum)?

- Welke adviezen voor niet medicamenteuze behandeling zijn gegeven (inclusief datum)?
- Welke adviezen voor medicamenteuze behandeling zijn gegeven (inclusief datum)?

[Open tekstveld]

[vul naam in symptoom 3 etc.]:
- Is er diagnostiek verricht (inclusief datum) of gepland (inclusief datum)?
- Welke adviezen voor niet medicamenteuze behandeling zijn gegeven (inclusief datum)?
- Welke adviezen voor medicamenteuze behandeling zijn gegeven (inclusief datum)?

[Open tekstveld]

Indien nodig; kopieer bovenstaande.

Voor een volledig invulbaar formulier voor symptomatologie, zie bijlage 3.

4F SCENARIO'S

- **Beschrijf besproken mogelijke scenario's en bijbehorende wensen/afspraken.**
  [Open tekstveld]

Bovenstaande is besproken door [Open tekstveld] op [Klik of tik om een datum in te voeren].

4G COMPLEMENTAIRE ZORG

- **Overzicht complementaire zorg**
  [Open tekstveld]

- **Handelingen m.b.t. ontspanning**
  [Open tekstveld]

5. LEVENSEINDEZORG

5A LEVENSEINDE

- **Beschrijf de voorkeuren en afspraken m.b.t. levenseinde**
  [Beschrijf de voorkeuren (bijv. plaats) en bijzonderheden m.b.t. levenseinde. Denk ook aan gewoontes/rituelen/principes waar zorgverleners rekening mee kunnen houden.]

- **Afscheid en uitvaart**
  [Beschrijf de wensen rondom het afscheid en uitvaart. Denk ook aan afscheidsfotograaf. Besproken met kind?]

- **Eventuele obductie/donatie**
  [Besproken? Huidige wensen?]

- **Uitvaartverzekering**
  [Ja, welke? Nee, hoe wordt het dan geregeld?]

- **Uitvaartmaatschappij**
  [Is er een uitvaartmaatschappij betrokken? Besproken met gezin?]

- **Nazorg**
  [Besproken? Huidige wensen?]

5B VOCHT- EN/OF VOEDING ONTHOUDING

Zie voor aanbevelingen vocht en/of voeding onthouding in de laatste levensfase de Richtlijn vocht en voedingsonthouding.

- **Wensen van kind/ouders en doel m.b.t. (sonde)voeding**
  [Open tekstveld]

- **Wat is besproken/afgesproken omtrent het stoppen van vocht- en voeding aan het eind van het leven**?
  [Open tekstveld]

5C PALLIATIEVE SEDATIE

Wanneer er sprake is van een refractair symptoom, zie voor aanbevelingen palliatieve sedatie, stappenplannen en doseringsschema's de Richtlijn palliatieve sedatie.

- **Is er sprake van een refractair symptoom?**
  Kies een item.
- **Zo ja, van welk refractair symptoom is hier sprake?**
  [Open tekstveld]
- **Wat is besproken/afgesproken omtrent palliatieve sedatie?**
  [Open tekstveld]

Bovenstaande is besproken door [Open tekstveld] op [Klik of tik om een datum in te voeren.]
Voor starten behandeling afstemmen met [Open tekstveld]

B

Definitions of ROUGE score variants

- ROUGE-1

$$R_{\text{ROUGE-1}} = \frac{\textit{number of unigrams matches}}{\textit{total number of unigrams in the reference text}} \tag{B.1}$$

$$P_{\text{ROUGE-1}} = \frac{\textit{number of unigram matches}}{\textit{total number of unigrams in the generated text}} \tag{B.2}$$

- ROUGE-2

$$R_{\text{ROUGE-2}} = \frac{\textit{number of bigram matches}}{\textit{total number of bigrams in the reference text}} \tag{B.3}$$

$$P_{\text{ROUGE-2}} = \frac{\textit{number of bigram matches}}{\textit{total number of bigrams in the generated text}} \tag{B.4}$$

- ROUGE-Lsum

$$R_{\text{ROUGE-Lsum}} = \frac{\textit{longest common subsequence}}{\textit{total number of unigrams in the reference text}} \tag{B.5}$$

$$P_{\text{ROUGE-Lsum}} = \frac{\textit{longest common subsequence}}{\textit{total number of unigrams in the generated text}} \tag{B.6}$$

where:
The longest common subsequence (LCS) is defined as the longest sequence of words that appears in both texts, in the same order, but not necessarily contiguously.
Unlike ROUGE-L, which captures the LCS of the complete reference and generated texts, ROUGE-Lsum, first splits the generated and reference texts into individual sentences, after which the ROUGE-L score is computed for each sentence pair. The ROUGE-Lsum score is then obtained by averaging all ROUGE-L scores.

For all ROUGE variants, the F1 score is calculated as follows:

$$F1_{\text{ROUGE}} = 2 * \frac{R_{\text{ROUGE}} \cdot P_{\text{ROUGE}}}{R_{\text{ROUGE}} + P_{\text{ROUGE}}} \tag{B.7}$$

C

Average word count of generated texts across conditions

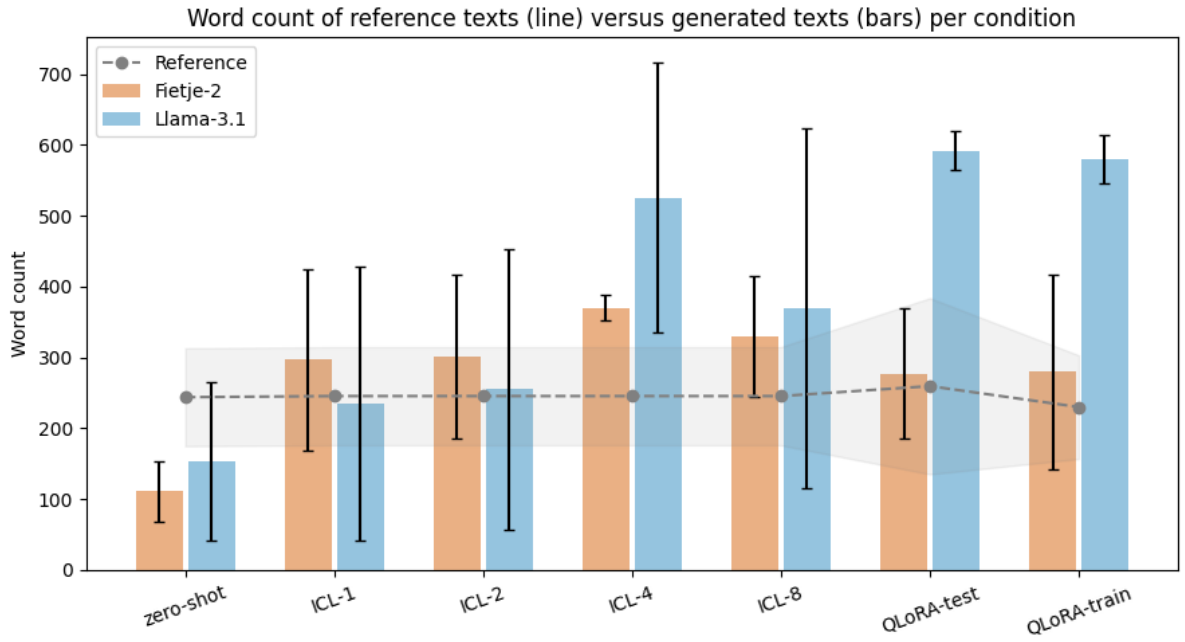


Figure C.1: Average word count of the generated texts by Llama-3.1 and Fietje-2, per fine-tuning condition, compared to the average word count in the corresponding reference texts. Error bars (generated texts) and shaded areas (reference texts) indicate standard deviations.

D

Distribution of MEDCON matches across conditions

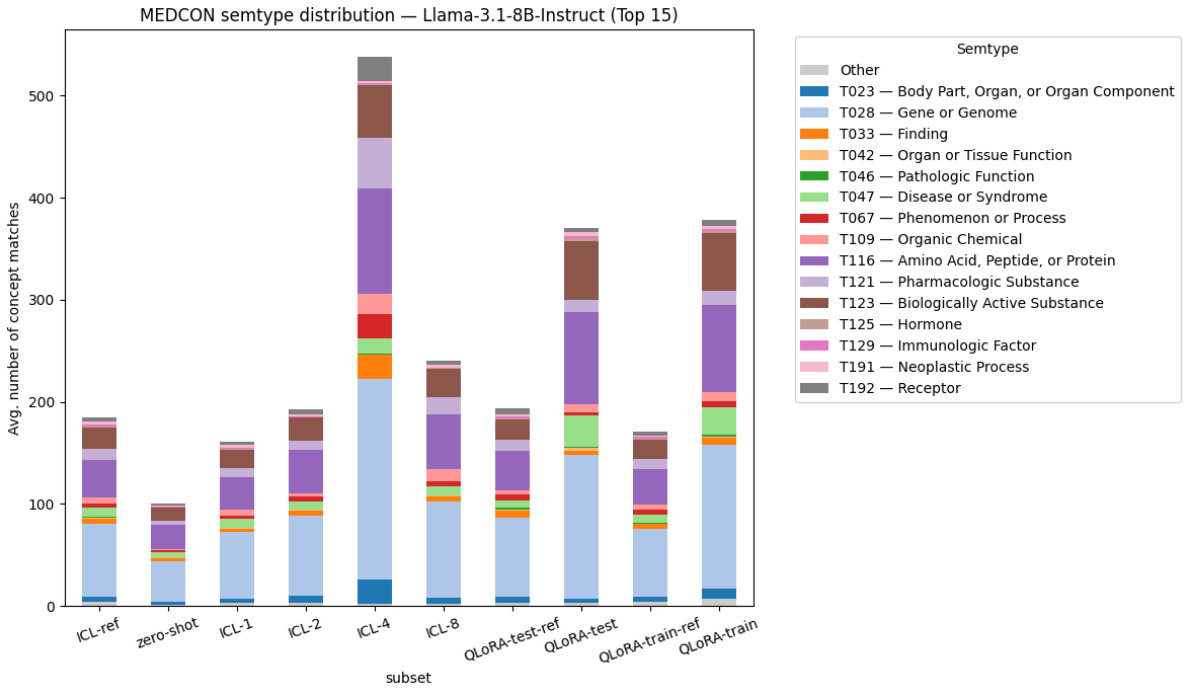


Figure D.1: The figure displays the average number of medical concept matches (semtypes) in the generated texts, compared to those in the reference texts, per condition for Llama-3.1.

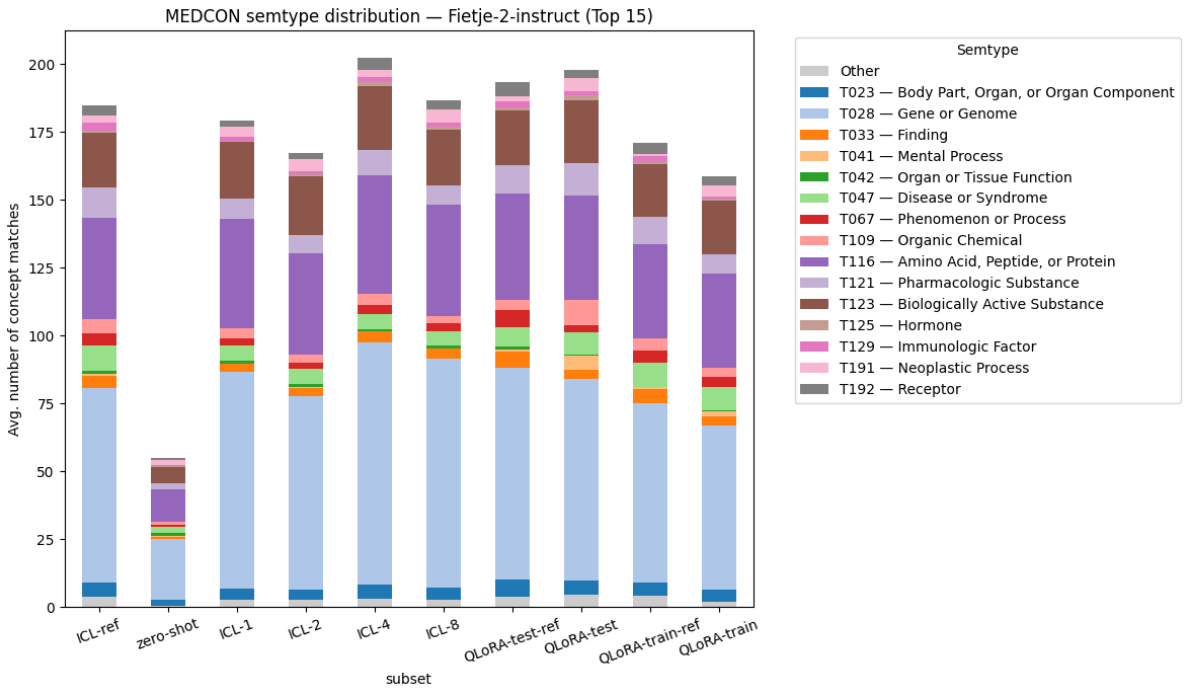


Figure D.2: The figure displays the average number of medical concept matches (semtypes) in the generated texts, compared to those in the reference texts, per condition for Fietje-2.

E

Training and validation losses during QloRA fine-tuning

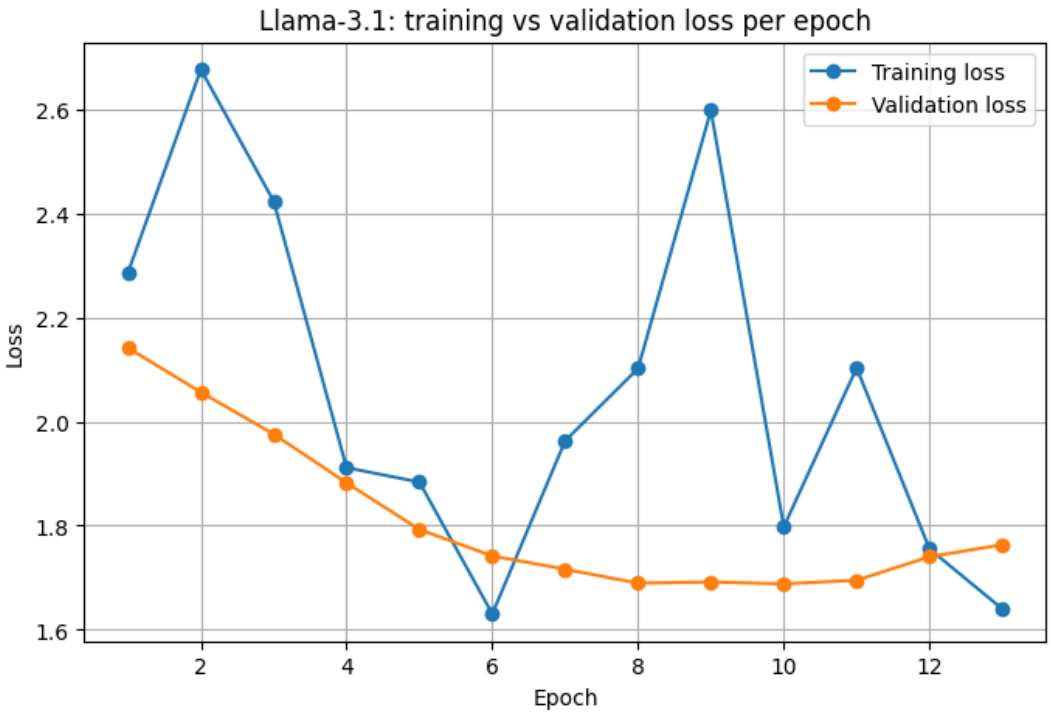


Figure E.1: Training and validation loss per epoch during QLoRA fine-tuning of Llama-3.1.

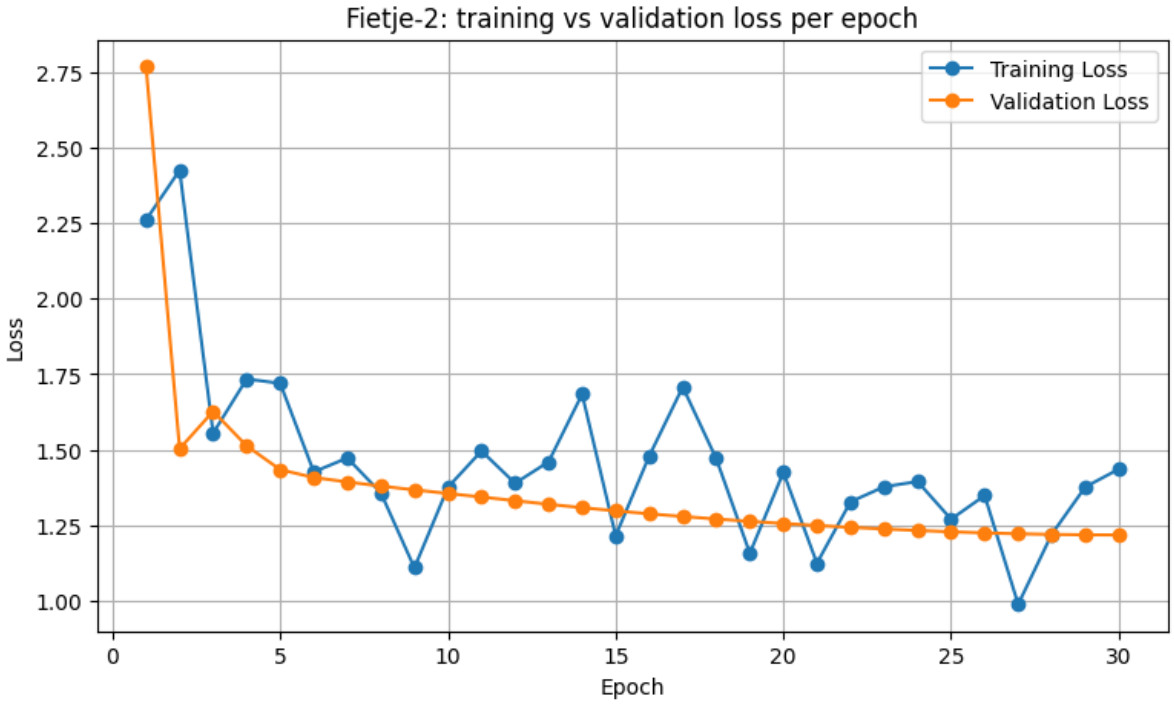


Figure E.2: Training and validation loss per epoch during QLoRA fine-tuning of Fietje-2.

F

Proof-of-Concept prompt reformulation

To examine whether the original formulation of the instruction in the prompt, consisting of three questions, contributed to the models' limited performance, we conducted a small proof-of-concept study. We rephrased the three questions into three instructions that directed the models more towards the desired answer and prompted the models with one ACP question per run, instead of with all three questions.

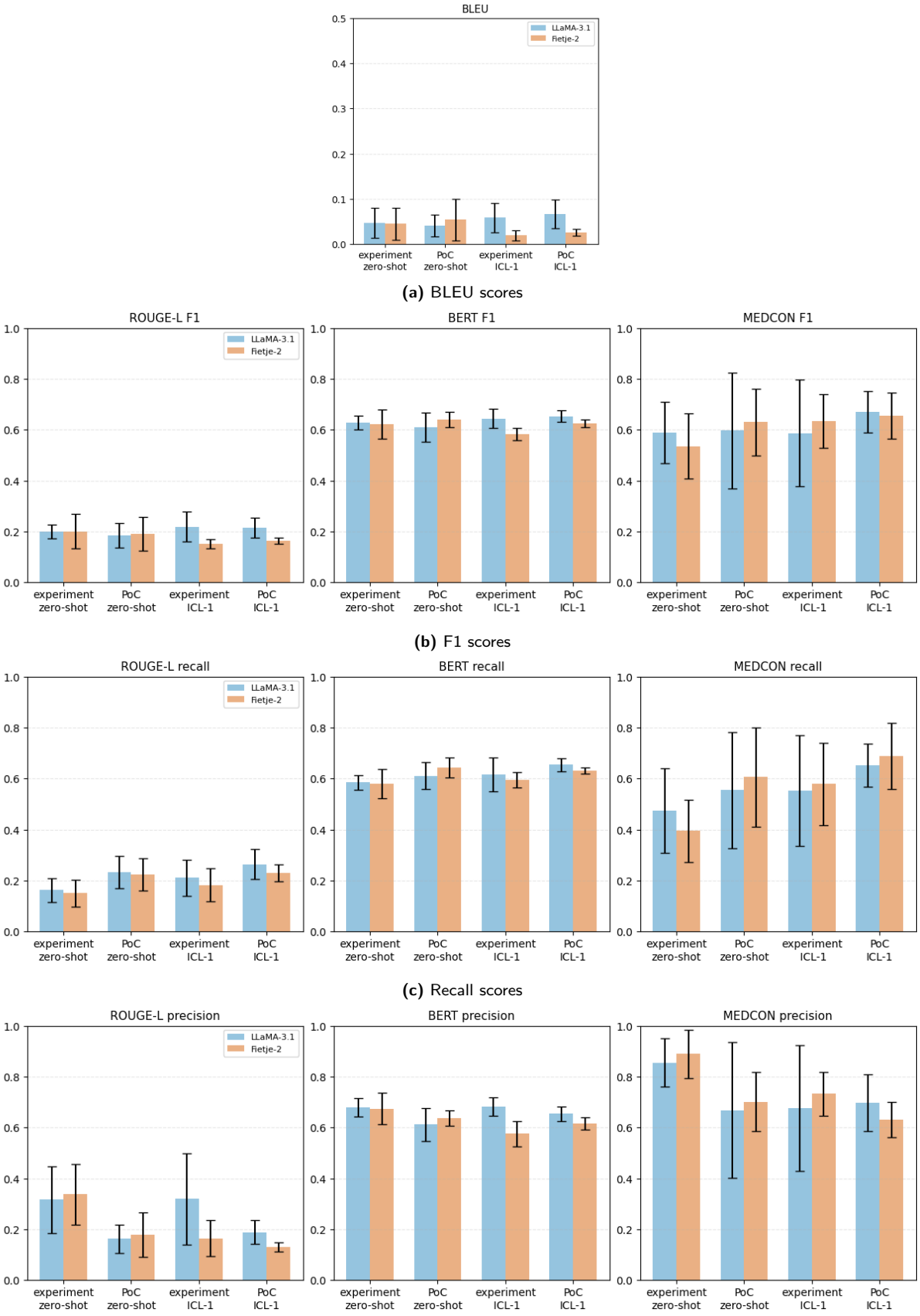
Instruction reformulation per ACP question:

1. Study the context and give a concise description of who the child and the family are, such as condition, hobbies, interests, organisation of daily care, and current points of care.
2. Study the context and give a concise description of what's of importance in the life of the child and their family.
3. Study the context and give a concise description of the goals, wishes and agreements for future care and treatment.

We performed zero-shot prompting and ICL-1 with both Llama-3.1 and Fietje-2 and compared the results to those obtained in the experimental evaluation (see Figure F.1).

We noticed no clear differences in BLEU and F1 scores between the two evaluations. Llama-3.1 shows a similar trend of improvement from zero-shot prompting to ICL-1 in the Proof-of-Concept compared to the experimental evaluation. Accordingly, the performance of Fietje-2 decreases from zero-shot prompting to ICL-1 in the Proof-of-Concept as well. The recall scores during zero-shot prompting in the Proof-of-Concept are generally slightly higher than the experimental evaluation, while the precision scores are slightly lower. During ICL-1, we see a similar trend, however, with even smaller differences between the results of the Proof-of-Concept and the experimental evaluation.

These findings indicate that a more directive instruction formulation in the prompt might slightly increase the amount of relevant information in the generated text. However, like in the experimental evaluation, this is accompanied by the addition of more irrelevant information. In addition, the overall model performance has not improved, suggesting that an explanation for the relatively low model performance lies elsewhere.



(a) BLEU scores

(b) F1 scores

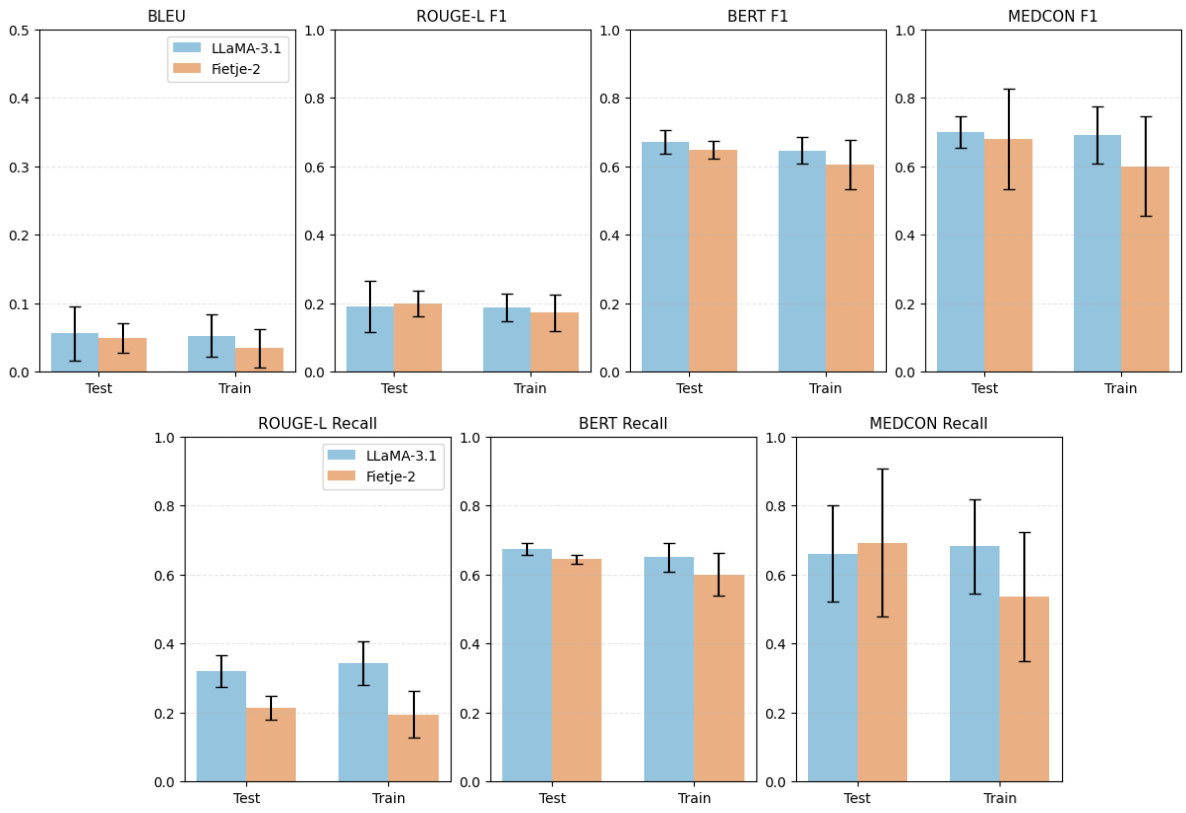
(c) Recall scores

(d) Precision scores

Figure F.1: Comparison of the average automatic metric scores of the experimental evaluation and the Proof-of-Concept study across zero-shot prompting and ICL-1. Error bars indicate standard deviations. (*Note that the BLEU y-axis limit is set to [0,0.5]*)

G

Comparison of automatic evaluation metrics results of the training and test sets after QLoRA fine-tuning



[figure continues on the next page]

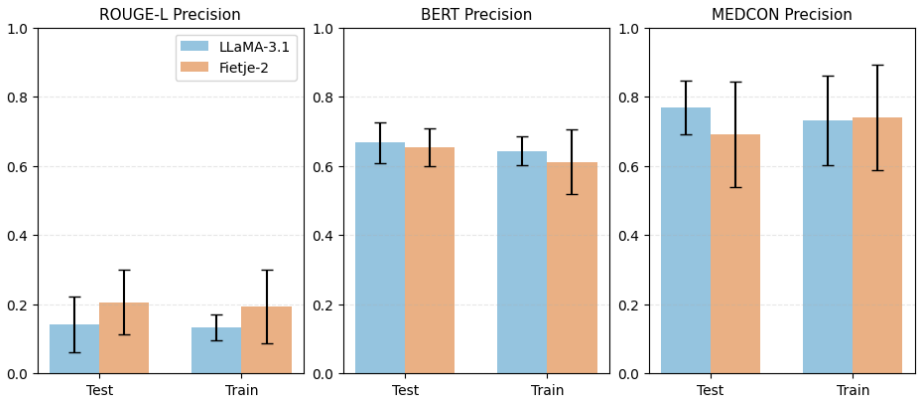


Figure G.1: Comparison of the automatic evaluation metrics results of the training and test set after QloRA fine-tuning. Error bars indicate standard deviations.
(*Note that the BLEU y-axis limit is set to [0,0.5]*)