universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Comparing Supervised Fine-Tuning and Direct Preference Optimization for Terminology-Aware English-to-French Medical Machine Translation

verfasst von | submitted by

Sandra Pörnbacher BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 587
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Multilingual Technologies
Betreut von | Supervisor:	Miguel Angel Rios Gaona PhD

Acknowledgements

I would like to express my gratitude to everyone who supported me throughout the process of writing this Master's thesis. First, I am deeply thankful to my supervisor, Mr. Angel Rios Gaona, PhD, for the continuous guidance and patience and support. I truly appreciate the time, feedback and encouragement you invested in helping me bring this work to completion. I also want to thank my family and friends for their never-ending support and motivation.

Abstract

Machine Translation (MT) in the medical domain carries risks that general-domain translation does not. A fluent but terminologically incorrect translation can mislead clinicians and patients, and such errors are easily overlooked because the surrounding text sounds natural. The named errors can even be fatal and underline the importance of studying and investigating this topic. This Master's thesis investigates whether Direct Preference Optimization (DPO) improves the accuracy of medical terminology translation in English-to-French MT beyond what Supervised Fine-Tuning (SFT) alone achieves. The work is empirical and comparative as it applies the same data and evaluation procedure to three multilingual decoder-only language models, namely TowerInstruct-7B, Ministral-8B-Instruct, and Qwen3-8B. Each model is first evaluated with a zero-shot approach as the baseline and then taken through SFT, a terminology-enriched SFT variant, and a terminology-aware DPO variant, so that any observed difference can be attributed to the specific step that was changed.

The models are fine-tuned with Parameter-Efficient Fine-Tuning (PEFT) methods (QLoRA) on the EMEA v3 English-French parallel corpus, which is reduced from roughly 1.09 million raw sentence pairs to about 287,000 through heuristic and semantic filtering and then down-sampled to 20,000 training, 1,000 validation, and 1,000 held-out test pairs. Additionally, the TICO-19 benchmark is reserved entirely as the main test set. Medical terms are extracted from the English source with a clinical named-entity recogniser and resolved to their French preferred terms through the Unified Medical Language System (UMLS) lookup, and these terms are injected as a glossary into the instruction prompt to form the terminology-enriched variant. For DPO, the supervised adapter is first merged into the base model and a fresh adapter is attached. Preference pairs are then built from eight hypotheses generated per source sentence and ranked with the preference strength selected through a sweep. Evaluation combines BLEU, chrF, and COMET-DA with a terminology-accuracy measure, paired bootstrap significance testing, and a manual inspection of system outputs.

The central finding is a clear negative result for DPO. Across all three models and both test sets, the terminology-aware preference-optimised system differs from the terminology-enriched supervised system by at most a few thousandths of COMET, and the choice of preference strength has little effect. For Qwen, the supervised and preference-optimised systems frequently produce really similar outputs. SFT

has a larger effect. In comparison to the zero-shot baseline it improves COMET for Tower and Qwen on the EMEA test set but falls below the baseline for all models on the TICO-19 test set. Terminology enrichment changes the scores inconsistently across both test sets and does not return a reliable improvement. The surface-level metrics and the semantic metric differ, which argues for reporting a nerual metric as the primary measurement.

The thesis contributes reproducible zero-shot, SFT and DPO systems for medical English-to-French translation, as well as the methodological steps required to train DPO stably on parameter-efficient models. Its broader implication is that the choice of base model and the strength of the base model matter more for terminological accuracy than the preference-optimisation stage studied here, and that reliable term adherence will likely require mechanisms that enforce, rather than only encourage, the use of correct terms.

Zusammenfassung

Die maschinelle Übersetzung im medizinischen Bereich birgt Risiken, die bei Übersetzungen in allgemeinen Bereichen nicht bestehen. Eine flüssige, aber terminologisch falsche Übersetzung kann medizinischem Personal und Patient:innen falsche Informationen vermitteln, und solche Fehler werden leicht übersehen, weil der gesamte Text durch die natürliche Sprache korrekt wirkt. Die genannten Fehler können sogar bis zum Tod führen und unterstreichen, wie wichtig es ist, dieses Thema zu untersuchen und zu erforschen. Die vorliegende Arbeit untersucht, ob Direct Preference Optimization (DPO) die Genauigkeit der medizinischen Terminologieübersetzung bei der maschinellen Übersetzung vom Englischen ins Französische im Vergleich zu Supervised Fine-Tuning (SFT) allein verbessert. Der Ansatz ist empirisch und vergleichend angelegt, da dieselben Daten und dasselbe Bewertungsverfahren auf drei mehrsprachige Decoder-only-Sprachmodelle angewendet werden, nämlich TowerInstruct-7B, Ministral-8B-Instruct und Qwen3-8B. Jedes Modell wird zunächst mit einem Zero-Shot-Ansatz als Baseline evaluiert und durchläuft anschließend SFT, eine terminologie-angereicherte SFT-Variante und eine terminologie-gestützte DPO-Variante, sodass beobachtete Unterschiede dem jeweils veränderten Schritt zugeschrieben werden können.

Die Modelle werden mitParameter-Efficient Fine-Tuning (PEFT)-Methoden (QLoRA) auf dem englisch-französischen Parallelkorpus EMEA v3 feinabgestimmt. Dieses Korpus wird durch heuristische und semantische Filter von rund 1,09 Millionen Satzpaaren auf etwa 287.000 reduziert und anschließend in 20.000 Trainings-, 1.000 Validierungs- und 1.000 Testpaare aufgeteilt. Zusätzlich ist der TICO-19-Benchmark vollständig als zentraler Testdatensatz reserviert. Medizinische Begriffe werden aus dem englischen Text mit einem klinischen Named-Entity Recogniser extrahiert und führen über eine Abfrage im Unified Medical Language System (UMLS) zur bevorzugten französischen Übersetzung. Diese Begriffe werden als Glossar in den Instruction Prompt eingefügt und bilden so die mit Terminologie angereicherte SFT Variante. Für DPO wird zunächst der Adapter in das Basismodell integriert und ein neuer Adapter angehängt. Anschließend werden Präferenzpaare aus acht generierten Hypothesen gebildet und gereiht, wobei die Präferenzstärke über einen Sweep ausgewählt wird. Die Evaluierung kombiniert BLEU, chrF und COMET-DA mit einer Berechnung der Terminologiegenauigkeit, gepaarten Bootstrap-Signifikanztests und einer manuellen Inspektion der Systemausgaben.

Das zentrale Ergebnis ist ein deutlich negatives Resultat für DPO. Über alle drei Modelle und beide Testdatensätze hinweg unterscheidet sich das terminologie-bewusste, DPO System vom terminologie-angereicherten SFT System höchstens um wenige Tausendstel für COMET-DA, und die Wahl der Präferenzstärke hat wenig Einfluss. Bei Qwen erzeugen das überwachte und das präferenzoptimierte System häufig sehr ähnliche Ausgaben. SFT hat einen größeren Effekt. Im Vergleich zur Zero-Shot-Baseline verbessert es COMET-DA für Tower und Qwen auf dem EMEA-Testset, sinkt jedoch für alle Modelle auf dem TICO-19-Testset unter die Werte der Baseline. Terminologieanreicherung verändert die Werte über beide Testsets hinweg inkonsistent und liefert keine verlässliche Verbesserung. BLEU und chrF unterscheiden sich von COMET-DA in ihrem Verhalten, was dafür spricht, eine neuronale Metrik für die primäre Messung zu verwenden.

Die Arbeit präsentiert reproduzierbare Zero-Shot-, SFT- und DPO-Systeme für die medizinische Englisch-Französisch-Übersetzung sowie die methodischen Schritte, die erforderlich sind, um DPO stabil auf parameter-effizienten Modellen zu trainieren. Die weitergehende Folge ist, dass die Wahl des Basismodells und dessen Stärke für die terminologische Genauigkeit wichtiger sind als die hier untersuchte Präferenzoptimierungsphase und dass verlässliche Genauigkeit in bezug auf die korrekt Terminologieübersetzung wahrscheinlich Mechanismen erfordern wird, die die Verwendung korrekter Begriffe erzwingen, statt sie nur zu suggerieren.

Contents

List of Tables **9**

List of Figures **11**

1 Introduction **1**
1.1 Research Question and Assumptions 2
1.2 Research Motivation and Objectives 3

2 Background and Related Work **5**
2.1 Large Language Models 5
2.1.1 Transformer Architecture 7
2.1.2 Pre-Training and Transfer Learning 9
2.1.3 Parameter-Efficient Fine-Tuning and LoRA 12
2.2 Supervised Fine-Tuning 13
2.3 Direct Preference Optimization 16
2.3.1 Reinforcement Learning and Reinforcement Learning from Human Feedback 17
2.3.2 Theoretical Foundation 18
2.3.3 Training Process 20
2.3.4 Advantages and Limitations 22
2.4 Machine Translation 23
2.4.1 Neural Machine Translation 24
2.4.2 LLM-Based Machine Translation 25
2.4.3 Evaluation Metrics 25
2.5 Medical Machine Translation 27
2.5.1 Terminology 27
2.5.2 Characteristics of Medical Language 28
2.5.3 Datasets and Benchmarks 30
2.6 Related Work 31

3 Methodology **37**
3.1 Pipeline 38
3.2 Dataset Description and Preprocessing 40
3.2.1 Terminology Annotation and Extraction 41

3.2.2 Terminology-Aware Instruction Data . . . 43
3.3 Model Selection . . . 43
3.4 Experimental Setup . . . 44
3.4.1 SFT Configuration . . . 45
3.4.2 DPO Configuration . . . 46
3.5 Evaluation Framework . . . 48

4 Results **51**
4.1 Automatic Evaluation Results . . . 51
4.2 Human Evaluation Results . . . 55
4.3 Error Analysis . . . 56
4.4 Training Efficiency Comparison . . . 58

5 Discussion **59**
5.1 SFT vs. DPO: Comparative Analysis . . . 59
5.2 Implications for Medical MT . . . 60
5.3 Limitations . . . 61
5.4 Future work . . . 62

6 Conclusion **65**
6.1 Key Findings . . . 66
6.2 Contributions . . . 66

Bibliography **69**

List of Tables

1 English-to-French medical term examples from the UMLS extraction used in this thesis. For each English source term, the middle column lists two of the French synonyms retrieved from UMLS, and the right column gives a fluent French phrase that does not appear in the retrieved synonym set. A candidate translation is credited for the term if it contains any retrieved synonym; thus *crise cardiaque* counts as correct for *myocardial infarction* because it is itself a retrieved synonym, whereas *dépassement de la posologie* does not count for *overdose*. 30

2 Results of remaining sentence pairs after applying heuristic and semantic filter to the EMEA v3 (Tiedemann, 2012) corpus. 41

3 Final SFT configuration. Architecture settings (QLoRA 4-bit NF4, rank 8, scaling 64, dropout 0.05, max. length 512, effective batch 32, 3 epochs, cosine schedule) are shared across all models. The learning rate, warmup, and adapted modules are adjusted per model. 46

4 A preference pair whose chosen and rejected candidates differ in the translation of the medical term *HIV infection*. The chosen candidate produces the correct French term *infection par le VIH*, whereas the rejected candidate fails to translate it, producing the incorrect form *infectie-virales*. MBR score gap 0.15. 48

5 Best configuration per model from the Optuna search (rank = 8, scaling = 64 in all cases). Validation loss is the minimised objective. 51

6 Supervised models validation results at the selected checkpoint per model. 52

7 Tower supervised models on the split validation subsets. 52

8 Tower supervised model and preference-optimised models across β values. 53

9 Final held-out test results on the EMEA test set. 53

10 Final held-out test results on the TICO-19 benchmark. 54

11 Terminology accuracy on the terminology-containing test sentences. N is the number of evaluated sentences. 55

12 A terminology substitution rewarded by COMET-DA but penalised by BLEU (Qwen terminology DPO, EMEA test set). Segment BLEU = 20.33 (corpus mean 50.14), segment COMET-DA = 0.931 (corpus mean 0.869). The injected glossary term was *diabetes mellitus* → *diabète sucré.* 57

13 Case in which the UMLS lookup injected an incorrect French term, so that a correct model output is recorded as a terminology miss. . 58

List of Figures

1 The Transformer encoder-decoder architecture, showing multi-head self-attention, position-wise feed-forward layers, residual connections, and positional encodings. (Vaswani et al., 2017, 3) 8

2 Decoder-only LLM architecture, illustrated with LLaMA, (a) decoder stack, (b) auto-regressive generation and (c) common PEFT operations (Han et al., 2024, 3). 9

3 Comparison of RLHF and DPO. (Rafailov et al., 2023, 2) 18

4 The different degrees of formalization from unstructured textual content to ontologies (Navigli, 2022, 519) 28

5 Methodology pipeline for terminology-aware medical MT. The EMEA corpus is filtered and split, terminology is injected into the prompts, and each model is first fine-tuned wiht SFT and then optimised with DPO on preference pairs. All systems are evaluated on the EMEA test set and the TICO-19 benchmark. The final outcome includes three variants per model, namely the SFT baseline, the term-enriched SFT variant and the terminology-aware DPO variant. 39

6 Illustration of the terminology annotation and extraction pipeline . 42

1 Introduction

Medical Machine Translation (MT) serves as an important bridge for communication between patients and clinicians in multilingual settings, where translation errors can have serious consequences for patient safety and the overall quality of healthcare (Mehandru et al., 2022). Clinicians often face barriers such as limited time and resources, cultural differences and varying levels of medical vocabulary that result in challenges for them and patients (Flores, 2005). These factors make reliable translation essential in medical settings.

One of the main challenges in medical MT is that current general-domain MT systems often fail to meet the specific and strict requirements of medical applications (Gaona et al., 2023). Neural Machine Translation (NMT) models can produce grammatically correct and stylistically appropriate text that hides serious semantic errors. Such errors are therefore easily overlooked because the output appears professionally written and linguistically natural and is especially hard to distinguish for non-specialists. Terminological errors are particularly concerning. Medical terminology is governed by international standardisation efforts and authoritative resources such as the Systematized Nomenclature of Medicine (SNOMED) (Bos and Donnelly, 2006), the International Classification of Diseases (ICD) (Fritz, 2000), and the Medical Subject Headings (MeSH) (Lipscomb, 2000). These vocabularies are brought together in the Unified Medical Language System (UMLS) (Bodenreider, 2004), a metathesaurus maintained by the US National Library of Medicine that integrates preferred terms, synonyms, and concept relations across a wide range of biomedical vocabularies.

Prior work has addressed these challenges with different approaches. Instruction-tuned Large Language Model (LLM)s that integrate medical terminology databases directly into training prompts can reach much higher terminological accuracy compared to standard fine-tuning (Rios, 2025). Comparative evaluations of commercial translation platforms have provided insights into current capabilities and their limitations for clinical use (Sebo and de Lucia, 2024). Approaches based on preference optimization have also been applied to MT more broadly, showing that preference-based training can produce consistent improvements in translation quality across multiple language pairs (Yang et al., 2024). However, it remains an open question whether preference optimization specifically improves medical terminology translation accuracy beyond what Supervised Fine-Tuning (SFT) alone achieves.

The challenge of rendering English medical terms with formally correct French equivalents has been documented in terminology translation research (Deléger et al., 2010). Even semi-automatic methods combining knowledge-based and corpus-based approaches achieve valid translation for only half of a target vocabulary. The other half requires manual intervention by medical experts. This underlines the difficulty of achieving formal terminology accuracy in English-to-French medical translation automatically.

This Master's thesis addresses that issue by investigating whether Direct Preference Optimization (DPO) (Rafailov et al., 2023) improves medical terminology translation accuracy in English-to-French MT compared to SFT and its underlying zero-shot baseline. The approach uses Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) of the multilingual pre-trained LLMs TowerInstruct-7B-v0.2 (Rei et al., 2025; Unbabel, 2024), Qwen3-8B (Yang et al., 2025; Qwen Team, 2025) and Ministral-8B-Instruct-2410 (Liu et al., 2026; Mistral AI, 2024) on the European Medicines Agency (EMEA) v3 EN-FR parallel corpus (Tiedemann, 2012). For each source sentence, medical terms detected with scispaCy (Neumann et al., 2019) are looked up in UMLS to retrieve their French synonym terms, which are added as a glossary to the instruction prompt so the model learns to follow terminology constraints. Preference pairs are then constructed by generating eight translation hypotheses per source sentence and ranking them by expected French UMLS terms they contain, so that the hypothesis with the correct terminology is chosen and one missing those terms is rejected. The preference-optimised models are trained with this preference data and evaluated alongside the SFT variants on the Translation Initiative for COVID-19 (TICO-19) benchmark (Anastasopoulos et al., 2020) and the EMEA test set using the automatic metrics BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Character n-gram F-score (chrF) (Popović, 2015), and Crosslingual Optimized Metric for Evaluation of Translation (COMET-DA) (Rei et al., 2020).

1.1 Research Question and Assumptions

Given the challenges in medical MT outlined above, the following research question guides this investigation.

> To what extent does DPO improve the quality of medical terminology translation compared to SFT when evaluated using automatic metrics and terminological accuracy?

This thesis rests on two main assumptions. First, ranking translation hypotheses according to whether they contain the expected French medical terms derived

from UMLS using the EMEA v3 EN-FR parallel corpus (Tiedemann, 2012) and UMLS as reference sources, produces preference pairs that offer a stronger learning signal than general quality-based translation ranking. Second, PEFT / LoRA fine-tuning of Tower (Rei et al., 2025; Unbabel, 2024), Qwen (Yang et al., 2025; Qwen Team, 2025) and Ministral (Liu et al., 2026; Mistral AI, 2024) on medical parallel data provides a suitable SFT system for measuring the effect of subsequent DPO training.

1.2 Research Motivation and Objectives

Medical MT presents unique challenges that distinguish it from general-domain translation. In healthcare contexts, translation errors can have serious consequences for patient safety and the overall quality of healthcare delivery (Mehandru et al., 2022). The focus of this research on terminology is motivated by the fact that terminological errors pose high risks in clinical settings and are often overlooked by conventional automatic evaluation metrics (Gaona et al., 2023). Additionally, this affects a vast amount of people since most of them are not specialised in the medical field or do not have enough domain-related knowledge to spot errors. This concern becomes even more important as LLMs are used by the general public, with people often turning to LLM outputs as a first source of medical information instead of consulting a medical professional for various reasons such as cost or limited access. In this context, an LLM that generates output which is incorrect regarding medical questions can be as dangerous for the people relying on it as if it was used in a medical setting (Jurafsky and Martin, 2026). Approaches that simply maximise the likelihood of reference translations may not adequately capture the nuanced requirements of medical terminology (Mehandru et al., 2022), and DPO may offer a complementary training signal by incorporating preference comparisons that encode formally correct terminology over informal or imprecise alternatives (Rafailov et al., 2023; Sun et al., 2025).

This thesis therefore makes the following contributions. First, a zero-shot baseline together with two SFT variants for English-to-French medical translation is established using the EMEA v3 EN-FR parallel dataset (Tiedemann, 2012) with UMLS-derived French terminology injected into the instruction prompts for the multilingual LLMs Tower, Ministral and Qwen. Second, a DPO training method will be developed using a medical terminology resource to construct preference pairs that encode formal medical terminology over informal descriptions. Third, it systematically compares the zero-shot baseline, SFT and DPO across the three models. For the DPO stage the preference-strength parameter β is treated as the main hyperparameter. Fourth, the evaluation combines the automatic metrics BLEU, chrF and COMET-DA with a terminological-accuracy measure, followed

by paired bootstrap significance testing.

In summary, the named contributions clarify under which conditions DPO improves medical terminological accuracy over the SFT systems and on which metrics any effect appears. The contributions also show where DPO produces no improvement or even degrades the overall translation quality in comparison to the zero-shot baseline. The full results and details, as well as their analysis are presented in Chapter 4 and Chapter 5.

2 Background and Related Work

This chapter provides the background necessary to understand the methods and experiments described in subsequent chapters, and reviews the work most directly relevant to the research question. Section 2.1 covers the transformer architecture and the pre-training and transfer learning paradigms that underlie the models used in this thesis. Section 2.2 introduces SFT forming the baseline of the experiments and Section 2.3 introduces DPO, another training paradigm, including its theoretical foundations and its relationship to Reinforcement Learning from Human Feedback (RLHF). Section 2.4 introduces NMT and the automatic evaluation metrics used to assess translation quality. Section 2.5 describes the specific challenges of medical MT, the datasets and benchmarks relevant to this work, and prior training approaches in this domain. Section 2.6 reviews studies directly relevant to the research question, covering preference optimization for MT and fine-tuning approaches for medical terminology translation.

2.1 Large Language Models

This section provides the architectural background for the models used in this thesis. The transformer architecture underlies Tower(Rei et al., 2025; Unbabel, 2024), Qwen (Yang et al., 2025; Qwen Team, 2025) and Ministral (Liu et al., 2026; Mistral AI, 2024) which are fine-tuned on medical open source data, as well as the evaluation metrics that are based on contextual embeddings. Pre-training and transfer learning explain how general-purpose language models are adapted to specialized domains such as medical translation.

LLMs are a specific kind of Deep Learning (DL) models, because they are deep Neural Network (NN)s and most modern variants are built on the transformer architecture further explained in Subsection 2.1.1. DL is a subset of Machine Learning (ML) in which computational models composed of multiple processing layers learn to represent data at increasing levels of abstraction (LeCun et al., 2015). By capturing subtle patterns, DL makes it possible for the LLM to generate fluent text. These layers of features are not designed by engineers but learned directly from data through the general-purpose procedure of backpropagation, which adjusts the model's internal parameters to minimise prediction error (LeCun et al., 1998). The availability of GPU computation and improved activation functions has made

it possible to train much deeper networks than before, and DL now forms the foundation on which LLMs rely.

LLMs build on a long line of neural sequence models. Early Recurrent Neural Network (RNN)s trained with backpropagation through time struggled to learn long-range dependencies, because the error signal tends to vanish or explode as it is propagated across many time steps (Hochreiter and Schmidhuber, 1997). Long Short-Term Memory (LSTM) networks addressed this by pushing information through a gated memory cell with a self-connection of weight, which allowed the gradient to flow over hundreds of steps and made LSTM the dominant architecture for sequence modelling for nearly two decades (Hochreiter and Schmidhuber, 1997). The transformer architecture later replaced sequential recurrence with parallelisable self-attention. It enabled the scale at which today's LLMs, including Tower, are trained (Vaswani et al., 2017).

As the name suggests, an LLM is a language model trained at very large scale. Its main training task is next-token prediction. It contains context windows of up to thousands of tokens- The model repeatedly predicts the most likely next token, and through this self-supervised objective it learns meaning, context and a broad base of linguistic and world knowledge during pre-training (Jurafsky and Martin, 2026). After pre-training, the model can be adapted to downstream tasks through prompting or fine-tuning without the need of a specific architecture such as MT. LLMs are part of the broader generative Artificial Intelligence (AI) category which also includes image generation, for example. General LLMs can be distinguished by their architectures which are either decoder, encoder or encoder-decoder (Jurafsky and Martin, 2026). The models used in this thesis, Tower (Rei et al., 2025; Unbabel, 2024), Ministral (Liu et al., 2026; Mistral AI, 2024) and Qwen (Yang et al., 2025; Qwen Team, 2025), are decoder-only models. They take an input series of tokens and generate one output token at a time working in a left to right flow where the model only predicts from previous tokens. Whereas encoder-only architectures return vector representations of each token and are not part of the generative models, but rather used for classification. The encoder-decoder framework maps an input sequence to an output sequence of tokens (Jurafsky and Martin, 2026). Since the models used in the experiments of this thesis are decoder-only models, references to language models in the following chapters and section should be understood as referring to decoder language models.

Prompting is a way to interact with a language model using a text string. The user input conditions the generation by providing a question or instruction. The prompt can also include additional data such as demonstrations, which are labelled examples drawn from the training set;. This procedure is known as few-shot prompting, a form of in-context learning (Jurafsky and Martin, 2026). Beyond prompting , the model itself can be adapted through instruction tuning, in which

it is fine-tuned on instruction-response pairs. Rios (2025) shows that instruction tuning is effective on medical data. Similarly, Zaid et al. (2025) fine-tune Qwen-1.5B with LoRA adapters for multilingual clinical dialogue summarisation and structured information extraction, producing English summaries, schema-aligned JSON outputs and multilingual question-answering responses, while improving fluency and factual accuracy under limited GPU resources. However, prompting and instruction tuning are only useful if the underlying model is already robust enough and pre-trained on high-quality and a large amount of data. The capabilities of the named fine-tuning procedures build directly on the capabilities of the underlying model. Most modern LLMs today are build on the transformer architecture, explained in detail in the following subsection.

2.1.1 Transformer Architecture

The transformer architecture has established itself as the most commonly used in comparison to RNNs (Jurafsky and Martin, 2026). The revolutionary work of Vaswani et al. (2017) provided a much simpler, faster and therefore also cheaper architecture which is used nearly universally by modern LLMs and illustrated in Figure 1.

The Transformer is a NN that does not use RNNs and CNNs, but instead uses multihead attention. With the transformer Vaswani et al. (2017) provide some key innovations revolutionising the state-of-the-art model architectures. Starting with multi-head attention, a process that runs all tokens in parallel and is able to capture long-distance dependencies. As the name already implies, multi-head attention does not use a single attention computation but multiple attention heads that operate in parallel. Each head has its own query, key and value weight matrices, which the model uses to obtain different aspects of the syntactic, semantic or positional relationship for the context (Jurafsky and Martin, 2026). Continuing with the transformer blocks. Each block has two sublayers with a residual connection and a normalisation layer. In detail, each block applies layer normalisation and multi-head attention and afterwards adds the result back to the input through residual connection. The feed-forward sublayer consists of two layers with a Recitified Linear Units (ReLU) activation (Agarap, 2018). Through this design the original token information is always present and stays throughout the layers. Additionally, the model is able to know the sequence of tokens even though they are not processed in order. This is due to postional encoding, since self-attention has no built-in notion of token order. The positional encoding is added to each input embedding. Recent models often use relative positonal encodings that are integrated directly in to the attention mechanism at eahc layer (Jurafsky and Martin, 2026). Lastly, the transformer includes a language model head. This is the output of the final layer which is passed through an unembedding matrix and softmax function to

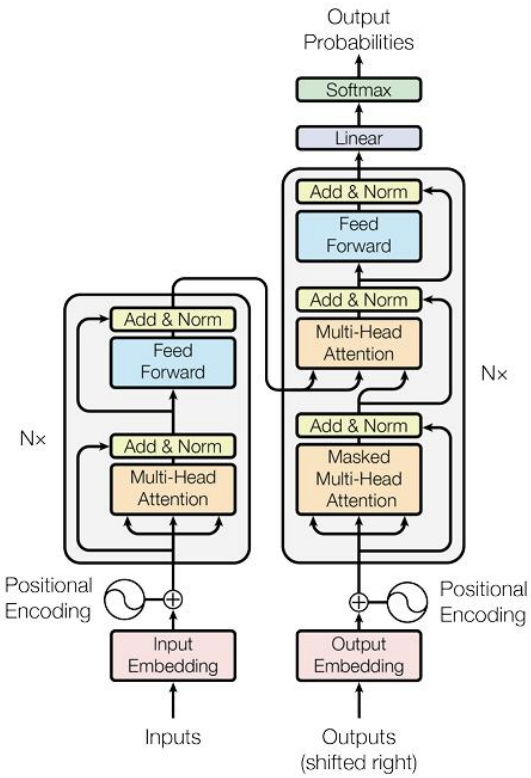

Figure 1: The Transformer encoder-decoder architecture, showing multi-head self-attention, position-wise feed-forward layers, residual connections, and positional encodings. (Vaswani et al., 2017, 3)

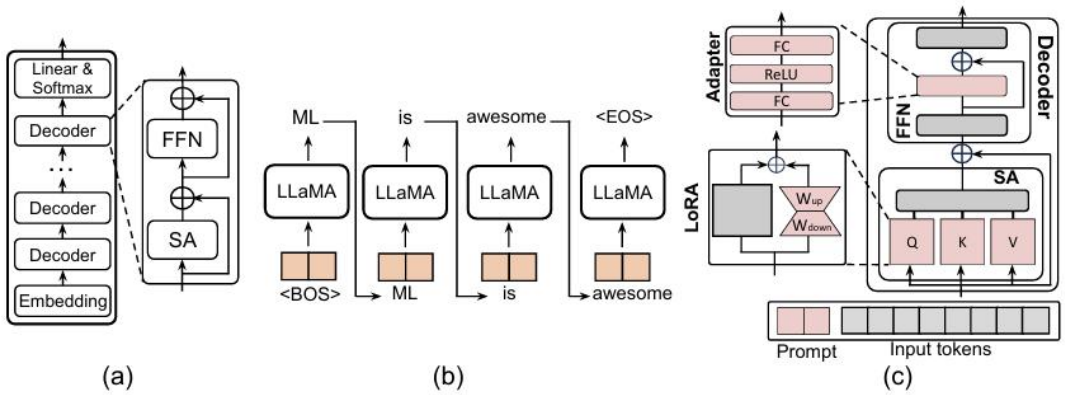

Figure 2: Decoder-only LLM architecture, illustrated with LLaMA, (a) decoder stack, (b) auto-regressive generation and (c) common PEFT operations (Han et al., 2024, 3).

produce a probability distribution over the vocabulary for the next token. Usually, the unembedding matrix is implemented as the contrary of the input embedding matrix. This technique is know as weight tying (Jurafsky and Martin, 2026).

Modern LLMs, including the ones used in the experiments of this Master's thesis, are based on the transformer architecture. They are though not based on the full encoder-decoder architecture but rather on the decoder-only variant, shown in Figure 2, together with additional design choices, which are explained in detail in Chapter 3.

2.1.2 Pre-Training and Transfer Learning

Pre-training is the first step of training a language model. A LLM is given a large amount of text in a corpus and performs next-token prediction. The process starts with random guessing but it learns from previous guesses and can therefore minimise the error over time. It can improve because the answer is always compared to the correct token in the actual corpus before the model continues with the next prediction. The error is usually computed using cross-entropy loss, after which internal weights are adjusted to obtain better results next time. This iterative process continues for millions of tokens and more (Jurafsky and Martin, 2026).

The data used for pre-training is self-supervised. This means the correct answer already exists in the text itself. As mentioned above, the next token to be predicted is already present in the corpus. Training corpora are often web-scraped and filtered content (Jurafsky and Martin, 2026). Through this process, the model acquires both a general understanding of natural language structure and a broad base of

world knowledge contained in the training data. The amount and variety of text that the model is exposed to during pre-training is determative for its. To increase the quality of the model output further, instruction-tuning, fine-tuning or other methods can be applied.

Pre-training and transfer learning are two closely connected concepts. In transfer learning, a model first is trained on a source domain that is usually a more general one or at least one that includes a large amount of data. Then the resulting weights are reused as a starting point for training on a target domain. Data in target domains is often scarce or hardly available (Neyshabur et al., 2020). The pre-trained weights remain preserved and are used as an informed initialisation that already captures useful language properties. Research shows that two main factors contribute to successful transfer. First, the reuse of features that have been learned during pre-training. And second, the low-level statistical properties of the data that the model has already internalised. Both help to get faster convergence and a better final performance in a specific domain or task compared to training from random initialisation (Neyshabur et al., 2020). Models that are fine-tuned from pre-trained weights remain in the same scale of the loss as the original pre-trained model. This helps to explain why fine-tuning typically does not erase knowledge that is learned during pre-training.

Medical MT is a specialised domain, where the language terminology differs a lot from general text used during pre-training. In such a case, transfer learning is often combined with domain adaption. A common approach is a mixed-domain pre-training that is also called continued pre-training. A model that has been pre-trained on a general-domain corpus is then trained on a domain-specific dataset. Through this process the model can keep the general language understanding and in the same time it can shift the internal representations in the direction of the target domain. Continued pre-training on domain-specific data can noticeably improve the models output in comparison to general pre-training. At the same time, the similarity between the source and target domain affect how effective transfer learning is. Neyshabur et al. (2020) observe that when the target domain is very different from the source in context of its linguistic style, the benefits of feature reuse decrease. But at the same time low-level statistical benefits and faster optimisation often remain present.

The general framework of transfer learning addresses in above is also part of modern Natural Language Processing (NLP). In modern NLP pre-training of language models on unlabelled data and transferring its parameters to downstream task is the dominant paradigm. Devlin et al. (2019) show that a single set of pre-trained weights can be used across multiple tasks with only minimal changes for each specific task. They provide this idea as a top-step process where transformer-based models are firstly pre-trained followed by fine-tuning. They have a specific

setup, where the pre-trained model is first initialised with weights learned form large unlabelled corpora. This happens through self-supervised objectives, like masked language modelling, and afterwards the model is fine-tuned with labelled data for each downstream task. This is essentially the NLP counterpart of the feature-reuse mechanism described by Neyshabur et al. (2020), and it confirms that initialising from pre-trained weights consistently outperforms training from random initialisation, particularly when in-domain data is limited (Devlin et al., 2019).

This is the paradigm followed in the present thesis. Tower has already been pre-trained on large-scale multilingual text, and the experiments described in later chapters do not train a new model from scratch but instead fine-tune these existing weights on medical English-to-French data. The pre-trained model is therefore the source of both the general linguistic knowledge and the translation capabilities that the fine-tuning step is meant to specialise.

Shi et al. (2025) refine this by treating domain adaption as part of a continual learning pipeline for LLMs. They describe modern LLM training as a sequence of stages along a vertical axis of continuity. This implies moving from continual pre-training on general data, through domain-adaptive pre-training on unlabelled domain-specific data, to continual fine-tuning on final downstream task. In this framework, models such as Tower being a general-purpose translation model, can be adapted to the medical domain A more recent line of work refines this picture by treating domain adaptation as part of a continual learning pipeline for LLMs. Shi et al. (2025) describe modern LLM training as a sequence of stages along a vertical axis of continuity, moving from continual pre-training on general data, through domain-adaptive pre-training on unlabelled domain-specific data, to continual fine-tuning on the final downstream task. In this setup, adapting a general translation model like Tower to the medical domain happens in the later stages of the pipeline, where the aim shifts from broad language learning to specialising the model for a much narrower, domain-specific dataset.

A key challenge identified in this literature is catastrophic forgetting. This is the tendency of a model to lose previously acquired knowledge when its parameters are updated on a new distribution. Shi et al. (2025) note that pre-trained LLMs are relatively robust to forgetting at the representation level, in line with the observation by Neyshabur et al. (2020) that fine-tuned models tend to remain in the same loss basin as their pre-trained starting point. Even so, fine-tuning can still erase specific task details and shift the model's behaviour away from its original distribution (Shi et al., 2025). For a thesis that fine-tunes the same base model under two different objectives this matters in two ways. The medical fine-tuning data must be informative enough to actually shift the model towards medical terminology, while the update should not be so aggressive that the model loses the general translation competence it acquired during pre-training. The use of PEFT,

discussed in Section 2.1.3, is also relevant here, since restricting updates to a small set of additional parameters limits how far the model can drift from its pre-trained state (Shi et al., 2025).

Taken together, these observations support the choice of starting from a pre-trained multilingual translation model and adapting it to the medical domain through fine-tuning, rather than training a specialised model from scratch. They also motivate a careful comparison between SFT and DPO as adaptation objectives, since both rely on the same underlying pre-trained representations but differ in how strongly and in what direction they adjust the model.

2.1.3 Parameter-Efficient Fine-Tuning and LoRA

Fine-tuning is the continuation of pre-training a LLM. If an LLM is needed for a specific task or dataset on which it has not been pre-trained, fine-tuning is used to train the model further (Jurafsky and Martin, 2026). SFT (Section 2.2) is a well-known fine-tuning technique which performs well but using it on LLMs is expensive, because they have billions of parameters. Updating them during training requires big amounts of memory, processing power and a lot of time. PEFT methods, such as LoRA, are able to reduce this cost this drastically.

PEFT is used to update only a small subset of parameters while fine-tuning, instead of all of them. The parameters that are not touched during fine-tuning remain frozen, meaning they keep their pre-trained values LoRA is a specific PEFT method which is directly applicable to the transformer architecture (Hu et al., 2021). The large weight matrices inside the transformer, including query, key, value and output matrices, are frozen during the fine-tuning and are not updated directly. Instead, for each frozen matrix W of size $[k \times d]$, where k and d are the dimensions of the original weight matrix, two smaller matrices A of size $[k \times r]$ and B of size $[r \times d]$ are introduced and trained in its place, where r is the rank of the low-rank approximation. The rank r is chosen to be much smaller than k and d, which is what makes the update low-rank. The product AB approximates the change ΔW that full fine-tuning would have applied to the original matrix, so that the effective weight becomes $W + AB$ (Hu et al., 2021). The reduction of the trainable parameters significant because in the original LoRA experiments they applied they method to a 175B-parameter model. This process reduced the number of trainable parameters and cut the GPU memory requirements massively in comparison to full fine-tuning.

This approach has several advantages. It requires far less memory, because the gradients only need to be computed for the small matrices A and B instead for one big one W. Additionally, it adds no extra cost at inference time because the two trained matrices can be added simply into the original weights. And lastly, LoRA is really flexible. This means that different LoRA modules can be trained for

different tasks and or domains. They can also be changed by adding or subtracting them form the base weights (Jurafsky and Martin, 2026).

LoRA contains two main key points that make it appropriate for the experimental setup in this Master's thesis. First, Tower is a model with approximately seven billion parameters, Qwen and Ministral with approximately eight billion parameters. In the experiments, a SFT variant is compared to a DPO variant built on the same base model. Full fine-tuning of both variants would require a big amount of computational resources and memory, which are not realistically available in this context. Thanks to LoRA, those runs are feasible on a single GPU and since the base model remains frozen in both terms, the SFT and DPO adapters are trained on top of identical pre-trained weights. This enables the isolation of the effect of the training objective from a drift in der base model. This is the central comparison in this Master's thesis.

Connecting this background to the work that follows, the experiments in this Master's thesis contain several constant elements that make it able to have a controlled comparison. Most importantly, the same pre-trained base models, Tower, Qwen and Ministral, are used, their original weights remain frozen, and adaptation is carried out with LoRA adapters on the same medical English-to-French data. What differs between the conditions is the training objective, SFT versus DPO, and, within the DPO setting, the use of terminology-based preference pairs as opposed to plain preference pairs as the learning signal. The detailed experimental configuration, including data preprocessing, hyperparameters and evaluation is described in Chapter 3.

2.2 Supervised Fine-Tuning

This section provides an overview of SFT as a training paradigm. SFT serves as the primary method for adapting LLMs to domain-specific translation tasks, including medical terminology. Understanding its underlying principles, assumptions, and limitations is essential for interpreting the baseline performance against which later methods, such as preference-based optimization, are compared to.

This method builds directly on the primary paradigm in modern NLP, in which large models are first pre-trained on often unlabelled text using a self-supervised objective, such as next-token prediction. It is continuously specialised through additional training runs on smaller, curated and domain-specific dataset that serve as an example for the desired behaviour (Ouyang et al., 2022; Stiennon et al., 2020). Therefore, SFT can be places in the broad sequence of stages through which modern LLMs are trained. The stages are pre-training and SFT. This two-stage approach is relative simple and in combination with its empirical effectiveness it has made SFT the default starting point for almost every modern alignment

pipeline. It also frequently serves as a basis for pipelines that ultimately employ Reinforcement Learning (RL) (Subsection 2.3.1) or preference optimization in later stages (Rafailov et al., 2023).

SFT treats the adaption problems internally as a standard sequence-to-sequence learning task (Andrew and Richard S, 2018). Given a dataset of input-output pairs, where each input represents a prompt, instruction, or source sentence and each output represents the gold target sequence that the model should produce, the parameters of the pre-trained model are updated to maximise the conditional likelihood of the target outputs given the inputs. In the experiments of this Master's thesis, the input is a translation instruction with an English source sentence and the output is the French translation of that sentence. The used prompt format contains a glossary of medical terms and their target-language equivalents illustrated below.

> "Glossaries: "diabetes mellitus" → "diabète sucré". Translate the source text From English to French following the provided translation glossaries. Rules: output strictly the French translation. Do NOT repeat the English text. Do NOT include labels like Ënglishör Ërench:. Do NOT add quotation marks, explanations or any extra text. Now translate the following: English: Optisulin is used to reduce high blood sugar in adults, adolescents and children of 6 years or above with diabetes mellitus. French:"Optisulin est utilisé pour diminuer le taux de sucre dans le sang chez l' adulte, les adolescentes et les enfants à partir de six ans en cas diabète sucré."

This format explicitly states the terminology that needs to be used in the sentence. Formally, the objective minimises the negative log-likelihood of each token in the target sequence conditioned on the input and all preceding target tokens. This corresponds to a standard cross-entropy loss applied autoregressively (Ouyang et al., 2022). This objective is mathematically identical to the pre-training objective in causal language modelling. The main difference is that the loss is typically computed only over the target portion of each example rather than the entire sequence. This ensures that the model learns to generate the desired completions rather than to reconstruct the prompts themselves (Ouyang et al., 2022; Stiennon et al., 2020).

Instruction tuning is the way to make a LLM respond better by following instructions included in the prompt. It is a form of supervised learning. While other approaches like pre-training and PEFT (Subsection 2.1.3) update all parameters in the LLM, instruction tuning stands out with a low training cost of complete model training (Jurafsky and Martin, 2026). The instructions themselves serve also as limitations for the model because they can include certain restrictions, such as length limitation, for the output sequence. While the general idea of SFT

is to make the model learn and adapt to a certain task or domain, the focus of instruction tuning is not to teach the model a task but rather teach the model to be able to follow the instructions in general (Jurafsky and Martin, 2026).

The quality, diversity and representativeness of the fine-tuning dataset are decisive factors in determining the resulting model behaviour. High-quality demonstrations that accurately reflect the intended task distribution allow the model to generalise beyond the specific examples seen during training. Whereas noisy, biased or insufficiently diverse data tends to produce models that overfit to superficial patterns or fail out-of-distribution inputs (Ouyang et al., 2022). Recent work emphasises that the construction of instruction datasets requires careful attention to coverage across tasks, domains and linguistic styles. In addition, filtering and deduplication procedures, often based on similarity metrics or on heuristic quality rules, are essential components of a successful fine-tuning pipeline (Ouyang et al., 2022; Stiennon et al., 2020). Hybrid approaches that combine manually curated seed examples with automatically generated instances have emerged as a pragmatic compromise between the high quality of human annotation and the scalability of synthetic generation (Agrawal et al., 2024).

SFT plays a dual role in the experiments of this Master's Thesis. First, it serves as the principal baseline against which the proposed DPO approach is evaluated. The research question of whether preference-based optimization can improve terminology-aware medical MT beyond what is achievable through SFT alone presupposes a strong and well-tuned SFT model as the point of comparison. Without a competitive SFT baseline, any improvement attributed to preference optimization would be confounded with deficiencies in the underlying supervised model (Henderson et al., 2018). Consequently, considerable methodological care must be taken to ensure that the SFT model used in the experiments represents a fair and competitive instantiation of the standard fine-tuning paradigm, trained on the EMEA corpus (Tiedemann, 2012) with appropriate hyperparameter selection. This is explained in detail in Chapter 3.

SFT also serves as the necessary initialisation step for the preference optimization stage that follows (Ouyang et al., 2022; Rafailov et al., 2023). As established in the RLHF pipeline introduced by Ouyang et al. (2022) and subsequently adopted as the standard template for language model alignment, preference-based methods do not operate on raw pre-trained models but rather on models that have already undergone SFT (Stiennon et al., 2020; Rafailov et al., 2023). The supervised stage is responsible for giving the model the basic task competence required to produce candidate outputs of sufficient quality that meaningful preference judgements can be made between them. A model that has not yet learned to produce coherent translations in the target domain would generate outputs so poor that pairwise preference comparisons would carry little information. SFT therefore establishes

the operational regime within which subsequent preference optimization can refine the model's behaviour. The same principle applies whether the subsequent stage is RLHF or DPO (Ouyang et al., 2022; Rafailov et al., 2023).

SFT plays a central role but still has limitations which motivate the research for better SFT or alternative approaches for training language models. One issue is for example that maximum likelihood training assumes that every token has the equal amount of correctness (Christiano et al., 2017). This is essentially important for tasks such as MT where multiple translations of one source can be valid and the translation depends on register or domain conventions and context. In this case, a model trained only on maximum likelihood gets high scores but might still not achieve the domain-specific terminology. This limitation is particularly concerning in the medical translation context and also motivates the experiments of this Master's thesis and is further discussed in Chapter 2.5, talking about characteristics of medical text.

Another limitation of SFT is that is not possible to include negative examples. Each example is treated as a positive training example and also treated independently. So the model gets signals on what it should produce but not on what it should not (Rafailov et al., 2023). RL and preference optimization in comparison are designed to use this information and this different learning structure is one of the main why it can outperform SFT on tasks where grained distinctions matter (Ouyang et al., 2022).

All of the above show that SFT is a necessary part of an alignment pipeline but might be not enough for every task and domain. It provides a good basis and offers also a strong baseline against what the rest can be compared. In the following Sections RLHF and DPO will be explained and how these address the limitations mentioned above.

2.3 Direct Preference Optimization

This section focuses on DPO which is the central preference-learning method applied in the experiments of this Master's thesis. To give the necessary background, the section first introduces the RL fundamentals that motivated DPO and then describes the method itself including its theoretical foundation, the training process, as well as advantages and limitations. Throughout the section, the main focus lies on aspects of DPO that are most relevant for fine-tuning a language model on medical preference data.

Following the introduction of SFT and instruction-tuning from the previous Section 2.2, DPO is a way to further improve a language model through learning from preferences, for example. This approach is considerate more detailed because with preferences, the model can improve according to certain qualities. It does not

rely on a specific instruction, which already implies knowing what the instruction should look like in order to get the desired output. Additionally, it also does not require the knowledge of what the output must contain because preference learning shows what the end goal is and what quality the output should have. This happens through indication of which response is preferred over another in a set of hypotheses (Jurafsky and Martin, 2026).

2.3.1 Reinforcement Learning and Reinforcement Learning from Human Feedback

RL is the theoretical and methodological foundation on which preference-based language model alignment is built (Rafailov et al., 2023). In the framework of RL an agent learns to select an action that maximises the cumulative reward signal over others (Andrew and Richard S, 2018; Bertsekas, 2019). In this framework, a policy is a rule that tells an agent what to do in each situation. For every state that it could be in, the policy gives a set of possible actions and also how likely the agent is to choose each of them. The final goal of RL is to find the best policy, this means the one that helps the agent get as much reward as possible over time.

While SFT assumes labelled training data that specifies the correct output for each input, RL relies on the reward signal (Andrew and Richard S, 2018). This makes it possible for RL to capture the scale of correctness better, which cannot always determined by one single reference output. This is the case for translation, where a sentence or word can have multiple correct translation, depending on context, register and more. What matters in this case is not the exact match with the gold reference but the conformity of qualities like terminological accuracy, fluency and faithfulness.

However, it is difficult to scale RL because the results are often unstable or hardly unproduceable. Small changes of a random seed, hyperparameters or the environment can lead to different outcomes even if the changes are really small (Henderson et al., 2018). These reproducibility problems can also appear in language model alignment. The training of large models is so expensive, that running many ablation studies to identify what actually improves the models substantially becomes out of range (Ouyang et al., 2022). This combination of cost and instability is a part of the motivation to develop DPO.

A further extension of the framework to align language models is RLHF (Ouyang et al., 2022). It uses human-annotated preference data instead of a manually made reward function. First, the reward model is trained to predict which output of two a human annotator would prefer. Then, the language model is optimised with RL, usually Proximal Policy Optimization (PPO) (Schulman et al., 2017), to maximise the predicted reward. In parallel, a Kullback-Leibler (KL)-divergence term keeps

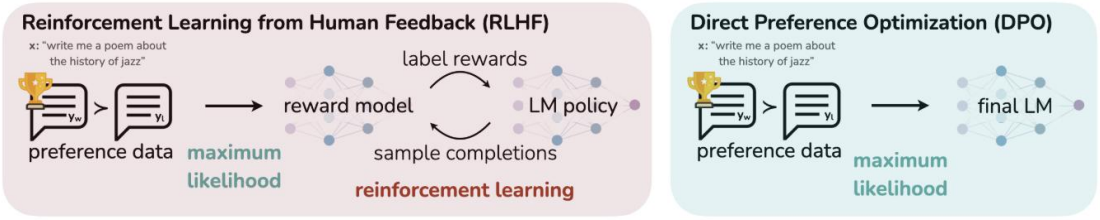

Figure 3: Comparison of RLHF and DPO. (Rafailov et al., 2023, 2)

it close to the refernce policy. This approach has become a standard method for aligning LLMs, but it still includes the instability and cost of RL and adds the complexity of a separately trained reward model. How DPO addresses these issues is further explained the upcoming Subsection 2.3.2.

Building on what is explained above, DPO can be understood as a streamlined alternative to traditional RLHF pipelines. While RL and RLHF rely on reward models and iterative policy optimisation, DPO uses preference data directly to adjust model behaviour without explicit reward modelling. Therefore, DPO can be grouped with the broader reinforcement approaches but with some conceptual differences, practical advantages and relevance for improving translation quality, particularly in the context of medical terminology.

2.3.2 Theoretical Foundation

DPO (Rafailov et al., 2023) optimises language models for human preferences without including an explicit reward model or RL. Figure 3 shows how DPO reduces the complexity of the RLHF pipeline (Ouyang et al., 2022). It removes the separate reward model training, RL optimisation and sampling during training. The much simpler DPO pipeline uses an analytical mapping between reward functions and optimal policies instead. The language model and the reward model are both represented in the policy network within one single step. This shows that a constrained reward maximisation can be solved with a single-stage policy training.

Rafailov et al. (2023) provide a simplified version of the RLHF framework while the goal, maximising the expected reward while keeping the model close to a reference policy, remains the same. The goal is to find the policy that gets the highest expected reward while staying close to the reference policy. The parameter β controls how strongly this deviation is penalised as shown in Equation 1.

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta(y|x)} \left[r_\phi(x, y) - \beta \log \sum_{y'} \pi_{\text{ref}}(y' \mid x) \exp\left(\tfrac{1}{\beta} r_\phi(x, y')\right) - \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \right] \tag{1}$$

Human preferences are modelled with the Bradley–Terry framework (Bradley and Terry, 1952), which says that the chance of choosing one output over another depends on the difference in their rewards. In this model, the probability of preferring output y_1 over output y_2 follows a logistic function of their reward difference as shown in Equation 2. When the policy-based reward formula is plugged into this framework, the partition function cancels out. This means preference probabilities using can express only the policy and the reference policy, without ever computing the intractable partition function.

$$p^*(y_1 \succ y_2|x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \tag{2}$$

This cancellation leads directly to the DPO loss, which increases the log-probability that the model assigns higher likelihood to preferred outputs than to dispreffered ones. This is scaled by β and the ratio to the reference policy. Rafailov et al. (2023) show that this formulation is fully expressive because any reward function consistent with the Bradley–Terry model can be represented through the policy-based reparameterisation of DPO. An important regularization mechanism in DPO is KL-divergence (Kullback, 1951), denoted D_{KL}. It measures how much one probability distribution differs from another. KL-divergence prevents the optimised policy from drifting too far from the reference policy during training which ensures that the model continues to produce coherent output.

In comparison to training pipelines in RLHF and SFT, DPO differs significantly in complexity. SFT uses a single stage of supervised learning, only learns from preferred outputs and ignores the preference structure in the data. It can teach a model to produce fluent and or task-specific outputs but it has no mechanism for distinguishing between outputs of different quality. There is only right or wrong, because all training example are treated as equally correct. In a translation setting SFT cannot get more information about which of two translation handles terminology more accurately. The loss function treats a precise and an imprecise translation as the same if they both are included in the training data.

Continuing with RLHF (Ouyang et al., 2022), that addresses this issue by including human preference signal into training. But this process consists of three separate stages. At first, and initial SFT stage is used to produce a model that is capable to generate reasonable output. Second, a reword model is trained on human-annotated preference pairs using the Bradley-Terry framework (Bradley and Terry, 1952). Third, the language model is optimised using RL, more precisely

PPO. In this stage the reward model provides the training signal and KL-divergence constrains how far the policy can drift from the reference. This three-stage approach is expensive regarding computational costs and also requires sampling of new outputs from the model during training. Additionally, it involves a big number of hyperparameters across alls stages which need to be tuned separately and carefully. The reward model itself introduces another source of error, because it is trained on a finites set of human annotations. This is the reason it may not generalise reliably to all output produces by the language model (Ouyang et al., 2022).

DPO simplifies this and reduces the process to two stages. One is the optional SFT fine-tuning to create a reference policy. The other is the direct optimization on prefence data using binary cross-entropy loss (Rafailov et al., 2023). The explicit reward model does not exist in this process, there is no sampling during training and there are only a minimal hyperparameters in addition to the learning rate and β. The main idea of DPO is that the reward-maximisation problem used in RLHF acutally has a closed-form solution that can be written directly in term of the policy itself. All this happens without the need to compute the reward model. Through that it is possible to train on preference data in one single supervised step. The problem is now treated as a classification task over pairs of preferred and dispreferred outputs. Rafailov et al. (2023) also show that DPO achieves a better reward-KL frontier than PPO across all KL values and it is more stable and more efficient computational-wise.

This simplification is also meaningful for the present Master's thesis. Training Tower, Qwen and Ministral with LoRA adapters already operates near the limits of the available hardware. So adding a separate reward model training stage with an online sampling loop would make a full RLHF pipeline not feasible in this setting. The two-stage pipeline provided from DPO is not only simpler but also a necessary choice given the constraints.

2.3.3 Training Process

The training process of DPO can be split up into three stages that are connected with each other. First, the construction of the preference data. Second, the initialisation of the policy and the reference model. And third, the optimisation of the policy on the preference data. Each of the mentioned steps has implications for the experiments that will be captured in detail later in this Master's Thesis in Chapter 3.

The preference dataset for DPO consists of triplets of the form (x, y_w, y_l). In this case, x refers to the input prompt, y_w is the preferred or chosen output (winner), and y_l is the dispreferred or rejected output (loser) (Rafailov et al., 2023). In RLHF these triples come from human annotators who compare two or more candidate output for the same prompt. However, this is not the case in practice, where

perference pairs are often constructed through other processes like for example by ranking the outputs with a stronger model, using heuristic quality scores or pairing a clean reference translation with a worse version of it. The quality and the structure of the preference data have a direct effect on what the model learns. Pairs where y_w and y_l differ in many ways only teach the model a broad sense of quality. In contrast, pairs that differ in just one specific aspect, like a single terminology choice, help the model learn much more precise distinctions. For medical translation, this second type is especially important because the goal is to improve a narrow but critical property of the output.

DPO requires a reference policy, denoted π_{ref}, against which the optimised policy, denoted π_θ, is regularised. According to Rafailov et al. (2023), the SFT model that generated or scored the preference data is used as the reference policy. This minimises the distribution shift between the policy being optimised and the policy that produced the data. In this Master's thesis, this means the the SFT model trained on the EMEA corpus serves both as the starting point for DPO fine-tuning and as the reference policy, explained further in Chapter 3. Both π_θ and π_{ref} are therefore initialised from the same checkpoint, and only π_θ is updated during training.

For each batch of preference triplets, the training procedure computes the log probabilities that the current policy π_θ and the frozen reference policy π_{ref} assign to the preferred and dispreferred outputs. These four log probabilities are combined into the DPO loss. This is a binary cross-entropy term on the difference between two log-probablity ratios. One is the ratio for the preferred output and the other is the ratio for the dispreferred output. Each of them is between π_θ and π_{ref}. When the current policy assign relatively more probability to y_w and less to y_l than the reference policy does, the loss is low. The parameter β scales this difference and thereby controls how strongly the model is pushed to deviate from the reference (Rafailov et al., 2023). The gradients of this loss are then used to update the parameters of π_θ via standard gradient-based optimization.

Several practical aspects of the DPO training process are worth highlighting. First, because the reference policy is frozen and used at every step, both π_θ and π_{ref} must be available in memory during training. When using PEFT methods such as LoRA, this overhead can be reduced by keeping a single base model in memory and applying the trainable LoRA adapters only to π_θ, while π_{ref} uses the base model without adapters. Second, the choice of β has a direct and non-trivial effect on training behaviour. Small values allow the policy to deviate more strongly from the reference and tend to improve performance, but values that are too small can cause the model to degenerate (Liu et al., 2025). Third, the quality of the preceding SFT stage shapes the entire training trajectory, since it determines both the reference policy and the initialisation of π_θ (Feng et al., 2024). These considerations are

taken up in more detail in the next subsection.

2.3.4 Advantages and Limitations

DPO has several practical advantages in comparison to RLHF that make it attractive for fine-tuning LLMs in settings with constrained means. DPO does not only remove the need for a separate reward model, it also replaces the iterative RL optimisation with a single binary cross-entropy loss. Through that it substantially reduces the complexity of training and computational cost (Rafailov et al., 2023). In the context of the present Master's thesis, this is especially interesting, because fine-tuning 7B and 8B-parameter models with PEFT methods such as LoRA already demands a lot from the available hardware. Hence, a two-stage pipeline without online sampling is far more tractable than a full RLHF setup. Another advantage consists in DPO's training stability. Without the need of the exploration-exploitation dynamics that are inherent to policy gradient methods, it is able to avoid the instability and reproducibility problems that have been documented for RL-based alignment. Rafailov et al. (2023) show in a theoretical point, that DPO achieves better reward-KL trade-off than PPO across all KL values. They also suggest that the simplification does not come at the cost of alignment quality.

Despite various strengths and advantages, research identified several limitations of DPO that are relevant to the experimental context of this Master's thesis (Feng et al., 2024; Liu et al., 2025). The asymmetry of how the DPO loss function updates the model is one of them. Feng et al. (2024) provide a theoretical analysis using gradient vector fields and show that the DPO loss decreases the probability of generating dispreferred responses at a faster rate than it increases the probability of generating preferred responses. This means that the model learns rather to avoid bad outputs than to produce good outputs. This can be especially problematic when the preferred and dispreferred responses are lexically similar, since the gradients pulling the preferred output are partly counteracted (Feng et al., 2024). In a medical translation context, where the difference between a correct and an incorrect terminology choice may be a single word, this asymmetry is a meaningful concern.

Another limitation is that DPO results are very sensitive to the quality of the preceding SFT stage. Feng et al. (2024) show that where the model starts in the DPO loss landscape, determined by how well SFT aligned it, strongly affects the final results. If SFT is too weak, the model begins in a poor region of the gradient field and may not learn to produce better preferred responses, even after DPO training. This matters here because the SFT model trained on the EMEA v3 corpus is both the reference policy and the starting point for DPO, so any weaknesses from that stage carry over.

The role of the reference policy and the KL-divergence constraint pose an

additional limitation, even though a more subtle one. Liu et al. (2025) investigate this in their research and find that DPO is sensitive to the strength of the KL constraint, controlled by the hyperparameter β. A smaller β usually helps by letting the model move further away from the reference policy, but if β becomes too small, performance drops or the model can even break down. This is difficult in practice because the best β is unknown and must be found experimentally. They also show that although the constraint acts at the sequence level, it does not stop extreme probability shifts at the token level, where some tokens can be downweighted by several orders of magnitude, which can destabilise the model.

Liu et al. (2025) also identify a performance ceiling introduced by the reference policy itself. The reference policy is usually the SFT model that should be improved by DPO, the KL constraint ends up penalising movement away from a model that is already suboptimal. This creates a built-in tension where the constraint is needed to prevent model collapse, but it also limits how far the model can shift toward better outputs. In this Master's thesis, this means that DPO's improvements are partly bounded by the quality of the SFT model trained on the EMEA v3 corpus, and one contribution of the later experiments is to measure this upper bound in practice.

Taken together, these limitations show that DPO is not automatically better in all settings. Its success depends on careful tuning and on the quality of the earlier training steps. The theoretical results from Feng et al. (2024) and Liu et al. (2025) help frame the findings of this Master's thesis, meaning improvements in medical terminology translation may be limited by the asymmetric gradient behaviour of the DPO loss, the strength of the SFT starting point, and the method's sensitivity to hyperparameters.

2.4 Machine Translation

This section introduces the MT background that is relevant for this Master's Thesis. It focuses on NMT, on LLM-based MT and on the automatic evaluation metrics used in the experiments. MT is known as one of the oldest tasks of computer science and is still evolving, because translation quality still varies strongly across language pairs, corpora and training conditions (Specia and Wilks, 2022). Jurafsky and Martin (2026) explains, that MT is most commonly used to give people access to information provided in other languages. This also applies translation in the medical context, where patients may need to read a patient information leaflet in another language or communicate with clinicians when they live abroad or do not know the used language well enough. MT therefore also reduces the asymmetry of access according to languages on the web and supports a more equal access for speakers of different languages.

In the following subsections, different parts of MT are discussed in more detail. Starting with Subsection 2.4.1 focusing on the sequence-to-sequence framework, attention and transformer-based architectures in the translation context (Jurafsky and Martin, 2026; Specia and Wilks, 2022). Followed by the LLM-based MT in Subsection 2.4.2, a paradigm used in this Master's Thesis. Subsection 2.4.3 describes the automatic evaluation metrics applied in the experiments and discusses their strengths, as well as their limitations in general and in domain-specific settings (Papineni et al., 2002; Rei et al., 2020; Popović, 2015).

2.4.1 Neural Machine Translation

In the long history of MT, NMT is one of the youngest paradigms. It arised in the 2010s and treats translation as a sequence-to-sequence problem that benefits strongly from the attention mechanism. The field is evolving rapidly and is the dominant paradigm and research and deplyed in many systems (Specia and Wilks, 2022). Most NMT systems are built on the encoder-decoder framework, where an ecoder maps the source sentence to a continuous representation and a decoder generates the target sentence conditioned on that representation (Jurafsky and Martin, 2026).

As mentioned above, sequence-to-sequence models treat translation as a mapping from an input sequence of source tokens to an output sequence of target tokens (Sutskever et al., 2014). The two sequences can have different lengths and derive from different vocabularies (Jurafsky and Martin, 2026). The model factorises the conditional probability of the target sentence as a product of token-level probabilities. In this case, each target token is predicted given the source sentence and all previously generated target tokens. In the original encoder-decoder framework, the encoder compresses the entire source sentence into a context vector with a fixed size. This is then passed to the decoder (Sutskever et al., 2014). The decoder generates the target tokens one after the other. This way each prediction is conditioned on the context vector and on the tokens it already has produced. However, this fixed-size context vector has to encode the meaning of a whole source sentence in to a single representation. Through this it also becomes harder as sentences get longer. This limitation motivated the attention mechanism following this.

To address the issue of long source sentences needed to be put into one context vector, the attention mechanism was introduced. It does not rely on a single representation, because the decoder allows to look back at all the encoder states and assigns weights to every single source token at every encoding step (Chorowski et al., 2015). Therefore the encoder representations are a wighted combination computed by the decoder. The weights reflect how relevant each source position is for predicting the next token. This soft alignment between source and target tokens has two important effects for translation. First, through this process the model has

access to detailed source information at every step, improving translation quality for longer sequences. Second, an implicit aligned is produced that is useful for the analysis and also debugging of translation outputs (Chorowski et al., 2015).

Modern NMT is based on the transformer architecture introduced in a previous Section (2.1). It was groundbreaking and replaced the encoder-decoder models because it only relies on self-attention and not on sequential recurrence. The training is therefore more parallelisable and models can capture long-range dependencies more effectively (Vaswani et al., 2017). In transformer-base NMT the encoder and the decoder consist of stacked self-attention and feed-forward layers. The cross-attention between encoder and decoder has the role of the attention mechanism described above. With transformer-based encoder-decoder style as basis, two other MT systems evolved. First, one that keeps the encoder-decoder structure and trains it specifically as a MT system. And second, a decoder only LLM that treats translation as one of many generation tasks. In the following subsection, the latter is discussed in detail.

2.4.2 LLM-Based Machine Translation

LLMs can be used for MT by using zero-or few-shot setting or prompting (Jurafsky and Martin, 2026). When treating translation as a text-generation task, the source sentence is giver to a decoder-only LLM together with an instruction. The model then produces the translation as part of the generated output. This fundamentally different from the ecnoder-decoder NMT systems, where each model is traned from scratch for translation only.

The use of general-purpose LLMs for translation has advantages and limitations. One one hand, LLMs can use general world knowledge and linguistic competence acquired during pre-training which is helpful when encountering rare words and stylistic variation (Jurafsky and Martin, 2026). Through prompting with little data or PEFT they can be adapted easily. On the other hand, general-purpose LLMs are usually not designed for translation and their performance on domain-specific or low-resource langauge pairs can be worse than NMT systems. This issue has been addressed by pre-training and fine-tuning LLMs on mutlilingual data and translation tasks. One example is the TowerInstruct model (Rei et al., 2025; Unbabel, 2024). It is a decoder-only LLM adapted for translation between many language pairs including English and French. This makes it a suitable base model for these experiments.

2.4.3 Evaluation Metrics

To capture the different aspects of translation quality, MT systems require multiple complementary metrics. This Master's Thesis uses BLEU (Papineni et al., 2002),

chrF (Popović, 2015) and COMET-DA (Rei et al., 2020) as automatic evaluation metrics. These three metrics differenr in what they measure and hwo well they correlate wiht human judgement. Their combination makes them an informative for the comparison of SFT and DPO.

BLEU (Papineni et al., 2002) is a lexical overlap metric and one of the most widely used and well-known automatic metrics in MT. It compares system outputs to one or more reference translation using n-gram precision. It is efficient to compute, which is why it is broadly adopted. In older tuning procedures like minimum error rate training it was used to optimise the target by optimising weights iteratively so the translation that was closest to the reference was ranked on top of the list (Specia and Wilks, 2022). However, BLEU has also limitation such us its dependence on the number of references, maximum n-gram length and smoothing technique. All together, this makes a direct comparison across studies unreliable without standardisation. Post (2018) address this issue with SacreBLEU, where they apply a fixed metric-internal preprocessing pipeline. Another limitation of BLEU is that it relies on exact n-gram matching. This excludes morphological variations like synonyms and weights all matched n-grams equally. Therefore, BLEU often does not correlate well with human judgement (Specia and Wilks, 2022) and for this reason it alone is not a sufficient basis for comparing SFT and DPO.

chrF (Popović, 2015) improves BLEU by looking at character n-grams instead of word n-grams. This allows it to capture morphological similarity and makes it less sensitive to small variations. This is useful for morphologically rich target languages. However, like BLEU, chrF is a reference-based surface-level metric and still struggles to distinguish acceptable paraphrases form domain-specific errors (Gaona et al., 2023).

Neural metrics are based on contextual embeddings and are more promising for domain-specific evaluation. COMET-DA (Rei et al., 2020) is one of them and uses cross-lingual representations and correlates more strongly wiht human judgements that surface-level metrics when evaluated against Mutlidiemntional Quality Metrics annotations (Lommsel et al., 2013). But alls reference-based metrics, including COMET-DA inherit biases and possible errors of the reference translations they are compared against (Gaona et al., 2023).

Recent work has also explored using LLMs as a judge for general quality assessment and for accuracy in translation (Jurafsky and Martin, 2026). This paradigm is instantiated in approaches such as the GPT Estimation Metric Based Assessment (GEMBA) (Kocmi and Federmann, 2023) and offers a reference-free or reference-light alternative to the other introduced metrics. It is not used in this Master's thesis but it represents one of the most active areas in MT evaluation.

2.5 Medical Machine Translation

Medical MT presents unique challenges that distinguish it from general-domain translation. Within healthcare contexts, translation systems serve as important intermediaries in communication between patients and clinicians, where translation errors can have serious consequences for patient safety and healthcare delivery (Mehandru et al., 2022). Current general-purpose MT systems often fail to meet the specific requirements of medical applications (Gaona et al., 2023), and NMT output can appear fluent while containing serious semantic errors that are difficult to detect because the text reads naturally.

It is proven that MT of specialised domains benefits from domain-adapted systems. This also refers to translation service providers that often use terminology databases in order to handle specialised content. For professional workflow, terminological accuracy and consistency is one of the main concerns. Quality assurance tools detect incorrect terminology according to a reference glossary and most Computer-Assisted Translation (CAT) tools have terminology recognition integrated in their environment (Bowker and Corpas Pastor, 2022).

The limitations of SFT discussed in Section 2.2 apply especially to the medical translation domain. Medical terminology contains strict conventions that are collected in databases such as UMLS (Bodenreider, 2004). The correct rendering of terminology can be critical for patient safety and otherwise can lead to harm or even be fatal. Translations that are fluent and stick to the source can still be using incorrect terms but may receive a high likelihood if the model has been trained on general medical parallel text. This is especially tricky if the data itself contains inconsistent terminology. Skianis et al. (2020) confirm that automatic metrics often do not measure terminological accuracy perfectly and that models that are solely optimised on those metrics may underperform in terminology-specific evaluations.

2.5.1 Terminology

Terminology poses some special issues in comparison to general language translation. In the evaluation of MT, terminological accuracy can be measured by comparing terms to the gold standard. The issue is that this approach often fails to account for snyonyms, which can be treated as equivalent even if they are not present in the gold standard (Korkontzelos and Ananiadou, 2022). UMLS (Bodenreider, 2004) offers a huge database of medical terms sorted in ontologies. Those are formal repersentations of knowledge which are organised in maps of meaning (Navigli, 2022). Knowledge can be represented in different from s starting from plain textual content up to ontologies as seen in Figure 4. Ontologies contain different building blocks such as concepts or classes, instances and relations and are divided in sections. Whereas there exist upper ontologies that do not belong to a specific domain and

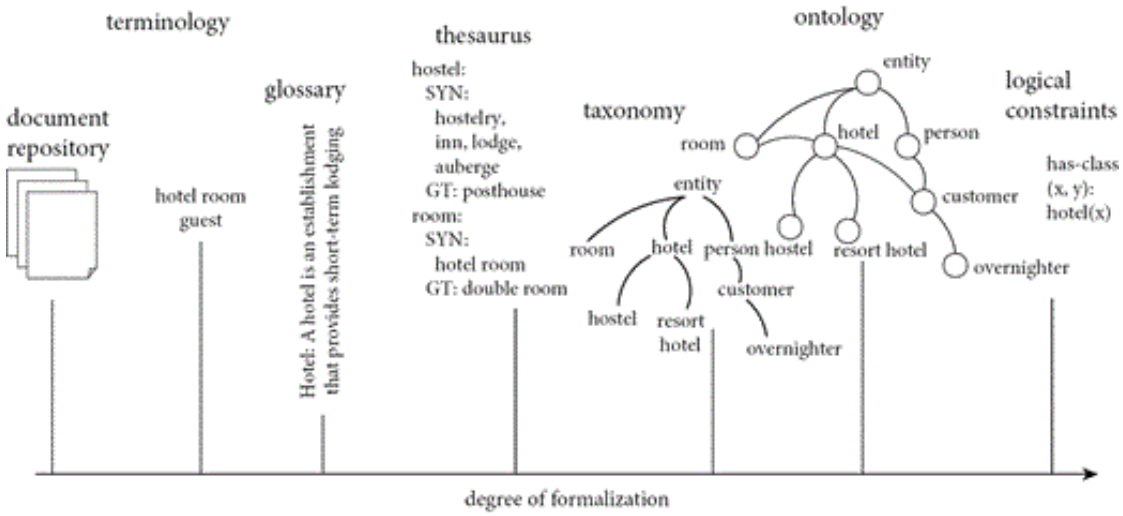

Figure 4: The different degrees of formalization from unstructured textual content to ontologies (Navigli, 2022, 519)

contain mostly general concepts, UMLS (Bodenreider, 2004) is a domain ontology specialized on medical concepts (Navigli, 2022).

Biomedical ontologies and controlled vocabularies or glossary including UMLS (Bodenreider, 2004), ICD (Fritz, 2000), MeSH (Lipscomb, 2000) and SNOMED (Bos and Donnelly, 2006) provide a standardised mapping for medical concepts. UMLS alone covers over one million concepts and 2.8 million terms. Which makes this resource a plausible gold standard for the comparison of DPO and SFT on medical texts (Korkontzelos and Ananiadou, 2022). An important distinction needs to be made between term variation which means the same concept has different forms, and term ambiguity which means a single from can map to multiple concepts. Those are fundamental challenges in term recognition and it also includes abbreviation. A medical MT evaluation must therefore decide whether the model output that produces a valid variant of the reference counts as correct or not and is in this case accurate enough (Korkontzelos and Ananiadou, 2022).

2.5.2 Characteristics of Medical Language

Medical language differs from general language through consistent grammatical and terminological features. At the lexical level, medical terminology largely derives from Greek and Latin roots, forming a morphological system in which prefixes, roots, and suffixes each carry a stable meaning (Banay, 1948).

For example, the root *cardi(o)-* means heart-related and *hepat(o)-* liver-related, while the suffixes *-itis* refers to inflammation and *-ectomy* to a surgical removal.

This means, combined terms such as *hepatitis* can be split into *hepat-* (liver) and *-itis* (inflammation). Because this Greco-Latin layer is largely shared between English and French, many such terms are closely related and similar in their expression across the two languages, like *hepatitis / hépatite*, *cardiology / cardiologie*, and *gastrectomy / gastrectomie*. At the conceptual level, the same kind of regularity is captured explicitly by UMLS (Bodenreider, 2004), which groups every surface form that denotes a given concept across synonyms, source vocabularies, and languages under a single Concept Unique Identifier (CUI). For instance, the UMLS concept *Diabetes Mellitus* (CUI C0011849) gathers English atoms such as *Diabetes Mellitus* (MeSH) and *Diabetes mellitus (disorder)* (SNOMED) together with their French equivalents, including the preferred form *diabète sucré* alongside the shorter variant *diabète*. Because all of these variants share one identifier, UMLS can be used both to recognise a medical term in the English source and to retrieve its formally preferred French equivalent, the mechanism underlying the term-enriched prompts and preference pairs used in this thesis. At the syntactic level, medical texts frequently use nominalised constructions, passive voice, and impersonal formulations, while patient-facing documents use simpler sentence structures suited to varying levels of health literacy (Zappatore and Ruggieri, 2024).

In comparison to general-domain NLP, the biomedical domain does not prefer well generalised results. In this specific domain abbreviations are often used, along to specific morphology and a high rate of out-of-vocabulary tokens. Therefore it also behaves differently than general-domain MT under the same training regimes (Cohen, 2022). This also causes frequent struggles for MT in this domain and especially on MT for assisting health communication. Another issue or less negative characteristic of medical language is, that it can be seen as a sublanguage. They differ from other in their linguistic dimensitons, which causes the not possible generalisation of biomedical texts like for general texts. Sublanguages are domain specific and serve a special purpose. They often use distinctive workd classes in sentences, are consistent and have a certain economy of expression which also results in different work-co-occurrence patterns as in general-domain data (Kittredge, 2022).

Terminology is the most important dimension for MT quality in this domain. Medical terminology is governed by authoritative resources such as SNOMED (Bos and Donnelly, 2006), ICD (Fritz, 2000), and MeSH (Lipscomb, 2000). A key challenge for MT systems is to render source-language terms with their formally preferred target-language equivalents rather than acceptable but non-standard paraphrases, since terminological inconsistency can affect interoperability across healthcare systems (Zappatore and Ruggieri, 2024; Naveen and Trojovský, 2024). This is also the basis for the terminology signal used in this thesis where for each English source term and a set of French synonyms is retrieved from UMLS, and a candidate translation is credited when it contains one of these synonyms. Table 1

illustrates this with examples drawn from the data used in this work. It shows an English term, two of its retrieved UMLS French synonyms and a French example that falls outside the retrieved set. In multilingual clinical settings, general-purpose MT systems may produce fluent output that still contains terminological errors with clinical consequences, such as a mistranslated dosage instruction or an imprecise description of a side effect (Mehandru et al., 2022; Zappatore and Ruggieri, 2024; Cohen, 2022).

English source term	**UMLS French synonyms (examples)**	**Outside the synonym set**
diabetes mellitus	diabète sucré; diabète	sujets diabétiques
overdose	surdosage; surdosage SAI	dépassement de la posologie
myocardial infarction	infarctus du myocarde; crise cardiaque	-

Table 1: English-to-French medical term examples from the UMLS extraction used in this thesis. For each English source term, the middle column lists two of the French synonyms retrieved from UMLS, and the right column gives a fluent French phrase that does not appear in the retrieved synonym set. A candidate translation is credited for the term if it contains any retrieved synonym; thus *crise cardiaque* counts as correct for *myocardial infarction* because it is itself a retrieved synonym, whereas *dépassement de la posologie* does not count for *overdose*.

2.5.3 Datasets and Benchmarks

The TICO-19 (Anastasopoulos et al., 2020) is a publicly available benchmark dataset providing test and development data across over 30 languages. Documents are translated from English and cover COVID-19-related content. All translations have been reviewed through a strict quality assurance process, achieving average quality scores above 95% across translation directions (Anastasopoulos et al., 2020). In this thesis, TICO-19 is used exclusively as the held-out test set.

The EMEA corpus (Tiedemann, 2012) consists of biomedical documents from the European Medicines Agency, covering medicinal product documents and their translations into official EU languages. In this thesis, EMEA is used as training and validation data for domain adaptation. Two further resources commonly used in medical MT research are the WMT Biomedical Translation Shared Task (Neves et al., 2024), which provides test sets for six language pairs based on PubMed abstracts, and the Khresmoi Summary Translation Test Data (Dušek et al., 2014),

which provides development and test sets for MT of medical article summaries across eight languages including French and English.

SFT on medical translation datasets helps models adapt to domain-specific vocabulary, but approaches that simply maximise the likelihood of reference translations may not adequately capture the nuanced requirements of medical terminology (Mehandru et al., 2022). Rios (2025) present an instruction-tuning approach that integrates medical terminology constraints from authoritative resources directly into training prompts. Combined with LoRA, this approach is computationally efficient and shows substantial improvements in terminological accuracy across multiple language pairs (Rios, 2025). Sun et al. (2025) address the data acquisition challenge for preference optimization by using the model's own quality assessment capabilities to generate preference pairs, requiring initial training on domain-relevant error annotation data so that models can recognize medical terminology errors and semantic shifts (Sun et al., 2025).

2.6 Related Work

This section reviews work directly relevant to the central research question of this thesis: whether direct preference optimization improves medical terminology translation accuracy compared to supervised fine-tuning. The goal is not only to summarise prior work but make explicit how each line of research relates to the two-stage setup used in this Master's thesis and how it informs the interpretation of the results.

Although SFT is the standard way to adapt a model for MT, it is limited especially in the context of correct terminology rather than overall fluency. SFT maximises the likelihood of a single reference sentence and treats every token in that sentence as equally important, which also ends as them count the same in the loss. Through that, the method gives the model no signal that separates clincal tersm from their informal or approximate synonyme (Mehandru et al., 2022; Rios, 2025). This matters in medical translation, because the difference between a fluent paraphrase and the exact term is the difference containing clinical risk. Instruction-tuning wiht built-in terminology constraints addresses this by adding terminological information directly to the training signal. This has been shown to improve terminological accuracy over standard fine-tuning across several language pairs (Rios, 2025). This is relevant because it confirms that terminology is something fine-tuning can improve and it suggests the improvement comes from giving the model clearer signals than just likelihood maximisation. Preference-based training can be seen as another answer to this problem. Instead of an explicit terminology constraint it gives the model a learned preference between chosen and rejected output.

The method behind this preference signal was build on RLHF. In this approach, a separate reward model is first trained to predict human preferences and the language model is then optimised using methods such as PPO. It also includes a KL divergence term that stops the model from drifting too far from its starting point (Feng et al., 2024). This works but is still complex because it needs a reward model in addition to the RL step that is often unstable (Feng et al., 2024). DPO was introduced to get similar results but without an explicit reward model and without the RL loop. It shows that the chosen version can be included directly in the terms of the model's own probabilities. Those so called preference pairs can be used to train the model with a simple loss that is similar to classification. So both SFT and DPO fine-tune the same model, but preference optimisation also uses the contrast between a chosen and a rejected output. This contrast makes it a good candidate for teaching correct terminology.

Feng et al. (2024) analyse the DPO objective. Their main finding is that the loss lowers the probability of rejected outputs more than it raises the probability of the chosen. This is relevant for the present experiments because preference training can be good for reduced occurrence of wrong terms while it does not put the same amount into raising the occurrence. Having fewer errors is not the same as having more correct terms. They also find that the method depends on the quality of the SFT stage before it. If that stage is of poor quality or just weak it is harder for the next stage to perform well. This is a reason to treat the SFT model in these experiments as a baseline and not just as a necessary side product of the DPO process, because the results of preference optimization depend on it.

Liu et al. (2025) focus on the role of the reference policy and make the same point as discussed above but from another view. In standard DPO, the reference policy is usually the SFT model itself and the KL-divergence term constraints it. So the supervised model also sets the limit of how far the training can change the model. Liu et al. (2025) show that this performance is sensitive to how strong the constraint is. If it is too loose the model can break and on the other side it does not learn properly. They also found that a strong reference policy can help but for that to happen it needs to be compatible with the model that is trained. This is possible if they share the same base model or training data for example. Feng et al. (2024) and Liu et al. (2025) show that there is no direct comparison between SFT and DPO because one methods depends on the other. It is rather a comparison between a baseline and the process that is build on that baseline. In this Master's thesis this setup is used.

Another key factor that decides if preference training is useful is the building of preference pairs. Agrawal et al. (2024) provide a methodological basis to construct preference data. They show that quality estimation metrics can be used to rank outputs and this by making it similar to human judgement. The model-training on

this metric-ranked preference data improves translation quality in comparison to the SFT baseline. The most important results of their research are the following. First, SFT on only chosen hypotheses does not improve the final results as preference training on the same data. This supports the idea that it is the contrast between chosen and rejected the improves the model. Second, they find that for preference optimisation to return better results it is necessary to have a regularisation or SFT stage before. For the present experiments the preference pairs are built to encode correct terminology.

The results discussed above are mostly shown on general-domain instructions or overall translation quality. Medical translation adds constraints that make the terminology question harder and more important. The specific problem of English-to-French medical terminology translation was studied by Deléger et al. (2010). They combine knowledge-based and corpus-based methods finding valid French translations for only about half of the target medical vocabulary. This is useful today because it shows how limited fully automatic terminology translation was for this specific language pairs at the time and sets a baseline difficulty against which modern methods should be judged.

Adapting a general model for medical translation is part of a domain adaption process. The survey by Shi et al. (2025) shows the trade-offs involved in said process. The main point is the tension between gaining domain knowledge and keeping the general ability of the base model. They show this in the medical case where training a model on biomedical or clinical text improves medical performance but at the same time the general performance drops throughout the training process. They also note that PEFT methods such as LoRA are widely used to add domain knowledge while limiting this drop. They do not focus on terminology directly but adapting the model to terminology includes the same factors that drive domain adaption in general.

Whether preference optimization carries over from general translation quality to terminology is the open question that the medical literature only partly answers. Yang et al. (2024) combine DPO with Minimum Bayes Risk (MBR) and report better trainslation quality across language pairs and metrics. However, their target is overall quality and not terminology. Savage et al. (2025) find that DPO is better than SFT in complex tastks such as pattern recognition int hte context of clinical NLP. A benchmarking study of DPO for medical vision-language models reaches a similar conclusion. Gains over SFT were inconsistent across tasks and models, and reliable improvement required domain-specific preference construction (Kim et al., 2026). Taken together with the theoretical insights on the supervised baseline and the limits imposed by the reference policy, these medical results point to a key expectation tested in this thesis. Preference optimization is likely to help with medical terminology only when the preference pairs are explicitly designed to

encode correct terminology, and only when applied on top of a strong supervised baseline and not simply by using the method itself.

A common theme across all of this is that the improvements may not appear in the standard metrics, which means the evaluation challenge is tied directly to the research question itself. Standard automatic metrics were made to measure overall similarity to a reference or overall quality. They are not made to track terminological accuracy. A translation can still be fluent and more or less correct while using the informal term instead of the correct one and still get a high score from those metrics. But this error is exactly what matters in clinical settings (Gaona et al., 2023). If a metric cannot differentiate correct and incorrect terminology, the real gain from preference training could be lost inside a small change in the overall score, causing for the method to be judged wrongly useless. Newer evaluation frameworks that combine several model-based quality judgements have been proposed for more reliable scores (Junczys-Dowmunt, 2025). But also these frameworks aim to general quality and not to terminology specifically. Comparing SFT and DPO with automatic metrics alone is not enough. For this reason, automatic metrics need to be combined with manual terminology analysis. Since this is not feasible for this thesis, it is rather a manual inspection than a full analysis.

The most direct basis of this thesis is in the work of Rios (2025), on which the present methodology builds. Rios (2025) adapts Tower-7B, Llama-3-8B and FLAN-T5 to the medical domain using parameter-efficient instruction tuning with Quantized Low-Rank Adapter (QLoRA) (Dettmers et al., 2023), trained on a 60k-segment split of the EMEA corpus across English-Spanish, English-German and English-Romanian. Terminology is introduced into the training signal by matching term pairs from the IATE health domain against each parallel segment, and any matched pairs are inserted into the prompt as a glossary before the source text. The results show that this form of instruction tuning improves both general translation quality, measured with BLEU, chrF, and COMET, and terminology accuracy, measured as exact match against reference terms. Tower-7B in particular benefits from the procedure and, for English-Spanish and English-German, reaches the highest scores on both general and terminological measures. Two further findings are directly relevant. First, an automatic error analysis with XCOMET shows that instruction tuning reduces critical and major errors, meaning the gain is not only a metric shift but a reduction in the kind of errors that matter clinically. Second, terminological accuracy is computed as exact match against the reference. Rios (2025) flags this explicitly as a limitation, since exact matches ignore the context and may miss terminologically valid variants.

This Master's thesis builds on Rios (2025) and examines whether, given that supervised instruction tuning with terminology-aware prompts already improves terminology accuracy, applying DPO on top of such an instruction-tuned model

can get an additional, terminology-specific improvement. The setup is similar to that of Rios (2025) using the same base model family (Tower), an EMEA training corpus and the same parameter-efficient adaption strategy, being LoRA. It adds a preference optimisation stage in which preference pairs are constructed to encode terminological correctness rather than general quality. In this sense, the SFT baseline of this thesis is not an arbitrary baseline but the method established by Rios (2025), and the comparison with DPO is a test of whether preference-based training adds value on top of an already terminology-aware supervised stage.

3 Methodology

In the context of this Master's thesis, pre-training and transfer learning form the basis for adapting Tower, Qwen and Ministral to medical English-to-French translation. The base models have already acquired general translation or multilingual capabilities through large-scale pre-training, and the following the fine-tuning steps described in Chapter 2 build on these pre-trained weights rather than training a new model from scratch. This makes the assumptions and findings about transfer learning discussed above directly relevant to the experimental setup of this work. The code underlying the experiments has been published on GitHub[1] to make them reproducible and transparent.

This Chapter contains the methodology used for the experiments about terminology-aware fine-tuning for English-to-French medical MT researched in this Master's Thesis. It includes the research design, system architecture, data and its preparation, the choice of models, the training procedure for SFT and DPO fine-tuning, the terminology-pipeline and the evaluation framework. The experiments will be described in detail to make them reproducible. The experiments are done using LLM-based translation, so they are all using decoder-only LLMs instead of encoder-decoder translation models. The models are injected with medical glossary entries into the training prompt and are then compared against a baseline.

To compare the models against each other, for each model there is a supervised baseline that is trained first and also forming the base for the subsequent preference-optimization process. The data underlying the experiments stays the same throughout the models and they are evaluated using the same metrics. In the following sections the research design, the system architecture, the data and its preprocessing, the model choice, the configuration of the fine-tuning methods and the evaluation framework are described.

The empirical research is designed to be comparable. The central research question is whether DPO improves over SFT for medical MT. For this, the experiments also look at whether terminolgogy-enriched SFT training variants and DPO on terminology-aware preference pairs can improve the model performance further, particularly on sentences that contain such medical terms. To answer the question, the same fine-tuning and evaluation criteria are applied to the three decoder-only LLMs.

[1] https://github.com/sandrapoer/medical-mt-dpo

The research design also connects to the theoretical background of the previous chapters and is motivated by the findings discussed there. For the fine-tuning stage, this means that one variable at a time is isolated instead of maximizing a single system in order for each methodological choice to be evaluated independently. Every model is first evaluated on its untuned zero-shot form as the baseline and then goes through the same order of processes, starting with SFT, a terminology-enriched SFT variant, and a terminology-aware DPO variant. Those variants are built on the supervised model. This strict order makes it possible to attribute any observed differences to each specific step that was changed.

A clear separation between the training the the final evaluation data is maintained. The EMEA corpus is used for all training and validation runs. The TICO-19 benchmark is used only for final testing. Medical terminology is extracted with scispaCy and resolved throuhg UMLS producing a bilingual terminology cache used to annotate a part of the data. Both corpora include data from the medical domain and are available in multiple languages. The full dataset statistics and preprocessing details are presented in Section 3.2

3.1 Pipeline

The pipeline built to run the experiments named above starts with input of a raw corpus and ends with the final evaluation. The whole process, depicted in detail in Figure 5, starts with a data stage that performs heuristic filtering, semantic filtering and terminology injection. After this step a SFT trainer is used to fine-tune each model using QLoRA. Then the hypothesis generator produces eight translation candidates for each source sentence using the fine-tuned model. Those are then ranked through translation quality and medical terminology coverage and after this process formatted into preference pairs. The DPO stage fine-tunes the merged SFT model on these pairs followed by the computation of the automatic metrics, while also storing the raw hypotheses for later inspection. Finally, a separate analysis involves statistical significance and terminology accuracy. The untuned zero-shot models are evaluated as the baseline.

In detail, the raw EMEA corpus is filtered and split, then converted into a chat-style format with two prompt variants being a plain translation instruction prompt or said prompt including a terminological glossary. The SFT model is trained on this data and merged. from the merged model, candidate translations are generated, ranked and formatted into preference pairs on which the DPO model is trained. Both the SFT and the DPO models are finally evaluated on the EMEA test set as well as on the TICO-19 benchmark. The latest step is followed by a significance and terminological accuracy analysis. Two major architectural decision of this pipeline, namely the merging of the supervised adapter into the base model

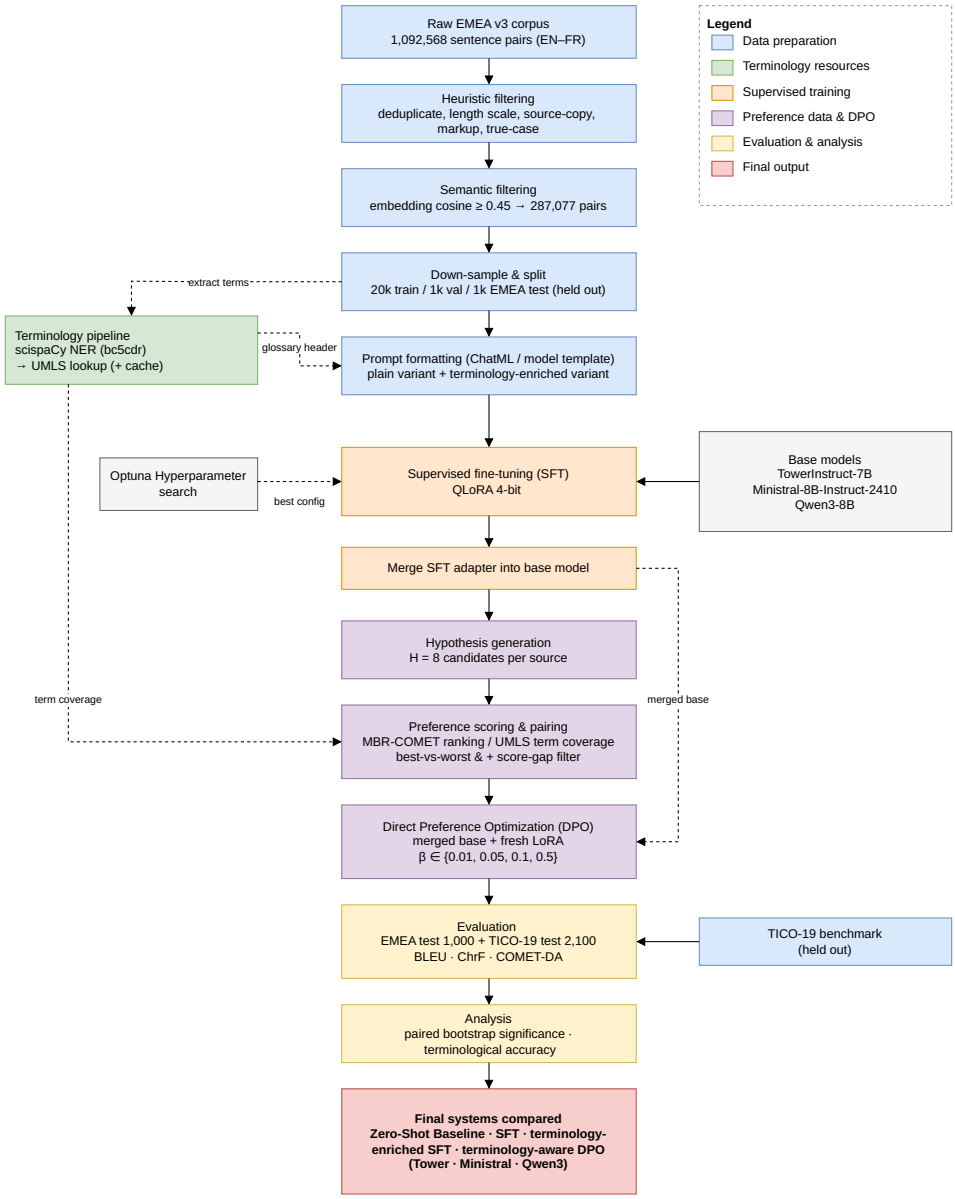

Figure 5: Methodology pipeline for terminology-aware medical MT. The EMEA corpus is filtered and split, terminology is injected into the prompts, and each model is first fine-tuned wiht SFT and then optimised with DPO on preference pairs. All systems are evaluated on the EMEA test set and the TICO-19 benchmark. The final outcome includes three variants per model, namely the SFT baseline, the term-enriched SFT variant and the terminology-aware DPO variant.

before DPO and the generation of the preference candidates, are discussed in detail a following Section 3.4.2.

3.2 Dataset Description and Preprocessing

The experiments are using two different datasets that also have different roles in the process. The EMEA parallel corpus of medical documents in English and French is used for training and validation, as well as an additional test set of the models. The corpus is distributed by OPUS (Tiedemann, 2012). The corpus used as benchmark and test set is TICO-19 (Anastasopoulos et al., 2020), a translation benchmark created for the COVID-19. The EMEA corpus is used for all training and validation steps and the TICO-19 corpus is reserved entirely for the final testing stage.

The inspection on TICO-19 resulted in 2.100 pairs. Corpus size and the fact that it is a translation benchmark in medical MT are the main reasons for using it as a test-only dataset and not to for training or validation purposes. The EMEA corpus was selected for training because it is large, it contains medical data and it can be filtered towards terminology. Sample size was one another reason why this specific version of the EMEA corpus was chosen.

Before any model-based filtering, heuristic filters were applied to the EMEA corpus in order to delete empty segments, duplicates and sentence pairs where the length of source and target had a large difference in terms of range. Here, a pair is treated as too long if one sentence exceeds 200 words or is more than twice the word length of the counterpart, and as too short if either side has three characters or fewer. The re-check is needed after the markup stripping because that can again leave empty cells that need to be removed. The purpose of these filters is to get rid of misaligned or degenerate pairs before training in order to not add additional noise to the training and to prevent repeating things multiple times. The heuristic filtering also removed pairs in which the source had simply been copied to the target, at the same time the text was true-cased and the rows were shuffled. The semantic filter applied afterwards embeds the source and target sentences while pairs with a low cross-lingual similarity were dropped if their embedding cosine similarity did not reach at least 0.45. The goal of this step is to keep only pairs that are truly translation of each others, because length filters alone cannot guarantee that. It is done to have in the final step a clean training signal, even though the corpus might be a bit smaller in size. Table 2 shows the effect of each filtering step. From the final set of filtered and styled data, 20.000 sentence pairs were selected for training and 1.000 pairs for validation, as well as a separate test set with 1.000 pairs that is reserved for the final stage.

After filtering, the pairs were reformatted into a chat-style message format which varies between the models. The templates for the three models are shown below,

Table 2: Results of remaining sentence pairs after applying heuristic and semantic filter to the EMEA v3 (Tiedemann, 2012) corpus.

Step	Rows removed	Rows remaining
Start (EMEA v3)	–	1.092.568
Empty cells deleted	121	1.092.447
Duplicates deleted	719.232	373.215
Source-copied deleted	21.503	351.712
Too-long deleted	38.399	313.313
Too-short deleted	6.427	306.886
After re-check and shuffle	82	306.804
After semantic filter (≥ 0.45)	19.727	287.077

where {`prompt`} is the instruction-and-source user turn and {`translation`} is the French target.

TowerInstruct-7B - ChatML template

```
<|im_start|>user
{prompt}<|im_end|>
<|im_start|>assistant
{translation}<|im_end|>
```

Ministral-8B-Instruct - instruction template

```
<s>[INST] {prompt}[/INST] {translation}</s>
```

Qwen3-8B - ChatML template (reasoning disabled)

```
<|im_start|>user
{prompt}<|im_end|>
<|im_start|>assistant
{translation}<|im_end|>
```

It contains a user and assistant turn written in JSON while preserving non-ASCII characters such as French accents. The final JSON file contains one object per line.

3.2.1 Terminology Annotation and Extraction

Another preprocessing decision involved filtering the data containing medical terminology of the domain. The reason is that training data needs to be relevant

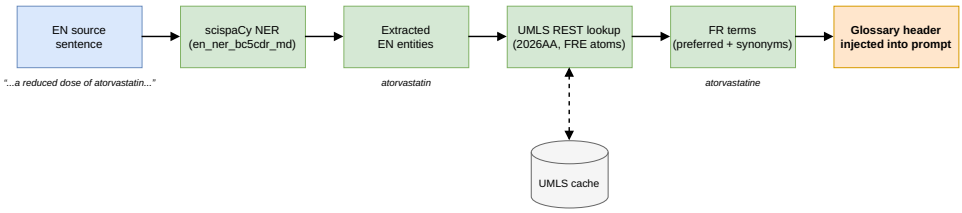

Figure 6: Illustration of the terminology annotation and extraction pipeline

to the domain it is evaluated in. This is done through a medical term extraction step that combines a clinical named-entity recogniser with a medical terminology database, working as follows. The clinical named entity recognition toolkit from scispaCy (Neumann et al., 2019) is used to extract medical terms in English. It is trained to recognise various chemical and disease entities, which works well for the selected corpus. Each extracted term is then looked up in the UMLS (Bodenreider, 2004) through its public web interface, which returns the French preferred terms and its synonyms. The lookup retrieves results page by page and stops at the first French match. Terms that have already been looked up are cached so that later lookups do not repeat the same query. The decision to involve UMLS in the extraction phase was made because exact term matching with other open source glossaries does not capture the synonyms that UMLS provides.

The extraction pipeline can be illustrated using a single soruce sentence from the training data. Given the English source "*In this situation a reduced dose of atorvastatin should be considered.*", the named-entity recogniser identifies the medical entity `atorvastatin`. This entity is then looked up through the UMLS API finding the French atom and returning the term pair `atorvastatin` → `atorvastatine`. This term pair is injected into the prompt as a glossary header for this specific sentence. If the UMLS lookup resolves in multiple French atoms, all synonyms are stored in the cache. For example the entity `diabetes` resolves to `diabète` together with variants such as `diabète sucré`. Figure 6 shows the full pipeline.

The complete lookup for the corpus produces an UMLS cache of about 5.000 English Terms mapped to about 7.000 French unique terms.

3.2.2 Terminology-Aware Instruction Data

The extracted terms are used to build an instruction-tuning variant that includes the terminology in the prompt, forming a new training set and following the work of Rios (2025). If a sentence contains a medical term, a glossary header is added at the beginning of the prompt listing the expected English-to-French term pairs, asking the model to follow the provided glossary. If no terms are found in one sentence, the original plain prompt is used. The prompt with included glossary is shown below.

```
Glossaries:
"atorvastatin" -> "atorvastatine"

Translate the English source text to French following the provided
translation glossaries. Rules: output strictly the French
translation.
Do NOT repeat the English text. Do NOT include labels like "English"
or "French".
Do NOT add quotation marks, explanations or any extra text.
Now translate the following:
English: In this situation a reduced dose of atorvastatin should be
considered.
French:
```

In the training set about 7.000 of 20.000 sentences received a glossary header which is roughly 36%. This term-enriched variant makes it possible to evaluate separately on two subsets, one containing a glossary header and the other containing the plain translation instruction prompts. The held-out test sets contain comparable proportions where 357 out of 1.000 EMEA test sentences ($\approx 36\%$) and 591 out of 2.100 TICO-19 test sentences ($\approx 28\%$) received a header. The same terminology step is later used to score translation candidates in the hypotheses during the preference-pair construction, as describes in Subsection 3.4.2.

3.3 Model Selection

The three models fine-tuned in the experiments are the following decoder-only LLMs: TowerInstruct-7B (Rei et al., 2025; Unbabel, 2024), Ministral-8B-Instruct-2410 (Liu et al., 2026; Mistral AI, 2024) and Qwen3-8B (Yang et al., 2025; Qwen Team, 2025). They were chosen because they are all designated multilingual models, with Tower even being a LLM specialised in translation, and because they are

of similar size. All three are autoregressive, decoder-only transformer models fine-tuned with PEFT.

Tower (Rei et al., 2025; Unbabel, 2024) is a translation model with seven billion parameters. It is build on a general open-weight language model base and trained on multilingual translation-related tasks. It was chosen for the experiments because the main task is translation and because other work builds on the same model family (Rios, 2025). It is know for its strong outputs regarding translation-related tasks and can be adapted easily further. It is a bit smaller than the other two models but it is trained on the specific task. It does not support as many languages as the other but it supports the main languages needed in the experiments, English and French.

Ministral (Liu et al., 2026; Mistral AI, 2024) is a multilingual general-purpose model. It is instruction tuned and has eight billion parameters, released by Mistral AI. The model is not specialised on translation tasks but is included in the experiments to check whether the methods generalise well even though the model is not specialised on translation. Regarding its general-purpose base one might think that it is not capable to behave as well in the translation domain as a specialised model.

Qwen (Yang et al., 2025; Qwen Team, 2025) is another multilingual general-purpose model that is instruction tuned and contains around eight billion parameters. The model has an optional "thinking" mode which is disabled during generation in the experiments because the model should produce translations directly without the internal reasoning block. Like Ministral, it is not specialised on translation directly.

3.4 Experimental Setup

All models were trained with PEFT. The low-rank adapters were trained on top of a 4-bit quantised base model with QLoRA (Dettmers et al., 2023). It combines 4-bit quantisation of the frozen base model with trainable low-rank adapters (Hu et al., 2021). This step was necessary because it keeps the number of trainable parameters small while the training ran on GPUs with limited available memory. The 4-bit configuration used the NF4 quantisation type with double quantisation and a bf16 compute dtype, and the low-rank adapters were applied to a model-specific the attention query, key, value and output projection. The same PEFT approach is used for SFT and DPO. Training ran on a shared server with three NVIDIA A100 80GB GPUs.

3.4.1 SFT Configuration

Each model underwent SFT on the formatted EMEA training set with an instruction prompt style. The prompt asks to model to translate the English source into French and the French sentence is the completion. During training, each model's own chat template is applied to the messages discussed in Section 3.2. Tower uses a ChatML template, Ministral uses the instruction template delimited by `[INST]` and `[/INST]` and Qwen is used with a disabled reasoning mode. The loss was computed on the completion, while the prompt tokens were masked so that the training signal comes only from the French target. A full training instance consists of an instruction and a source as the user turn and the French translation as the assistant turn, which is used to compute the loss. During training, the model's chat template is added around these messages.

During the training, models where evaluated on validation loss only because the installed version of the trainer did not support generation-based metrics during the training. Instead, the translation metrics were computed afterwards in a separate evaluation script. Hyperparamters were tuned with a small automated search using the Optuna framework (Akiba et al., 2019). The validation loss was minimised over a fixed numbers of trials per model. The learning rate ranged from 1×10^{-5} to 5×10^{-4} on a logarithmic scale, the warmup steps ranged from 50 to 200, the low-rank adapter rank was chosen from $\{8, 16, 32\}$, and the low-rank adapter scaling was chosen from $\{16, 32, 64\}$. Each model was tuned over about 15 trials, with one epoch per trial. The best configuration for all three models used a rank of 8 and a scaling of 64. The best per-model learning rates, warmup settings, and validation losses are reported in Chapter 4. For the models the low-rank adapters were applied to the attention query, key, value and output projection matrices. For the Tower model the search was additionally run with two different trainer implementations, and the one that achieved the lower validation loss was adopted for the final runs, this comparison is also reported in Chapter 4. Checkpoints were saved at fixed intervals during training, and the checkpoint used for each model was selected by its BLEU score on the validation set rather than by validation loss alone, because a lower loss does not necessarily correspond to a better translation.

The final supervised configuration is summarised in Table 3. The architecture-level settings were shared across all three models: QLoRA with 4-bit NF4 quantisation, double quantisation and a bf16 compute dtype, low-rank adapters of rank 8 and scaling 64 with a dropout of 0.05, a maximum sequence length of 512 tokens, a per-device batch size of 8 with a gradient-accumulation factor of 4. three training epochs, and a cosine learning-rate schedule.

Table 3: Final SFT configuration. Architecture settings (QLoRA 4-bit NF4, rank 8, scaling 64, dropout 0.05, max. length 512, effective batch 32, 3 epochs, cosine schedule) are shared across all models. The learning rate, warmup, and adapted modules are adjusted per model.

Shared QLoRA and training settings	
Quantisation	4-bit NF4, double quantisation
Compute dtype	bf16
LoRA rank (r)	8
LoRA scaling (α)	64
LoRA dropout	0.05
Max. sequence length	512
Per-device batch size	8
Gradient accumulation	4 (effective batch 32)
Epochs	3
Learning-rate schedule	cosine
Model-specific settings (learning rate / warmup / adapted modules)	
Tower	4.018×10^{-4} / 50 / q,k,v,o projections
Ministral	9.419×10^{-5} / 60 / `all-linear`
Qwen3	4.018×10^{-4} / 50 / `all-linear`

3.4.2 DPO Configuration

After SFT each model was further trained with DPO (Rafailov et al., 2023). In this step the models are optimised directly from pairs of preferred and dispreferred outputs under a Bradley-Terry preference model without training a separate reward model. This was done using the preference optimisation trainer from the TRL library.

A specific process is required to make DPO train correctly on the top of a PEFT supervised model. Training on an unmerged supervised adapter caused the training loss to collapse and stop decreasing. This happens because the supervised low-rank adapter had not been merged into the base model before a second adapter was stacked on top of it. So the supervised adapter is merged into the base model and saved as a merged model, then the clean merged model is loaded and attached with a fresh low-rank adapter for DPO. The reference model is left unspecified so that the trainer derives the reference from the merged state. This procedure was applied to all DPO runs.

The preference strength parameter β is treated as the main hyperparameter of the DPO stage and is tuned over the values 0.01, 0.05, 0.1, and 0.5. It controls the strength of the KL constraint that keeps the optimised policy close to the reference policy. Each β run was re-initialised from the clean merged SFT model, so that

the runs are independent and do not accumulate on top of one another. The value $\beta = 0.01$ was selected for all three models because it returned the best results.

The preferences pairs were constructed from generated translation candidates. For each source sentence, the supervised model generated 8 candidates referred to as $H = 8$ with a sampling temperature of 0.7. The value $H = 8$ was chosen because a reference method based on MBR decoding uses up to 32 candidates but this was too expensive and Yang et al. (2024) report that smaller values are justified. Two complementary ways of turning the hypotheses into preference pairs were implemented. The First reproduces the reference method and ranks the candidates by MBR scoring. In this case, each candidate is scored against all other candidates without using a reference. The utility function in this case is COMET-DA. Additionally, the candidates are ranked by how well they cover the expected French medical terms, using the UMLS pipeline described in Section 3.2. In this case, the preference signal is explicitly terminology-aware. From the final ranked hypotheses a schema of best-worse pairs was built.

A score-gap filter removes pairs whose preferred and dispreferred candidates have almost identical scores so that the model is only trained on contrasts that are informative. A pair is discarded if the MBR score gap between the chosen and rejected candidate is below 0.05. After applying the score-gap filter, the terminology-ranked set used for the final DPO run contained about ≈ 4.000 for Tower, ≈ 3.000 for Ministral and ≈ 4.000 for Qwen.

The preference pairs must be generated from each model's own output because the preferred and dispreferred signal has to match the output distribution of the own model's policy. Reusing one model's candidates to train another model is found to degrade performance badly, for exactly this reason. Therefore, the candidate generation is repeated separately for each model. The construction of a preference pair can be shown using a real example form the Tower candidate set shown in Table 4. For a source sentence describing the management of *HIV infection*, the chosen and rejected candidates of the preference pair differ a lot.

The example in Table 4 shows that the variation in the translation is terminological and not only stylistic. Both translation candidates are fluent but they diverge on the medical term. The chosen candidate produces the correct French term *infection par le VIH*, whereas the rejected candidate fails to translate it properly and returns *infectie-virales*. In this case, the preference signal rewards exactly the part that the terminology-aware approach targets, namely the production of correct terms rather than degenerated, omitted or more casual ones.

Table 4: A preference pair whose chosen and rejected candidates differ in the translation of the medical term *HIV infection*. The chosen candidate produces the correct French term *infection par le VIH*, whereas the rejected candidate fails to translate it, producing the incorrect form *infectie-virales*. MBR score gap 0.15.

Source	Treatment with Ziagen should be prescribed by a doctor who has experience in the management of HIV infection.
Chosen	Ziagen doit être délivré uniquement par un médecin expérimenté dans la prise en charge de l'**infection par le VIH**.
Rejected	Ziagen doit être délivré par un médecin expérimenté dans la prise en charge de l'**infectie-virales**.

3.5 Evaluation Framework

The quality of the translation is measured with automatic metrics namely BLEU (Papineni et al., 2002), chrF (Popović, 2015) and COMET-DA (Rei et al., 2020). While BLEU and chrF cover surface-level overlap with the reference at the word and character level respectively, COMET-DA surpasses the surface level. The choice of metrics was made to get a more round idea of the output that one single metric could not provide. Terminology-aware translations can be semantically closer even though they overlap less with the reference. COMET-DA is treated as the primary metric and the other metric are used alongside it. As already mentioned, all translation metrics are computed after training in a separate evaluation script, using greedy decoding so that the generation is reproducible. The untuned zero-shot baslines, the supervised models and the preference-optimised models are all evaluated under the same procedure.

The structure remains the same for each model. The untuned zero-shot model serves as the baseline. The plain supervised model which is compared against this baseline, the terminology-enriched supervised model is compared against the plain supervised model and the preference-optimised model is compared against the supervised model on which it is built. Through theses choices, the comparisons of the models are isolated and each design choice can be validated separately.

The statistical significance is computed with paired bootstrap resampling as provided by the sacreBLEU command-line tool (Post, 2018). It is using 1.000 resamples and reports 95% confidence intervals and p-values. The resampling is performed within each model, comparing each model's own SFT and DPO variants. A comparison across models of different baseline quality is not meaningful in this case. The comparison follows the order "SFT baseline → terminology-enriched

SFT → terminology-aware DPO”.

In addition to the general translation metrics, a dedicated terminological accuracy evaluation measures how well the systems follow the provided glossaries. For each test sentence that contained a glossary header in the prompt, the expected French terms are searched for in the hypothesis while using case-insensitive substring matching. This tolerates the French morphological variants and does not require and additional stemmer. Three quantities are reported, namely recall, the share of injected French terms that appear in the hypothesis, precision, the share of all UMLS French terms found in the hypothesis that belong to the injected set, and their harmonic mean, the F1 score. Recall is considered the more interpretable measure, because the precision denominator scans the entire UMLS French vocabulary and is inflated by common words, which systematically lowers precision.

The final step of the experiments is the testing on held-out data. Each system is tested on the reserved 1.000-pair EMEA test set, as well as on the designated TICO-19 benchmark. The TICO-19 set contains 2.100 sentence pairs.

4 Results

This chapter summarises the results of the experiments described in Chapter 3. The results are presented objectively, while most of their interpretation follows in Chapter 5. The chapter first reports the automatic evaluation results, that cover the supervised hyperparameter search, the validation results of the supervised and preference-optimised models, the final held-out test results on the EMEA test set and the TICO-19 benchmark, the terminological accuracy evaluation and the statistical significance analysis. Followed by a summary of the qualitative observations from an error analysis and a report of the training-efficiency observations. The untuned zero-shot model of each model family serves as the baseline, against which the SFT variants and the terminology-aware preference-optimized model are compared.

4.1 Automatic Evaluation Results

For the Supervised Fine-Tuning (SFT) stage in the pipeline, a hyperparameter search with Optuna was done. For every model, a low-rank adapter rank of 8 and a scaling of 64 were selected, while the learning rate, warmup and validation loss differed per model. For the Tower model the search was run with two trainer implementations, namely HFTrainer and TRLTrainer. The TRL implementation achieved the lower validation loss (0.4308 versus 0.4524) and was adopted for all final Tower runs, and the corresponding configuration is the one shown in Table 5.

Table 5: Best configuration per model from the Optuna search (rank = 8, scaling = 64 in all cases). Validation loss is the minimised objective.

Model	Val. loss	Learning rate	Warmup
Tower	0.431	4.018×10^{-4}	50
Ministral	0.530	9.419×10^{-5}	60
Qwen	0.447	4.018×10^{-4}	50

Table 6 reports the validation results of the supervised models at their selected checkpoints. The three models differ significantly on the surface metrics. Qwen reaches a BLEU above 51 and a COMET-DA of 0.874, whereas Tower and Ministral remain below a BLEU of 8 and a COMET-DA of around 0.35 to 0.38. A review

of the outputs confirmed that the gap is genuine and not caused by a flaw in the evaluation. The final test results reported in Tables 9 and 10 are computed in the held-out test sets and differ from these validation scores.

Table 6: Supervised models validation results at the selected checkpoint per model.

Model	Checkpoint	BLEU	chrF	COMET-DA
Tower	1250	7.46	40.94	0.349
Ministral	1875	6.92	40.29	0.382
Qwen	1875	51.24	71.82	0.874

To find whether the injected glossary header returned better results, the Tower validation set was split into two subsets. One containing only sentence with a glossary header and one containing only sentence with the plain translation instruction. Table 7 reports the comparison. The model that got instruction with injected glossary headers improved the COMET-DA score by 0.086 over the baseline. It also improves COMET-DA on the translation-only instruction prompt. The surface metrics return less consistent results. BLEU improves on both subsets, while chrF was mostly unchanged. For the other tow models the terminology-enriched supervised model produced only small changes on the validation set that were in both cases close to the corresponding baseline.

Table 7: Tower supervised models on the split validation subsets.

Model	Prompt	BLEU	chrF	COMET-DA-DA
SFT	terms-injected	9.74	47.84	0.351
SFT terms	terms-injected	10.98	47.33	0.437
SFT	plain-only	6.22	36.34	0.344
SFT terms	plain-only	6.92	36.75	0.388

Table 8 reports the Tower preference-optimised models on the validation set across the four β-values, together with the supervised model. DPO improves the surface metrics slightly, BLEU rises from 7.46 to about 7.84, but the primary semantic metric, COMET-DA, does not improve and is marginally lower than the supervised baseline. The choice of β has little effect within the tested range.

A similar pattern is found for the other two models on the validation set when valid, model-matched preference pairs are used. For Ministral, the DPO model ($\beta = 0.01$) reached 6.37/37.37/0.392, which is essentially flat relative to its supervised baseline. For Qwen, the preference-optimised models degraded clearly in comparison to the very strong supervised model, reaching about 37.9/69.9/0.717, and one configuration collapsed entirely ($\beta = 0.5$, COMET-DA 0.175). For Tower, a

Table 8: Tower supervised model and preference-optimised models across β values.

Model	BLEU	chrF	COMET-DA-DA
SFT (ckpt-1250)	7.46	40.94	0.349
DPO $\beta = 0.01$	7.84	41.03	0.346
DPO $\beta = 0.05$	7.84	41.14	0.344
DPO $\beta = 0.1$	7.83	41.02	0.347
DPO $\beta = 0.5$	7.83	41.03	0.347

terminology-aware DPO configuration also collapsed at $\beta = 0.1$ (BLEU and chrF fell to 0).

The final testing was run on the held-out EMEA test set and on the TICO-19 benchmark. Four systems per model were tested, namely the untuned zero-shot baseline, the supervised model, the terminology-enriched supervised model, and the terminology-aware preference-optimised model. Tables 9 and 10 report the results.

Table 9: Final held-out test results on the EMEA test set.

System	BLEU	chrF	COMET-DA-DA
Tower baseline	28.6	51.8	0.748
Tower SFT	**38.2**	**63.9**	**0.839**
Tower SFT + terms	14.2	52.8	0.509
Tower DPO	14.4	52.8	0.510
Qwen3 baseline	36.6	62.8	0.840
Qwen3 SFT	**50.1**	**71.8**	**0.869**
Qwen3 SFT + terms	49.1	70.6	0.865
Qwen3 DPO	48.3	69.9	0.863
Ministral baseline	**35.6**	**62.1**	**0.830**
Ministral SFT	12.7	53.1	0.460
Ministral SFT + terms	12.7	51.9	0.480
Ministral DPO	13.0	52.0	0.484

Several observations can be drawn directly from the Tables 9 and 10. First, in comparison to the untuned zero-shot baseline the SFT has a mixed effect because it raised COMET-DA on the EMEA test set, that can be considered *in-domain* as it is the same corpus the models are trained on, for Tower and Qwen, but it lowers COMET-DA on the TICO-19 benchmark, that can be considered *out-of-domain*, for all models. Second, the terminology enrichment of the supervised model through the injected glossary headers does not improve the models consistently. For Tower, it reduces minimally the EMEA COMET-DA score while it increases minimally the TICO-19 score. Also for the other models there are only small to no effects. Third,

Table 10: Final held-out test results on the TICO-19 benchmark.

System	BLEU	chrF	COMET-DA-DA
Tower baseline	**33.4**	**57.1**	**0.767**
Tower SFT	13.1	51.3	0.476
Tower SFT + terms	13.7	51.0	0.524
Tower DPO	14.0	51.2	0.525
Qwen3 baseline	**36.4**	**61.6**	**0.807**
Qwen3 SFT	30.2	55.1	0.780
Qwen3 SFT + terms	29.9	55.3	0.777
Qwen3 DPO	29.7	55.3	0.777
Ministral baseline	**35.7**	**60.6**	**0.797**
Ministral SFT	12.2	50.9	0.479
Ministral SFT + terms	12.3	50.5	0.499
Ministral DPO	12.4	50.4	0.498

the preference-optimised models does not improve over the terminology-enriched supervised model at all. COMET-DA differs only minimally for all models on both test sets. Fourth, the scores are in general much higher for Qwen than for the other two models.

Table 11 reports the dedicated terminology-accuracy evaluation on the sentences, where the prompt was injected with glossary headers. Recall is uniformly low with about 22% on TICO-19 and about 42% on EMEA, and is nearly identical across all systems. The difference in recall and precision between the supervised and the preference-optimised model is at maximum about half a percentage, which can be considerate as noise. The higher precision of Qwen reflects its better overall translation quality rather than better glossary-following. As discussed in Chapter 3, the precision figures should be read comparatively because the precision denominator scans the entire UMLS French vocabulary and is therefore inflated by common medical words.

The paired bootstrap significance tests on BLEU and chrF, following the order "supervised model → term-enriched supervised model → terminology-aware DPO model", are computed within each model. For Tower, the terminology-enriched supervised model is significantly worse on the surface metrics, even though it is better on COMET-DA. This divergence between the surface metrics and the semantic metric is the clearest pattern in the significance results. The only significant positive preference-optimization effect anywhere is the comparison of terms-enriched supervised Tower model to the terminology-aware DPO Tower model on TICO-19 chrF ($p = 0.014$). For Qwen nothing is significant in the expected direction. For Ministral, the terminology-enriched supervised model is significantly better only

Table 11: Terminology accuracy on the terminology-containing test sentences. N is the number of evaluated sentences.

Test	System	Recall (%)	Prec. (%)	F1 (%)	N
TICO-19	Tower SFT terms	22.87	8.28	12.16	591
TICO-19	Tower DPO terms	22.64	8.29	12.14	591
TICO-19	Qwen3 SFT terms	21.94	10.57	14.27	591
TICO-19	Qwen3 DPO terms	21.35	10.39	13.98	591
TICO-19	Ministral SFT terms	22.52	5.86	9.31	591
TICO-19	Ministral DPO terms	22.40	5.90	9.34	591
EMEA	Tower SFT terms	41.73	15.74	22.86	357
EMEA	Tower DPO terms	42.03	16.27	23.45	357
EMEA	Qwen3 SFT terms	41.58	24.31	30.68	357
EMEA	Qwen3 DPO terms	41.58	24.31	30.68	357
EMEA	Ministral SFT terms	42.03	12.00	18.67	357
EMEA	Ministral DPO terms	42.03	12.01	18.68	357

on the EMEA BLEU, while it is worse on TICO-19 and on chrF.

4.2 Human Evaluation Results

A manual qualitative inspection complements the results of the automatic metrics in Section 4.1. The inspection is reference-based because it is done only by one single annotator. Each selected hypothesis is read against its source and the human reference and judged on semantic accuracy, fluency, completeness, terminology, adherence, and error type. To support these judgements with quantitative evidence, every inspected sentence is reported with its sentence-level BLEU, chrF and COMET-DA computed with the same tools as the automatic evaluation.

The inspection covers two stages. The primary object is the final system outputs on the held-out EMEA and TICO-19 test sets. A complementary inspection of the Tower preference-pair candidate generation on the training sources is used to understand the behaviour of the preference data rather than of the final systems, and examples from this second stage are labelled accordingly.

A qualitative finding that stood out on the test set is that the terminology systems often produce translation that are fluent and semantically adequate but lexically far from the reference. This causes COMET-DA to still rank them high while BLEU does not. A clear example is Qwen on EMEA shown in Table 12 and discussed in Section 4.3. The hypothesis is a complete and accurate translation but the sentence BLEU is less than half of the test-set mean. During the inspection, a high number of translations with low BLEU and high COMET-DA were found.

This supports treating COMET-DA as the primary metric.

Another observation shows that the strong Qwen baseline and its preference-optimised version, as well as term-enriched supervised model were often nearly identical in the sentence level. For inspected sentences the SFT and DPO outputs were roughly the same. Additionally, failures to follow the injected terminology were common, but they were genuine paraphrases of a correct glossary term.

The inspection of preference-pair candidates for Tower showed that MBR ranking usually selects the most reference-like candidate. It also revealed a data-quality issue. Some rows that were no full sentences survived the filtering stage and caused the candidates to hallucinate.

The manual inspection supports the findings of the quantitative results. First, the gap between the surface metrics and the semantic metric is real. Second, the preference optimisation brings little to no change for the strong models and often does not change anything at the surface level. Third, cases where the terminology was incorrect, the model often provided a true but not register-adequate paraphrase. The detailed error categories and examples follow in Section 4.3.

4.3 Error Analysis

This section provides an overview of the qualitative findings organised in error categories and its examples. It begins with the relationship between the metric and the translations and then investigates the glossary.

The qualitative observations from inspecting the stored hypotheses align with the quantitative results. The most noticeable pattern is the gap between the surface metrics and the semantic metric for Tower. The terminology-enriched model produces translations that score higher on COMET-DA but share fewer words with the reference, which lowers BLEU and chrF. This suggests that the terminology header shifts lexical choices in ways that are semantically correct but lexically different from the reference. The terminology-accuracy results add a second pattern. Although terminology-aware training improves Tower's overall quality, the recall of the injected glossary terms stays low (about 22% on TICO-19, about 42% on EMEA) and changes little between the supervised and preference-optimised models. This shows that the models do not reliably copy the provided terms even when they appear in the prompt.

A third pattern is the behaviour of the two general-purpose models. Qwen is at a quality ceiling, so neither the terminology header nor preference optimisation changes its scores enough. Ministral, by contrast, collapsed to a low translation quality across all of its variants. A double-BOS token issue and the chat template were both investigated and ruled out as the cause, so the collapse is recorded as a consistent negative result rather than a fixable bug.

A clear instance of the divergence between the surface metrics and the semantic metric is the translation in Table 12, produced by the Qwen terminology DPO system on the EMEA test set. The hypothesis is a fluent and semantically complete translation of the source, and COMET-DA scores it well above the corpus mean (0.931 versus 0.869). Its sentence-level BLEU, by contrast, is far below the corpus mean (20.33 versus 50.14), because the hypothesis diverges from the reference at several surface points at once. The prompt injected the UMLS term *diabète sucré* (the traditional full form of *diabetes mellitus*), but the model produced the contemporary clinical shorthand *diabète de type 2*, where the Roman numeral *type II* became the Arabic *type 2*. *Anti-diabétique* was expanded to *médicament antidiabétique*, while *dans le traitement du* was shortened to *pour traiter*, and *non insulino-dépendant* lost its hyphen. The central change is terminological because both *diabète sucré* and *diabète de type 2* are correct French, but the model preferred the modern term over the injected one, and BLEU penalises this lexical mismatch while COMET-DA recognises that meaning is preserved. This single example should not be over-interpreted because the terminology substitution is the most interpretable factor in the BLEU drop, but it is not the only one, since the numeral style, article choice and phrasing all contribute to the surface mismatch as well.

Table 12: A terminology substitution rewarded by COMET-DA but penalised by BLEU (Qwen terminology DPO, EMEA test set). Segment BLEU = 20.33 (corpus mean 50.14), segment COMET-DA = 0.931 (corpus mean 0.869). The injected glossary term was *diabetes mellitus* → *diabète sucré*.

Source	Glubrava tablets are an anti-diabetic medicine used to treat type 2 (non-insulin dependent) diabetes mellitus.
Reference	Glubrava comprimés est un antidiabétique utilisé dans le traitement du *diabète sucré* de type II (non insulino-dépendant).
Hypothesis	Glubrava comprimés est un médicament antidiabétique utilisé pour traiter le *diabète de type 2* (non insulinodépendant).

At some points, the UMLS lookup injected the wrong concept, which penalised the model. In each case the model produces the standard French term and the metric miss it. Table 13 shows examples.

For some examples, the glossary does not work at all meaning that the correctly injected terms were paraphrased anyway. This is also consistent with the low recall reported in Section 4.1. An example that represents this issue from the terminology candidate data is the source "`Dizziness (light-headed)`", for which the glossary injected *sensation vertigineuse* and the reference uses *sensation vertigineuses*. However, the model produces *sens de vertige* which is an acceptable paraphrase but does not reproduce the prescribed term.

Table 13: Case in which the UMLS lookup injected an incorrect French term, so that a correct model output is recorded as a terminology miss.

English source	**Injected UMLS term**	**Problem**
deaths	*heure de la mort*	wrong concept (“hour of death”)
overdose	*mauvaise utilisation des médicaments*	distant concept
ethanol	*méthylcarbinol*	archaic synonym, not used clinically

Finally, the inspection showed that even after the filtering, there are rows that did not include full sentences. They contained, for example, numeric fragments which cause the models to hallucinate. This is not essentially concerning for the final system but especially for the heuristic filtering stage, and shows that it did not in fact remove all degenerate rows. This pollutes the preference data and shows the need of a concrete data-cleaning improvement.

4.4 Training Efficiency Comparison

The use of QLoRA with 4-bit quantisation made it possible to fine-tune all three models on a single A100 GPU. The SFT stage of Tower took approximately one hour and fifty minutes. The most expensive part of the DPO pipeline was the generation of the candidate translations. For each sentence and each model separately, eight translation candidates were generated over the 20.000 training sentences. This step took several hours per model. The decision to use $H = 8$ instead of a higher number was motivated by exactly this cost and follows the finding that a smaller number of candidates captures already most of the benefit while the compute cost is just a fraction.

5 Discussion

This chapter interprets the results of Chapter 4 in relation to the research question and the theoretical background. First, SFT and DPO are compared directly, followed by the implications for medical MT. Furthermore, it discusses the limitations of the study and the challenges encountered throughout the process. Finally, the directions for future work are outlined.

5.1 SFT vs. DPO: Comparative Analysis

The central research qeustion of this thesis is whether DPO improves over SFT for medical MT. The results of the experiments give a clear answer for this question. DPO does not provide a consistent improvement over SFT for any of the three models on any of the used metrics. On the held-out test sets, the terminology-aware DPO models differ from the term-enriched SFT variants in a few thousands of COMET-DA. The only statistically significant positive effect is found for Tower on TICO-19 chrF, which is not the main metric the experiments rely on, given it is a surface-level metric. The choice of the preference-strength parameter β had only small effect on the tested range. In comparison to the untuned zero-shot baseline, the most visible improvement comes from the SFT stage, which changes the quality far more than the subsequent stages. For Qwen and Tower it improves COMET-DA in domain but falls below the baseline on TICO-19 for all three models.

The negative results differ across the three models and need therefore to be interpreted independently. For Qwen the supervised model is already very strong with a COMET-DA score of about 0.84. This case shows the ceiling effect that leaves little to no room for preference optimisation to add further value. The generated candidate translations are often similar and of high quality, meaning there are only small quality gaps between best and worse translation which makes it hard for the DPO model to distinguish properly. For Ministral the supervised models collapsed across all variants. Because of this, the preference pairs are too noisy and cannot provide a useful training signal. Tower, on the other hand, can be shown as an intermediate example, where the supervised baseline is not too strong nor too bad but the gaps between the translation candidates are still too small. Again, this limits how much the preference pairs can teach. In all three cases, the preference signal was either unnecessary, unreliable or too weak.

The effect of terminology enrichement on the supervised model does not show any gneeral positive results and depends on the test set used. As reported in Table 9 and 10, for Tower it lowers COMET-DA on the in-domain EMEA test set from 0.829 to 0.509, while it raises on the out-of-domain TICO-19 benchmark from 0.476 to 0.534. This case, however, is still scored lower than the baseline on the TICO-19 test set. For Qwen the changes remain scmall because of the mentioned ceiling effect. Ministral collapsed during SFT and leaves no room for improvement. DPO does not improve any supervised model variant.

The difference between the surface-level metrics and the semantic metrics is visible for Tower on the out-of-domain benchmark but is not visible as a global effect. On TICO-19 the term-enriched Tower model raise COMET-DA while the BLEU and chrF scores barely move. In Table 12 the same pattern is seen at the sentence level. This findings support reporting COMET-DA as the primary metric alongside the surface-level metrics, since in a terminology-focused setting and n-gram overlap metric can move independently to a semantic adequacy.

5.2 Implications for Medical MT

The most important takeaway is the connection between general translation quality and the correct use of terminology. Across all systems the recall of the injected glossary terms stays low. It remained around 22% on the TICO-19 test set and 42% on the EMEA test set. It was nearly identical across all systems including before and after DPO. On one hand, the models improve handling sentences with a lot of terminology. On the other hand, they still cannot consistently reproduce the specific terms they are given. This specific distinction might not matter for other fields, but in regulated clinical settings, where a specific approved term must be used, the glossary-injected instruction prompts are evidently not enough to guarantee adherence and even if the semantic metric improves, it does not guarantee that the terms are translated correctly.

A further implication concerns the evaluation of terminology-focused medical MT. As seen in the results in Chapter 4, the surface-level metrics and the semantic metric diverged for Tower on TICO-19. A study that reports only BLEU or chrF could misjudge a terminology-oriented change. This is the main reason to use a neural, semantic metric such as COMET-DA as the primary measure, while still having the surface-level metrics alongside it instead of replacing them. At the same time, all of the named metrics used in this study are not able to measure terminological correctness. This is made clear by the low recall. For the medical field, a dedicated terminology evaluation is a must have to capture what matters consistently.

Finally, the results depend strongly on the base model and the test domain. In

domain, SFT helps the stronger models Tower and Qwen on the EMEA test set. But out of domain, every fine-tuned system scores below the untuned zero-shot baselien on TICO-19. Minsitral, on the other hand, cannot be seen as an example for generalisation because its baseline run collapsed. This implies again that the starting model matters more than the preference optimisation step itself. A strong multilingual base generalises across the medical domain, while a weak base cannot be repaired by terminology-enrichment or DPO.

5.3 Limitations

This work inlcudes several limitations. First, the data, namely the EMEA corpus, contains sentences that recur across documents, which raise the possibility of overlap during training and validation splits and can be seen as a type of leakage, even after the initial filtering. The duplicates are often not identical but are closely related in their content and wording. The held-out TICO-19 benchmark is not affected by this but the used EMEA test set should be interpreted with this knowledge. This also explains the higher results of the EMEA test set in comparison to the TICO-19 benchmark.

A second limitation is the evaluation in the terminological accuracy measurement. The precision is computed against the entire UMLS French vocabulary, meaning that common medical words inflate the denominator and deflate the precision. Therefore, precision should be only interpreted comparatively, while recall is more interpretable. Additionally, the substring matching that is used to detect the terms tolerates morphological variations but it does not distinguish a correctly used term from a random string match. So the recall numbers should be seen as estimates and not as precise numbers.

Third, the limitation during the preference-data construction. the candidates were ranked by scoring each candidate against the other candidates. This produces only small quality gabs because the stronger models generate similar candidates. A reference-based ranking and scoring each candidate against the gold reference, could produce larger and cleaner gaps, resulting in stronger preference signals. The small gaps between the preference candidates are a contributing cause of the weak to null results of DPO and limit how strongly the results can be generalised beyond the ranking scheme that was used.

The experiments involved several methodological problems. They have been resolved but this still shapes the final results. The first was a failure of DPO to learn at all. The training loss reached a plateau because the preference optimisation was applied on top of an unmerged supervised adapter. So the trainable and the reference policies were not separated. By merging the supervised adapter into the base model before attaching a fresh adapter for DPO, this was resolved and this

was a precondition for every subsequent DPO result.

Another problem was the initial use of preference pairs generated by one model for all others. The first preference optimisation runs for Qwen and Ministral used candidates that had been generated by Tower. DPO requires the preferred and dispreferred candidates from each model's own output distribution, resulting in invalid runs in this case. The correction required regenerating the candidates separately for each model. However, even after this change, the DPO results did not improve. This case still shows how easily an invalid configuration can produce a misleading result.

A third problem involved an early version of the β-runs that did not reset between each other. So each β run built on the previous one rather on a clean supervised model. Therefore, the early β comparisons are invalide and re-run with a proper reset in-between them. A major problem was the collapse of the Ministral model. After an investigation and minor corrections of the BOS tokens and the chat-template, the issue could still not be resolved. For this reason, the collapse was accepted as a consistent negative result.

5.4 Future work

The limitations show several points that can be realistically studied in the future. The most direct idea is to replace the candidate ranking with reference-based ranking and scoring each candidate against the gold reference with COMET-DA. Through this change, larger quality gaps can be obtained which produce a stronger preference signal. A complementary step is to make the score-gap filter even tighter in order to get bigger preference contrasts.

Another direction addresses handling the terminology. Future work should investigate mechanisms that enforce rather than encourage the use of the injected terms. One option is constrained or lexically constrained decoding. This can guarantee that approved terms appear in the output. An additional option is to move from offline, pairwise DPO to an online reward-base policy-optimisation method with an explicit terminology reward. Yirmibeşoğlu Balal and Güngör (2026), for example, optimise a group-relative policy with a multi-objective reward that combines COMET-DA and BLEU with a task-specific reward. This policy optimisation can substantially improve the explicitly rewarded property without degrading overall quality. This applies to a setting where SFT alone was not able to handle it. So a promising idea for the present experiments would be the design of an analogous terminology-coverage reward that is computed against UMLS in the ranking step and optimising it directly with such a method.

Finally, the model-dependence of the results suggests future work on the interaction between the strength of the supervised baseline and the value of DPO. This

includes testing of additional models and varying the amount and the quality of the supervised data to clarify whether there is a common thread in which preference optimisation does help medical MT or whether the observed results generalise across model strengths.

6 Conclusion

This thesis investigated whether DPO improves the quality of medical terminology translation in English-to-French MT compared to SFT with the untuned zero-shot model as the baseline. To motivating problem is that fluent MT output can cause that terminological errors are overlooked and therefore carries real clinical risk. The objectives were to establish a untuned zero-shot baseline together with SFT and a terminology-enriched SFT variant for three multilingual LLMs, to develop a terminology-aware DPO method on top of those supervised systems, to compare the training paradigms systematically across the models and to complement the automatic evaluation with a qualitative error analysis. The overall approach was empirical and comparative. The same data and evaluation procedure were applied to each model so that any observed differences could be attributed to a specific training step that was changed.

The training and the validation data were parts of the EMEA v3 EN-FR parallel corpus. It was filtered using heuristic and semantic filtering and downsampled to a final set of 20.000 training examples, and 1.000 examples each for the test and validation set. Additionally, the TICO-19 benchmark was reserved entirely for the final testing stage. Medical terms were extracted form the English source using scispaCy and resolved to their French preferred terms through a UMLS lookup. Those terms were then injected as a glossary header into the existing translation instruction prompt to form a terminology-enriched training variant.

Three decoder-only models, namely TowerInstruct-7B, Ministral-8B-Instruct-2410 and Qwen3-8B, were fine-tuned with QLoRA. Each model was going through the same pipeline, consisting of a untuned zero-shot baseline, SFT, a terminology-enriched SFT variant and a terminology-aware DPO variant built on the supervised model. For DPO, the supervised adapter was first merged into the base model, a fresh low-rank adapter was attached and the preference pairs were constructed from eight hypotheses generated per source sentence. Finally, $\beta = 0.01$ was selected for all three models returning the best results. Evaluation used BLEU and chrF, as well as COMET-DA as the primary metric. The last step consisted of a dedicated terminological accuracy measurement, a paired bootstrap significance testing and a manual inspection of system outputs.

6.1 Key Findings

The central key finding of the experiments is that DPO did not provided a consistent improvement over SFT for medical terminology translation. On the test sets, the terminology-aware preference-optimised model differed from the terminology-enricher supervised model by only a few thousandths in COMET-DA for any model. Also the different choices of the β values had little to no effect on the training procedure. For the strongest model, Qwen, the supervised and preference-optimised terminology systems produced nearly identical outputs. This confirms that there is no big difference between them.

By contrast, SFT had a larger effect. In comparison to the untuned zero-shot baseline, it improved COMET-DA in domain for Tower ($0.748 \rightarrow 0.839$ on EMEA) and Qwen ($0.840 \rightarrow 0.869$). However, it scored below the baseline on out of domain data for all three models. The terminology-enrichment process did not consistently improve the quality. It lowered Tower's EMEA COMET-DA ($0.839 \rightarrow 0.509$) but improved it slightly on TICO-19 ($0.476 \rightarrow 0.524$). The change in the other systems even smaller. A consistent observation across all models is that the difference between the surface-level metrics and the semantic metric is large.

The terminological accuracy measurement showed that the recall of the injected glossary terms remained low with $\approx 22\%$ on the TICO-19 test set and $\approx 42\%$ on the EMEA test set. The numbers were nearly identical across the supervised and the preference-optimised systems.

6.2 Contributions

The work includes four contributions that are related to the research objectives. First, it establishes a untuned zeroShot baseline together with plain and terminology-enriched SFT systems for English-to-French medical translation across three multilingual models. Second, it develops a terminology-aware DPO pipeline, including the methodological steps required for stable preference optimisation on top of PEFT models. This happens by merging the supervised adapter before attaching a fresh one and generating preference candidates from each model's own outputs instead of reusing another model's outputs. Third, it provides a systematic comparison of SFT and DPO across models supported by automatic measure, a terminological-accuracy measure, a significance testing and it reports clear negative results for DPO under the studied configuration. Fourth, it contributes insight into terminology failures and documents the divergence between surface-level and semantic metrics. This part is directly relevant to how medical MT should be evaluated. Practically, the findings show that the choice of the base model and the quality of the terminology resource matter more for medical terminology translation

than the additional preference-optimisation stage.

This Master's thesis goal was to determine whether DPO improves medical terminology translation beyond SFT. Across three models and two test sets it found that it does not reach the assumptions under the studied configuration. SFT has a mixed effect that can fall below the untuned zero-shot baseline and terminology-enrichment does not improve consistently. Another important remark is that the surface-level metrics can move opposite to the translation quality in this setting and that the low terminology recall shows that the automatic metrics do not directly capture terminological correctness. The final conclusion of this work is that the starting model and the quality of the terminological resource are important, and that meaningful progress on terminological accuracy requires mechanisms that enfore instead of only encourage the use of correct terms.

Bibliography

Abien Fred Agarap (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M Guerreiro, and Andre Martins (2024). Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14503–14519, Miami, Florida, USA. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Barto Andrew and Sutton Richard S (2018). Reinforcement learning: an introduction.

George L Banay (1948). An introduction to medical terminology i. greek and latin derivations. *Bulletin of the Medical Library Association*, 36(1):1.

Dimitri Bertsekas (2019). *Reinforcement learning and optimal control*, volume 1. Athena Scientific.

Olivier Bodenreider (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl$_1$) : $D267 - -D270$.

L Bos and K Donnelly (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290.

Lynne Bowker and Gloria Corpas Pastor (2022). Translation technology. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Ralph Allan Bradley and Milton E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kevin B. Cohen (2022). Natural language processing for biomedical texts. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Louise Deléger, Tayeb Merabti, Thierry Lecrocq, Michel Joubert, Pierre Zweigenbaum, and Stéfan Darmoni (2010). A twofold strategy for translating a medical terminology into french. In *AMIA Annual Symposium Proceedings*, volume 2010, page 152.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondřej Dušek, Jan Hajic, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdenka Uresova, and Daniel Zeman (2014). Machine translation of medical texts in the khresmoi project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei (2024). Towards analyzing and understanding the limitations of dpo: A theoretical perspective.

Glenn Flores (2005). The impact of medical interpreter services on the quality of health care: a systematic review. *Medical care research and review*, 62(3):255–299.

April G Fritz (2000). *International classification of diseases for oncology: ICD-O*. World health organization.

Miguel Angel Rios Gaona, Raluca-Maria Chereji, Alina Secara, and Dragos Ciobanu (2023). Quality analysis of multilingual neural machine translation systems and reference test translations for the English-Romanian language pair in the medical domain. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 355–364, Tampere, Finland. European Association for Machine Translation.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Sepp Hochreiter and Jürgen Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). Lora: Low-rank adaptation of large language models.

Marcin Junczys-Dowmunt (2025). GEMBA v2: Ten judgments are better than one. In *Proceedings of the Tenth Conference on Machine Translation*, pages 926–933, Suzhou, China. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

Dain Kim, Jiwoo Lee, Jaehoon Yun, Yong Hoe Koo, Qingyu Chen, Hyunjae Kim, and Jaewoo Kang (2026). Benchmarking direct preference optimization for medical large vision–language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5052–5067, Rabat, Morocco. Association for Computational Linguistics.

Richard I. Kittredge (2022). Sublanguages and controlled languages. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Tom Kocmi and Christian Federmann (2023). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Ioannis Korkontzelos and Sophia Ananiadou (2022). Term extraction. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Solomon Kullback (1951). Kullback-leibler divergence. *Tech. Rep.*

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton (2015). Deep learning. *nature*, 521(7553):436–444.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Carolyn E Lipscomb (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi,

Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi (2026). Ministral 3.

Yixin Liu, Pengfei Liu, and Arman Cohan (2025). Understanding reference policies in direct preference optimization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8037–8052, Albuquerque, New Mexico. Association for Computational Linguistics.

Arle Richard Lommsel, Aljoscha Burchardt, and Hans Uszkoreit (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi (2022). Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2016–2025, New York, NY, USA. Association for Computing Machinery.

Mistral AI (2024). Ministral-8B-Instruct-2410. `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`. Hugging Face model repository. Accessed: 2026-06-27.

Palanichamy Naveen and Pavel Trojovský (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.

Roberto Navigli (2022). Ontologies. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes (2024). Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang (2020). What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team (2025). Qwen3-8B. `https://huggingface.co/Qwen/Qwen3-8B`. Hugging Face model repository. Accessed: 2026-06-27.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn (2023). Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins (2025). Tower+: Bridging generality and translation specialization in multilingual llms.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Miguel Rios (2025). Instruction-tuned large language models for machine translation in the medical domain. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 162–172, Geneva, Switzerland. European Association for Machine Translation.

Thomas Savage, Stephen P Ma, Abdessalem Boukil, Ekanath Rangan, Vishwesh Patel, Ivan Lopez, and Jonathan Chen (2025). Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: Comparative evaluation. *J Med Internet Res*, 27:e76048.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov (2017). Proximal policy optimization algorithms.

Paul Sebo and Sylvain de Lucia (2024). Performance of machine translators in translating french medical research abstracts to english: A comparative study of deepl, google translate, and cubbitt. *Plos one*, 19(2):e0297183.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang (2025). Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5).

Konstantinos Skianis, Yann Briand, and Florent Desgrippes (2020). Evaluation of machine translation methods applied to medical terminologies. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 59–69, Online. Association for Computational Linguistics.

Lucia Specia and Yorick Wilks (2022). Machine translation. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Haoxiang Sun, Ruize Gao, Pei Zhang, Baosong Yang, and Rui Wang (2025). Enhancing machine translation with self-supervised preference data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23916–23934, Vienna, Austria. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jörg Tiedemann (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Unbabel (2024). TowerInstruct-7B-v0.1. `https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1/tree/main`. Hugging Face model repository. Accessed: 2026-06-27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu (2025). Qwen3 technical report.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne (2024). Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu Balal and Tunga Güngör (2026). Diversity-aware literary machine translation with multi-reward policy optimization. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)*, pages 249–262, Linz, Austria. European Association for Machine Translation.

Kunwar Zaid, Amit Sangroya, and Jyotsana Khatri (2025). Multilingual clinical dialogue summarization and information extraction with qwen-1.5B LoRA. In *NLP-AI4Health*, pages 69–74, Mumbai, India. Association for Computational Linguistics.

Marco Zappatore and Gilda Ruggieri (2024). Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, 84:101582.

Acronyms

AI Artificial Intelligence. 6

BLEU BiLingual Evaluation Understudy. 2, 3, 5, 6, 10, 25, 26, 45, 48, 51–57, 60, 62, 65

CAT Computer-Assisted Translation. 27

chrF Character n-gram F-score. 2, 3, 5, 6, 26, 48, 52–56, 59, 60, 65

CNN Convolutional Neural Network. 7

COMET-DA Crosslingual Optimized Metric for Evaluation of Translation. 2, 3, 5, 6, 10, 26, 47, 48, 51–57, 59, 60, 62, 65, 66

CUI Concept Unique Identifier. 29

DL Deep Learning. 5, 6

DPO Direct Preference Optimization. 2–6, 10–13, 15–23, 26, 28, 32–35, 37–40, 44, 46–49, 52–62, 65–67

EMEA European Medicines Agency. 2–6, 9–11, 15, 21–23, 30, 35, 38–41, 43, 45, 49, 51, 53–57, 60, 61, 65, 66

ICD International Classification of Diseases. 1, 28, 29

KL Kullback-Leibler. 17, 19, 20, 22, 23, 32, 46

LLM Large Language Model. 1–3, 5–7, 9, 11–14, 18, 22, 23, 25, 26, 37, 43, 65

LoRA Low-Rank Adaptation. 2, 3, 7, 12, 13, 20–22, 31, 33, 35, 46

LSTM Long Short-Term Memory. 6

MBR Minimum Bayes Risk. 9, 33, 47, 48, 56

MeSH Medical Subject Headings. 1, 28, 29

ML Machine Learning. 5

MT Machine Translation. 1–3, 5, 6, 10, 11, 15, 16, 23–31, 37, 39, 40, 59, 60, 63, 65, 66

NLP Natural Language Processing. 10, 11, 13, 29, 33

NMT Neural Machine Translation. 1, 5, 23–25, 27

NN Neural Network. 5, 7

PEFT Parameter-Efficient Fine-Tuning. 2, 3, 5, 9, 11, 12, 14, 21, 22, 25, 33, 44, 46, 66

PPO Proximal Policy Optimization. 17, 20, 22, 32

QLoRA Quantized Low-Rank Adapter. 3, 5, 9, 34, 38, 44–46, 58, 65

ReLU Recitified Linear Units. 7

RL Reinforcement Learning. 14, 16–19, 22, 32

RLHF Reinforcement Learning from Human Feedback. 5, 15–20, 22, 32

RNN Recurrent Neural Network. 6, 7

SFT Supervised Fine-Tuning. 1–6, 9, 11–17, 19–23, 26–28, 31–35, 37–39, 44–46, 48, 49, 51–56, 58–62, 65–67

SNOMED Systematized Nomenclature of Medicine. 1, 28, 29

TICO-19 Translation Initiative for COVID-19. 2–6, 9, 11, 30, 38–40, 43, 49, 51, 53–56, 59–61, 65, 66

UMLS Unified Medical Language System. 1–3, 5, 9, 10, 27–30, 38, 42, 47, 49, 54, 57, 58, 61, 62, 65