



Exploring the Capabilities of ChatGPT in Machine Translation

Maria Eleftheria Dimoka

Department of Foreign Languages, Translation and Interpreting (DFLTI), Ionian University

Master's Thesis

Supervision Committee: Dr. Vilelmini Sosoni, Dr. Maria Tsigou, and Dr. Katia - Lida Kermanidis

July 2024

Table of Contents

Abstract.....	3
1. Introduction	4
2. ChatGPT	5
2.1 Origins of ChatGPT	7
2.2 The Evolution of Large Language Models and Generative Pre-trained Transformers	9
2.3 Overview and Operational Framework of ChatGPT	15
2.4 Strengths and Weaknesses.....	18
2.5 Applications.....	24
3. ChatGPT and Machine Translation.....	26
3.1 Related Work.....	29
3.1.1 Comparative Machine Translation Evaluation.....	29
3.1.2 Prompt Engineering for Machine Translation	31
3.1.3 Machine Translation Evaluation for High-resource and Low-resource Languages.....	35
3.2 The Case of the Greek Language	38
4. Research Methodology.....	40
4.1 Domains.....	42
4.2 Zero-shot Experimental Layout	43
4.3 Automatic Evaluation	44
4.4 Human Evaluation	46
4.4.1 Error Classification Task.....	49
4.4.2 DEJ-based Evaluation Task	51
5. Key Findings	52
5.1 Error Classification Analysis.....	53
5.2 Human and Automatic Evaluation Analysis.....	55
6. Pivot Prompting	59
7. Conclusion	61
References	65
Appendices	77
Appendix A: Zero-shot Experimental Layout.....	77
Appendix B: Error Classification Scores	80
Appendix C: Human Evaluation and Automatic Evaluation Scores.....	81
Appendix D: Pivot Prompting Evaluation Scores	81

Abstract

Amidst the growing interest in AI-driven language models and their extensive application in translation tasks, this study seeks to provide empirical insights into ChatGPT's efficacy as a machine translation tool specifically for the Greek language, a low-resource language. Furthermore, this research explores the potential optimization of the model's translation capabilities through the implementation of various prompting strategies. The evaluation framework employed in this study involves a multifaceted approach, combining human evaluation and automatic evaluation metrics across diverse text domains. The primary objectives of this study are to assess the machine translation capabilities of ChatGPT, identify optimal prompting strategies that enhance the quality of the model's translation outputs and evaluate its potential for professional utilization within the translation industry.

1. Introduction

In recent years, the rapid progression of Artificial Intelligence (AI) and Natural Language Processing (NLP) has ushered in the development of advanced Large Language Models (LLMs), capable of performing a wide array of sophisticated tasks. Among these advancements, the advent of ChatGPT in 2022 marked a significant milestone in the field of LLMs and generative AI, exemplifying the rapid evolution of chatbot technology. Despite the ongoing debates concerning its usability, ChatGPT quickly garnered tremendous interest from researchers and professionals alike, owing to its multifaceted capabilities. One particularly intriguing aspect of ChatGPT is its potential in multilingual applications and Machine Translation (MT), which sparked significant interest within the field of translation studies and translation industry.

The primary objective of this research is to explore ChatGPT's efficacy as a MT tool, with a specific focus on discovering whether the model is capable of delivering high-quality translations for the Greek language. Furthermore, we have experimented with optimization prompting strategies aiming to enhance the quality of machine-generated translations. Our research entailed a comprehensive evaluation of the model through both theoretical and practical lenses, employing both human evaluation and automatic evaluation metrics, allowing us to gain deeper insights into the model's translation capacity, common errors, and potential applications in the translation industry.

To establish a solid foundation for our research and form our expectations accordingly, we conducted an extensive analysis of ChatGPT's operational framework, strengths, limitations, and various applications. This analysis was supplemented by a thorough review of related studies on ChatGPT's MT capabilities, with particular attention to prompting techniques and challenges associated with low resource languages, which are pertinent to our investigation.

2. ChatGPT

The conversational artificial intelligence (AI) chatbot, ChatGPT¹, stands out as a notable exemplar of the advancements in the field of Natural Language Processing (NLP), and a significant milestone in the evolution of AI. Following its launch in November 2022 by OpenAI², it has accumulated a user base exceeding 180 million individuals, with its official website now recording an average of 1.6 billion monthly visits in 2024 (Duarte, 2024). The fast-growing AI platform continues to attract attention, and it has gained unprecedented popularity across diverse user demographics, ranging from individual consumers to large-scale enterprises. Notably, the scientific and academic communities have shown significant interest in the research surrounding the development and various applications of ChatGPT. Ray (2023) clarifies that by “March 2023 more than 3000 articles, reports, news have been published in various journals, conferences, newspapers, blogs and media reports,” as documented by Google Scholar (p. 122).

ChatGPT is a sophisticated and versatile AI chatbot that is trained to interpret instructions provided in prompts and generate contextually appropriate responses in a conversational manner, allowing it to assist in diverse text-based tasks and applications, and adapt to a variety of contexts and styles (Ray, 2023; Nazir & Wang, 2023). The term “chatbot” refers to an autonomous computer program designed to interact with users and simulate human conversation through a user-friendly interface by responding accordingly to text or voice-to-text input. ChatGPT 3.5 chatbot is equipped with a conversational and easily accessible interface tailored for internet users, and its open-access website enables extensive individual utilization, thereby offering the crowd the opportunity to engage with AI applications. With its already advanced technology and impressive performance, ChatGPT holds the

¹ ChatGPT. (n.d.). OpenAI. <https://openai.com/chatgpt/>

² OpenAI. (n.d.). <https://openai.com>

potential to redefine human-machine interactions and provide alternative solutions for facilitation across real-world scenarios.

ChatGPT is based on the Generative Pre-Trained Transformer (GPT) architecture for language model development, which is typically used for natural language processing tasks like text generation and language understanding. It is important to highlight the disparity between the architecture of GPT models, and Generative Adversarial Network (GAN) models which are primarily used for image generation (Ray, 2023). These differences extend beyond their respective applications and pertain to their structural and functional features. In the context of GPT architecture, ChatGPT 3.5 was developed as an advancement of its predecessors, earlier versions of GPT models developed by OpenAI, and serves as a direct progression of the InstructGPT language model, incorporating improvements in adaptability and functionality (Ouyang et al., 2022). Notable features that distinguish ChatGPT from initial GPT models include its dialogue-based format and its fine-tuning with the employment of both supervised learning and Reinforcement Learning from Human Feedback (RLHF) (Ray, 2023, p. 125).

Since its release, ChatGPT has emerged as an efficient tool with extraordinary performance in task resolution, successfully assisting in a variety of applications including data analysis, content generation, machine translation, education and healthcare assistance. Despite the significant challenges and ethical concerns that prevail in preliminary evaluation studies, the renowned chatbot exhibits the potential to surpass them. Following the employment of GPT-4³ architecture in the platform in March 2023, ChatGPT's overall performance was significantly boosted, allowing it to unveil heightened capabilities and undertake more sophisticated tasks (Jiao et al., 2023). Even though the software upgrade would provide our research with deeper insights into the chatbot's outputs and capabilities, we considered evaluating the earlier version of ChatGPT due to its accessibility to professionals, thereby

³ GPT-4. (n.d.). OpenAI. <https://openai.com/index/gpt-4/>

facilitating extensive assessment of a widely available tool. This section aims to provide an in-depth overview of ChatGPT, delving into its background, operational framework, strengths, limitations, and applications.

2.1 Origins of ChatGPT

As previously noted, ChatGPT stands as one of the latest groundbreaking creations of OpenAI, representing a significant milestone in the organization's line of work. OpenAI is an AI research organization that was founded in 2015 by Elon Musk, Sam Altman, and other investors and is now headquartered in San Francisco. Within the broader framework of developing Artificial General Intelligence (AGI)⁴ for the benefit of humanity, OpenAI's primary objective is to facilitate the progression of AI by empowering researchers and developers with tools and resources to effectively innovate in AI technologies while fostering a safe and beneficial environment for AI utilization (Alawida et al., 2023; Ray, 2023). Since its inception in the field of AI, OpenAI has garnered significant support and investment, facilitating the production of innovative applications that extend beyond NLP. At the forefront of its development initiatives is the creation of the novel series of GPT language models, which sequentially encompass GPT-1, GPT-2, GPT-3, GPT-3.5, and GPT-4. Most of these versions and similar affiliated models are accessible through the OpenAI Application Programming Interface (API)⁵, which enables developers to integrate them into their applications and services. In this way, users can easily apply the AI technology provided by OpenAI for their tasks on "a general-purpose text-in and text-out cloud platform" (Lisowski, 2024). In this context, the OpenAI API gives access to a range of pre-trained AI models, each

⁴ Artificial General Intelligence (AGI) is the science-fiction version of artificial intelligence (AI), where artificial machine intelligence achieves human-level learning, perception, and cognitive flexibility (Mucci & Stryker, 2024).

⁵ The OpenAI Application Programming Interface (API) serves as a gateway for developers to access OpenAI's pre-trained AI models, including versions of GPT and others, via a cloud-based service hosted on Microsoft Azure. It allows easy integration of OpenAI's AI capabilities into various applications and services (Lisowski, 2024).

with distinct capabilities, which extend beyond the different versions of GPT. Presented below is a brief description of several noteworthy and innovative models developed by OpenAI in recent years.

- (1) *DALL·E*: In January 2021, OpenAI released DALL·E⁶, a transformer language model that utilizes the capabilities of GPT-3 in a multimodal setting as it is trained to generate and modify images based on natural language input prompts (Roumeliotis & Tselikas, 2023). A year later, in April 2022, OpenAI launched its successor DALL·E 2⁷, and in 2023 its latest and more advanced version DALL·E 3⁸ which was integrated with Chat-GPT. Through these iterative upgrades, the model is now capable of demonstrating incredible results in image generation, visual quality, and adherence to prompt semantics, effectively overcoming the substantial errors of its predecessors (O'Connor, 2023). Notably, these advancements also include the incorporation of ethical guidelines designed to limit the generation of violent and adult content and the production of images that resemble the style of living artists.
- (2) *Whisper*: Whisper⁹ is an advanced Automatic Speech Recognition (ASR) model designed for the multilingual transcription and translation of spoken language. Since its release in 2022, the model has been trained on 680,000 hours of diverse speech datasets, enabling it to achieve human-like accuracy in speech recognition. Whisper's exceptional performance includes its ability to recognize various languages and accents, distinguish audio amidst background noise, and decipher fast speech. Nevertheless, the model is primarily English-centric, and its performance remains suboptimal in speech recognition and translation for low-resource languages (Radford et al., 2022).

⁶ DALL·E: Creating images from text. (2021, January 5). OpenAI. <https://openai.com/index/dall-e/>

⁷ DALL·E 2. (n.d.). OpenAI. <https://openai.com/index/dall-e-2/>

⁸ DALL·E 3. (n.d.). OpenAI. <https://openai.com/index/dall-e-3/>

⁹ Introducing Whisper. (2022, September 21). OpenAI. <https://openai.com/index/whisper/>

(3) *GPT-4o*: GPT-4o¹⁰ (omni) is the latest addition to OpenAI's AI language models, released in May 2024, and represents a significant advancement in AI technology. The model is introduced as an enhanced version of ChatGPT, branded as an all-in-one multimodal model capable of receiving text, audio, and image inputs and responding with a combination of text, images, and audio, thereby creating an immersive environment for natural human-computer interaction. The significant upgrades of GPT-4o over its previous versions mainly include improved processing and response speed, advanced translation capabilities, emotional intelligence, and an enhanced understanding of text, vision, and audio (Guinness, 2024).

2.2 The Evolution of Large Language Models and Generative Pre-trained Transformers

To evaluate the role of ChatGPT, organize our research, and form our expectations accordingly, we consider it essential to examine its development process and operational architecture. ChatGPT and its predecessors fall within the broad field of Generative AI, defined as “a class of artificial intelligence models that can create new data based on patterns and structures learned from existing data” (Ray, 2023, p. 121). Generative AI models, which have now become mainstream, are engineered to employ deep learning methods and neural network architecture to process and produce original content that closely resembles human-like outputs.

A specific subset within Generative AI models comprises Large Language Models (LLMs) which are primarily focused on Natural Language Processing (NLP) tasks. LLMs undergo training on immense amounts of text data to understand and generate natural language based on their training, revolutionizing the field of NLP by reaching unprecedented levels of quality in human language

¹⁰ Hello GPT-4o. (2024, March 13). OpenAI. <https://openai.com/index/hello-gpt-4o/>

processing. These models are designed to handle language in a similar way humans do and are capable of performing countless language-related tasks, such as translation, summarization, and question answering, without task-specific training (Ray, 2023, p. 134). Their dominance in their field is attributed to their large scale, in terms of the corpus size and the number of parameters (model size) they are trained on. LLM parameters are the processing guidelines that define how the model converts input data into output data, thereby influencing the manner in which input features affect the final output. The increasing number of parameters in an LLM broadens its spectrum of patterns, thereby advancing its complexity, diversity, and ability to predict and learn more comprehensively (Broadhead, 2023). However, the proficiency of LLMs is not entirely contingent on high parameter count but also depends on the quality and diversity of parameters, giving the model the ability to better navigate a wide range of sophisticated contexts and inputs. Notwithstanding specific ethical and interpretative challenges, LLMs are constantly evolving, demonstrating the potential to overcome past limitations and achieve new milestones.

Generative Pre-trained Transformers (GPT), a prominent subcategory of LLMs, have dominated the landscape since they were introduced by OpenAI in 2018. Before delving into GPT models' workflow, it is crucial to understand the origins of their architecture and their evolutionary progress over the years. The introduction of transformer architecture by Vaswani et al. in 2017 marked the starting point of the models' development and addressed the inherent limitations of conventional language models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). The transformer architecture is esteemed for its aptitude "to maintain contextual understanding and capture extended-range dependencies in text data, which is indispensable for tasks such as language modeling and text generation" (Roumeliotis & Tselikas, 2023, p. 3). The architecture is structured based on the employment of a series of transformer blocks (encoder and decoder blocks), each comprising a self-attention

mechanism and a feedforward neural network. The incorporation of self-attention mechanisms allows the models to evaluate the significance of different words (tokens) in a sentence according to their contextual relevance. On the other hand, the feedforward network grants the models the ability to comprehend non-linear correlations between input and output (Roumeliotis & Tselikas, 2023, p. 3).

ChatGPT's foundations lie in the GPT architecture, which is built upon the aforementioned transformer technology. The series of GPT models are designed to generate natural language output closely similar to human language, a feat facilitated by their innovative training approach that is based on unsupervised pre-training followed by supervised fine-tuning.

- *Pre-training*: One of the fundamental attributes of GPT models is their ability to undergo large-scale pre-training, enabling them to be exposed to vast datasets consisting of billions of tokens, utilizing advanced computational resources. During the pre-training phase, GPT models are trained on extensive corpora of text data (e.g., webpages, books) in an unsupervised mode, meaning that they do not require explicit labels or annotations for the training data (Ray, 2023, p. 123). Through this process, the models learn to predict the next word in a sequence based on a context window of preceding words, a task referred to as language modeling. Consequently, the models develop the capability to identify and generalize linguistic patterns and structures, having been exposed to a diverse and rich array of language representations (Nazir & Wang, 2023, p. 2).
- *Fine-tuning*: Following the initial phase of unsupervised training, fine-tuning is the technique employed to improve the performance of GPT models on specific downstream tasks. This process involves the adjustment of the models using a smaller, labeled dataset to improve their specialization and efficacy in particular tasks, such as text summarization or question answering. The fine-tuning dataset modifies the models' weights and biases, aligning them more closely

with the task requirements, in an attempt to reduce the discrepancy between the models' predictions and actual targets for the task (Ray, 2023, p. 123). This task-specific training of GPT models enhances their understanding of linguistic nuances relevant to the target task, allowing them to adapt while retaining the foundational knowledge gained during their pre-training on large and generic datasets.

The progressive evolution of GPT architecture has led to the creation of a powerful series of language models by OpenAI. With each version, gradual improvements and modifications have resulted in increasingly enhanced and contextually aware models, ultimately leading to the development of specialized variants like ChatGPT. Shortly described below is a brief overview of GPT iterations over the years.

- (1) *GPT-1*: Introduced by OpenAI in 2018, GPT-1 laid the groundwork for the development of its successors by demonstrating notable performance despite its early technological framework. GPT-1 was based on a transformer architecture and proved that large-scale unsupervised pre-training can produce language models competent in handling a variety of NLP tasks with minimal fine-tuning. The model, comprising 117 million parameters – a relatively small number compared to subsequent GPT versions – was pre-trained on an extensive corpus of text data that included books, webpages, and articles (Ray, 2023, p. 123). Despite its limited size, GPT-1 exhibited remarkable capabilities in a wide range of NLP tasks and text generation, thereby establishing the foundation for the development of subsequent GPT models.
- (2) *GPT-2*: Building on the foundation established by GPT-1, OpenAI released GPT-2 in 2019, significantly increasing the model's size and enhancing its performance. GPT-2 featured 1.5 billion parameters and was trained on a more extensive and diverse dataset, enabling it to capture a broader spectrum of linguistic knowledge. The model outperformed its predecessor by

generating more realistic and contextually relevant text that closely resembled human-produced content (Roumeliotis & Tselikas, 2023, p. 4). GPT-2's proficiency in understanding and generating high-quality natural language showcased the versatility of LLMs and raised awareness about the ethical concerns related to the potential benefits and risks associated with AI language models. Specifically, the model's capabilities sparked discussions about potential misuse, such as the production of misleading and harmful content, and underscored the necessity for research aimed at minimizing such risks, thereby endorsing responsible AI development (Ray, 2023, p. 123).

- (3) *GPT-3*: The third version in the GPT series, GPT-3, was launched in 2020 and marked a significant improvement over GPT-2 in terms of both scale and performance. Containing 175 billion parameters, GPT-3 was pre-trained on a massive dataset called WebText2, which included a diverse array of sources such as web pages, articles, and books (Ray, 2023, p. 127). It was the first model to be made accessible through the OpenAI API and due to its considerable size and complexity, GPT-3 exhibited impressive flexibility and adaptability across a wide range of NLP tasks, including language translation and code generation, with minimal fine-tuning required (Nazir & Wang, 2023). The model significantly excelled in generating high-quality natural language text that was more coherent, contextually aware, and realistic, even demonstrating rudimentary reasoning skills (Ray, 2023, p. 127).

GPT-3 introduced several innovative features, including multitask learning, which enabled the model to simultaneously perform multiple tasks. Moreover, GPT-3 was designed to perform tasks with minimal task-specific data by effectively leveraging few-shot learning techniques. In other words, the model could adapt to new tasks with just a few examples (few-shot), a single example (one-shot), or even no examples (zero-shot), relying solely on its extensive pre-training dataset (Yang et al., 2022).

Building upon the technological foundation and success of GPT-3, OpenAI released InstructGPT in 2022, a conversational AI language model that demonstrated improved performance in multiple aspects. This model was engineered to understand and generate user instructions, aiming to provide precise responses without the need for carefully crafted prompts as required by earlier GPT models. InstructGPT was specifically developed to address ethical concerns associated with GPT-3, striving to align language models more closely with user intent and produce content of heightened accuracy and reduced harmful content (Chan, 2022, p. 53). The model was built upon a fully trained GPT-3 model and underwent an additional training phase known as Reinforcement Learning from Human Feedback (RLHF). InstructGPT paved the way for the development of ChatGPT and is considered a significant milestone in the evolution of GPT models.

- (4) *GPT-4*: In 2023, OpenAI published the latest variant of its GPT language models, GPT-4, which was integrated into the ChatGPT platform (OpenAI et al., 2023). The new model is significantly more efficient than GPT-3, displaying human-like performance in several complex tasks within professional and academic domains (Ray, 2023). GPT-4 was branded as a multimodal model, capable of receiving both text and image inputs. Since its release, it has proven to be a versatile and competent tool, demonstrating advanced accuracy and creativity in real-world scenarios and complex questions, surpassing prior GPT variants. Among GPT-4's most impressive features are its enhanced capabilities for sentimental analysis, multitasking, and handling of dialects (Alawida, 2023, p. 9). The training and fine-tuning techniques applied to GPT-4 are similar to those used for GPT-3, with the primary difference being the parameter count; GPT-4 contains 1.76 trillion parameters. OpenAI has made noteworthy progress in scaling up deep learning, having developed the largest language model to date. Additionally, GPT-4 has been carefully

designed with a stronger emphasis on ethical considerations, incorporating mechanisms to promote responsible AI usage and limit biases.

Model	Year	Parameters	Training Data	RLHF	Features
GPT-1	2018	117 million	WebText (4.5 GB)	No	NLP tasks, text comprehension and generation
GPT-2	2019	1.5 billion	WebText (40 GB)	No	Advanced contextual understanding, realistic text generation quality (human-like)
GPT-3	2020	175 billion	WebText2, Wikipedia, Common Crawl (570 GB)	No	Complex NLP tasks, more advanced contextual understanding, enhanced text generation quality, reasoning capabilities, multitask learning, few-shot and zero-shot learning
InstructGPT	2022	1.3 billion (size can vary)	Variant of GPT-3 (570 GB)	Yes	Superior at following and generating instructional text, alignment with human intent, advanced contextual understanding and appropriateness of responses due to RLHF
GPT-3.5	2022	6.7 billion	Variant of GPT-3 (570 GB)	Yes	Improved alignment with human intent, suitable for interactive applications, enhanced text generation quality with fewer parameters, focus on safety and human values
GPT-4	2023	1.76 trillion	Published and Licensed data (1 TB)	Yes	More complex NLP tasks, advanced reasoning capabilities, improved coherence and contextual relevance, sentiment analysis, expanded multilingual understanding, specialization across diverse domains, emphasis on ethical concerns

Table 1: GPT models

2.3 Overview and Operational Framework of ChatGPT

ChatGPT is a form of conversational AI, a rapidly evolving technological field that combines machine learning and Natural Language Processing (NLP) to imitate human interactions. The development of this chatbot was inspired by InstructGPT, which is a structurally similar model with essential functional differences. Both are conversational AI models and variants of the GPT-3 model, but they differ in their applications, training data, and interaction style. InstructGPT is designed to generate instructional text and provide users with stepwise guidance for executing various tasks, while ChatGPT is a conversation-oriented and flexible agent, able to facilitate more natural and flowing interactions and execute a wider variety of text-based tasks (Alawida et al., 2023).

OpenAI released ChatGPT in 2022 and branded the chatbot as a language model that can interact conversationally (Introducing ChatGPT, 2022). However, it is crucial to distinguish the language model from the ChatGPT website and its additional features¹¹. ChatGPT is not a Large Language Model (LLM) itself, but an application powered by the GPT model behind it and displayed in the form of a chatbot website that is provided with extra components (e.g., chatbot interface, chat history) in order to be marketable and user-friendly.

ChatGPT is an AI conversational application that generates human-like responses to user inputs and proficiently executes NLP tasks in an interlocutory setting. It is based on the GPT-3.5 architecture, a smaller, adapted version of the GPT-3 language model, modified to align with human intent. Despite having 6.7 billion parameters, significantly fewer than GPT-3 (175 billion), GPT-3.5 excels at text generation, classification, and language modeling. The employment of GPT-3.5 architecture for ChatGPT aims to ensure a safer and more natural application of AI technology, by conforming to policies that prioritize human values in order to overcome past ethical concerns.

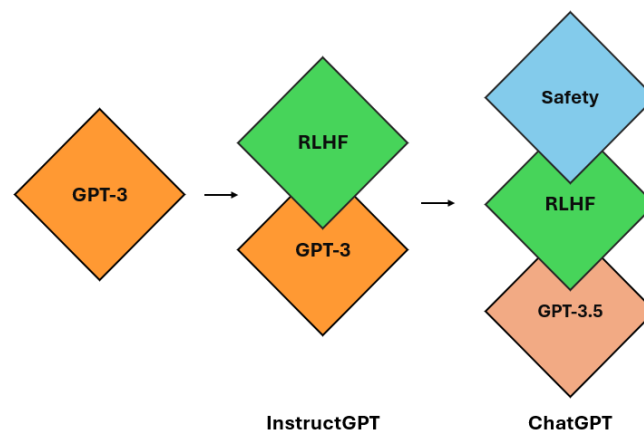


Figure 1: Stages of evolution

¹¹ In this paper, ChatGPT will be referred to as “model” for practical reasons, often indicating the chatbot’s underlying GPT architecture.

ChatGPT 3.5 acquires the characteristics of GPT architecture and “is developed through a two-phase process of supervised pre-training followed by supervised fine-tuning” (Roumeliotis & Tselikas, 2023, p. 1). During the pre-training phase, the model is trained on a massive corpus of raw, unlabeled text data in an unsupervised mode, utilizing innovative training algorithms. ChatGPT is strategically trained on selected corpora which entail conversations, reviews, social media texts, articles and more, as a way of optimizing the model for dialogue and interactive utilization. In this way, the model learns to predict the next word in a sentence based on a context window of preceding words, a task referred to as language modeling. This process enables ChatGPT to obtain a deep understanding of the structure of natural language, capturing its patterns, nuances, and complexities (Roumeliotis & Tselikas, 2023, p. 2). Thus, the unique training method of ChatGPT equips the model with the ability to process intricate and ambiguous contexts, generate accurate responses to diverse topics and maintain context in conversational settings.

ChatGPT, like InstructGPT, excels in interactive settings due to its unique methodology of supervised fine-tuning that incorporates human guidance into its learning process. The pre-trained GPT-3.5 model is fine-tuned using extensive Reinforcement Learning from Human Feedback (RLHF), a multiphasic method designed to optimize the model’s machine learning algorithms (Zhou et al., 2023, p. 4). This process begins with supervised fine-tuning on a curated dataset, where AI trainers provide conversational prompts and assist the model in composing responses based on model-generated suggestions (Nazir & Wang, 2023, p. 3). The second phase of RLHF involves collecting direct human feedback from AI trainers, such as quality rankings of outputs, in order to train a rewards model. This rewards model, a necessary component for the process, is heavily dependent on its human-provided training data to emulate a human evaluator. Its primary objective is to convert qualitative human evaluations into quantitative scores, enabling the model to learn to assess new outputs and generate

more desirable, human-favored responses (Bergmann, 2023). Finally, the rewards model is utilized alongside a reinforcement learning¹² algorithm, Proximal Policy Optimization¹³ (PPO), to further fine-tune ChatGPT. This process enhances the coherence, relevance, and overall quality of the outputs to align with human preferences (Alawida et al., 2023, p. 4).

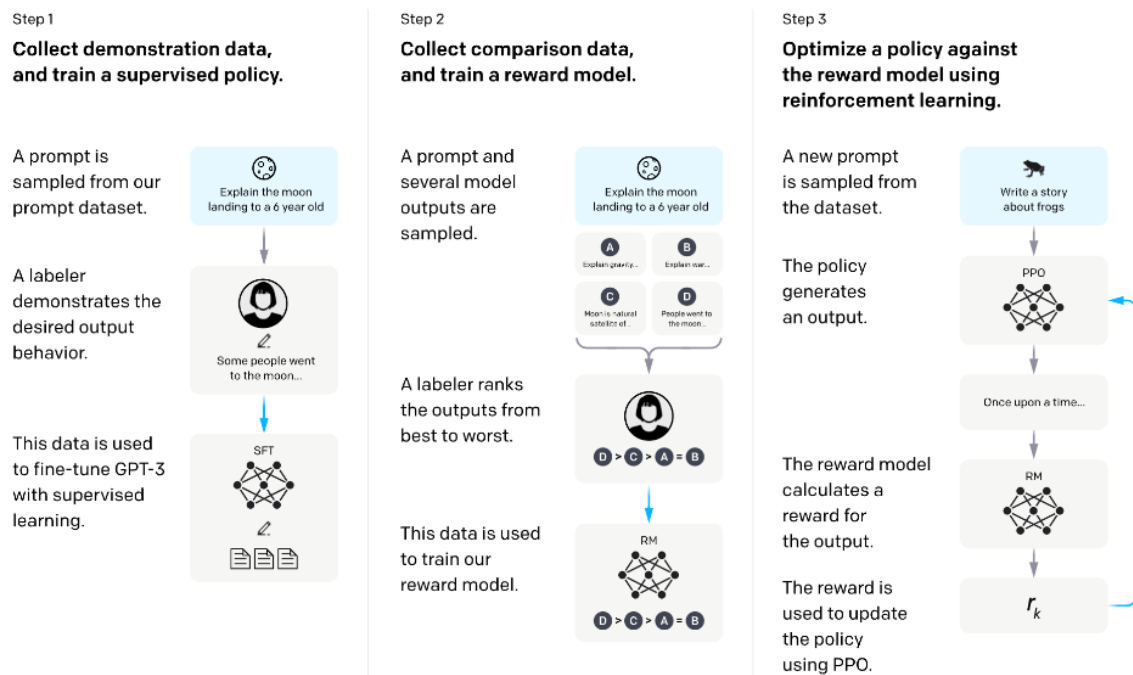


Figure 2: GPT-3.5 workflow (Introducing ChatGPT, 2022)

2.4 Strengths and Weaknesses

Understanding the architectural basis of ChatGPT provides deeper insights into the computational mechanisms that facilitate the model's extraordinary performance. Within the rapidly evolving AI landscape, ChatGPT remains among the leading conversational AI models of recent years, next to

¹² "Reinforcement learning is a machine learning training method based on rewarding desired behaviors and punishing undesired ones. In general, a reinforcement learning agent – the entity being trained – is able to perceive and interpret its environment, take actions and learn through trial and error" (Hashemi-Pour & Carew, 2023).

¹³ Proximal Policy Optimization (PPO) is a reinforcement learning algorithm, introduced by OpenAI in 2017, that strikes a balance between ease of implementation, sample complexity, and ease of tuning. PPO has become the default reinforcement learning algorithm at OpenAI because of its ease of use and good performance (Proximal Policy Optimization, 2017).

Gemini¹⁴ and Microsoft Copilot¹⁵. Its undeniable success significantly relies on its operational attainments, which make way for increasingly sophisticated human-machine interactions and drive revolutionary changes in businesses, academia, and everyday life. Listed below are some key functional features that support the diverse applications of ChatGPT:

- (1) *Natural Language Understanding (NLU)*¹⁶ and *Natural Language Generation (NLG)*: The advanced neural architecture of ChatGPT allows it to comprehend the meaning of natural language inputs, generate coherent and contextually appropriate responses and handle complex linguistic phenomena. This feature facilitates intuitive communication and highly realistic conversations with users, enabling the chatbot to maintain contextual information over prolonged interactions and effectively handle multiple inputs. Additionally, ChatGPT can decipher user intent and adapt to user preferences and conversational styles, generating increasingly fluent and personalized outputs (Roumeliotis & Tselikas, 2023, p. 16).
- (2) *Accessibility and Availability*: ChatGPT is an AI chatbot designed with a user-friendly interface and innovative features, such as chat history, to enhance the user experience. ChatGPT 3.5 is widely accessible through its web interface, mobile application, and OpenAI API, providing consistent assistance and reliable performance. While the latest version, GPT-4o, offers limited access for non-subscribers, ChatGPT 3.5 remains open-access despite occasional server overload challenges and lower performance (Rane et al., 2024, p. 72). Users can fully explore the chatbot's capabilities and communicate in natural language without requiring technical expertise or carefully crafted prompts. ChatGPT is designed to ensure availability

¹⁴ Gemini. (2024, May 14). Google DeepMind. <https://deepmind.google/technologies/gemini/>

¹⁵ Microsoft Copilot: Your everyday AI companion. (n.d.). Microsoft. <https://copilot.microsoft.com/>

¹⁶ Natural Language Understanding (NLU) is a branch of AI that uses computer software to understand input in the form of sentences using text or speech (Gillis, 2023).

across a diverse user base and allows seamless integration into various applications. It is a valuable tool for developers, professionals, educators, and individual users alike, promoting widespread adoption in both professional and everyday settings.

- (3) *Versatility and Scalability*: ChatGPT's extensive pre-training on sizeable datasets and its fine-tuning capabilities enable developers to adapt the model for specific tasks and domains, so that it generates tailored and relevant responses across diverse industries (Giray et al., 2024, p. 44). As a highly versatile tool, ChatGPT can be customized for a wide range of applications, including creative writing, translation, code generation, customer support, healthcare, education, and legal services. Its adaptability ensures it meets the specific needs and standards of various sectors, making it an invaluable assistant for both individuals and professionals. The chatbot's scalability enables it to handle extensive volumes of queries simultaneously, catering to applications with varying requirements, from routine tasks to large-scale projects (Ray, 2023, p. 126). Hence, the versatile architecture of ChatGPT ensures it remains a relevant and essential tool for meeting the evolving demands of modern industries, facilitating increased productivity and time efficiency.
- (4) *Multilingual Capabilities*: ChatGPT is trained to deliver coherent and contextually relevant responses in multiple languages, supporting a wide range of linguistic contexts. Its multilingual capabilities make ChatGPT an effective tool for cross-lingual applications, bridging language and cultural barriers, and promoting cross-cultural communication (Roumeliotis & Tselikas, 2023). The chatbot excels in language translation, multilingual content generation and customer service queries, ensuring that users can interact naturally in their native languages. Additionally, ChatGPT can generate culturally appropriate content based on suitable prompts, facilitating global accessibility and effortless interactions in

diverse settings. However, ChatGPT's performance can vary across different languages with its strongest capabilities centered around English (Robinson et al., 2023).

Since its release in 2022, ChatGPT 3.5 has evolved to be an indispensable conversational agent utilized across multiple and diverse applications worldwide, consistently attracting new users and acquiring new skills. However, it is essential to acknowledge that the chatbot and its underlying architecture exhibit several functional deficiencies. The growing interest in the platform has spurred extensive research, evaluating its performance on NLP tasks and highlighting various weaknesses. According to Ray (2023), ChatGPT holds the potential to address these challenges through ongoing research, development and optimization.

In fact, preliminary reviews of the latest version, GPT-4, launched in 2023, indicate remarkable advancements in addressing many of the limitations of its predecessor, such as improved handling of ambiguity and enhanced factual accuracy. These improvements are largely attributed to its more sophisticated architecture, expanded training data and increased model size. The conversational AI model is under continuous innovation, aiming to achieve new milestones and overcome its existing limitations. To effectively structure our research and set realistic expectations, it is crucial to examine related work concerning these limitations of ChatGPT 3.5. Some of the more frequently highlighted weaknesses that were pointed out in academic studies include:

- (1) *Contextual Understanding*: While ChatGPT excels among its competitors and predecessors in generating context-aware, realistic and fluent responses, it struggles in maintaining coherence across extended interactions and comprehending long-term, intricate contexts. These limitations can lead to occasional irrelevant or incomplete outputs, undermining the overall user experience. Additionally, ChatGPT 3.5 may provide unsatisfactory responses to ambiguous inputs, necessitating more specific prompts and potentially leading to time-

consuming interactions. The chatbot's incompetency in deciphering ambiguity and interpreting context is often attributed to its reliance on training data and user input, both of which can be insufficient. Thus, ChatGPT is prone to missing genuine comprehension and generating inconsistent or unrelated outputs, when prompts contain data outside its training scope (Giray et al., 2024, p. 45).

- (2) *Biases and Overreliance on Training Data:* ChatGPT is trained on vast amounts of curated textual data, often reflecting the biases and inaccuracies incorporated in its training sources. Due to the model's overreliance on the quality and diversity of its training data, ChatGPT generates undesirable responses significantly influenced by the flaws of its initial training dataset (Nazir & Wang, 2023). These biases propagated by the AI model can be factual, social, linguistic, or cultural, promoting inappropriate content and reinforcing existing stereotypes. This undermines the credibility of ChatGPT in domains where objective and unbiased outputs are essential, such as healthcare, scientific research, and legal contexts. Additionally, ChatGPT lacks true critical thinking and reasoning capabilities, limiting its ability to discern between reliable and inaccurate information (Bang et al., 2023). This highlights the general inability of AI to independently detect misinformation beyond its training data (Dale, 2020). The extensive use and popularity of ChatGPT further risk perpetuating and entrenching these biases, emphasizing the necessity for ongoing innovation and refinement of the model.
- (3) *Factuality and Hallucination:* A persistent challenge with ChatGPT, as documented in recent studies, is its susceptibility to generating nonfactual and untruthful information, a phenomenon referred to as hallucination (Ji et al., 2023). These hallucinations can manifest as plausible and realistic sensory experiences, including seemingly accurate and fluent outputs, yet they are factually fallacious and do not correspond to real-world scenarios

(Alkaissi & McFarlane, 2023). Bang et al. (2023), evaluated ChatGPT through tasks such as Machine Translation (MT) and Question Answering (QA) and identified considerable instances of extrinsic hallucinations. The model's tendency towards inaccuracy undermines its reliability, misleading users and unintentionally disseminating misinformation while mimicking cognizance. ChatGPT also lacks in-depth domain-specific knowledge and frequently generates outdated data, as it relies exclusively on its training data and user inputs, highlighting its unreliability in applications where accuracy is essential (Lecler et al., 2023, p. 272). Furthermore, ChatGPT's inability to authenticate the information it generates is another significant limitation, as it refuses to provide verification and references for its responses.

- (4) *Ethical Concerns*: The widespread utilization and advancement of ChatGPT 3.5 have raised ethical concerns, driving interest in addressing these issues through the implementation of policies and regulations in subsequent versions of the model. These ethical considerations include the employment of ChatGPT for malicious purposes, resulting in offensive, deceitful, inappropriate, or unauthentic responses (Chowdhury & Rahman, 2023). Additionally, the model is susceptible to privacy breaches, as it can access and record sensitive user data, such as medical records or financial information, prompting concerns about data protection and cybersecurity (Khan, 2023). Another critical issue involves the risks of plagiarism and academic fraud, as well as concerns about authorship attribution and intellectual property rights in instances where the AI model has contributed to the writing of a published document (Sallam, 2023; Guleria et al., 2023). Moreover, the extensive use of the chatbot has sparked concerns about overreliance on AI for routine, academic or professional tasks, potentially leading to a declining need for human expertise and a reduction in independent critical thinking skills.

2.5 Applications

As ChatGPT garners millions of users daily, its applications have expanded beyond individual use to encompass a broad range of professional settings. This transition underlines the model's versatility, enabling it to meet the complex demands of various industries and be tailored for specific professional applications. However, the extensive implementation of ChatGPT has raised concerns regarding the potential displacement of human expertise and workforce by AI technology, sparking significant debates about the chatbot's reliability and the extent of its exploitation for professional purposes.

Researchers have deemed the model suitable for assistant roles and limited applications to aid professionals in time-consuming tasks, highlighting its early stage of development and its unreliability in domains that require high accuracy and specialized knowledge (Sallam, 2023, p. 2). Another significant concern is the model's inability to be accountable for professional decisions, take initiative in critical situations, and maintain transparency in professional settings. Nevertheless, continuous innovations of the chatbot and the launch of its subsequent versions, GPT-4 and 4o, are consistently enhancing ChatGPT's role in the industry by gradually overcoming significant challenges and paving the way for potential revolutionary reforms in human-machine collaboration.

The development of ChatGPT 3.5 represents a milestone, signifying the incorporation of AI into everyday life and significantly reshaping our lifestyle. Extensive research has been centered around optimizing the AI chatbot within professional frameworks, suggesting guidelines for responsible utilization, credibility, and transparency. Many optimal applications of ChatGPT 3.5 aim to balance technological advancements with the maintenance of human expertise and supervision. Listed below are examples of the application of ChatGPT in several domains:

- a) *Language Translation*: ChatGPT aids in translating diverse content across its supported languages, producing translations with decent accuracy (Jiao et al., 2023). It also facilitates real-

time translation services and helps bridge language gaps. The chatbot performs exceptionally well in informal language translation tasks, though it exhibits limitations when applied in professional contexts (Lecler et al., 2023).

- b) *Content Generation*: ChatGPT can serve as a co-writer and creative agent successfully assisting in content creation tasks such as managing social media content and drafting emails and reports (Naumova, 2023). The chatbot has emerged as a valuable tool for professionals optimizing their workflows, saving time and effort, improving their productivity (Nazir & Wang, 2023, p. 5). However, research highlights the model's risks of biases and hallucinations suggesting its careful and supplemental role (Sallam, 2023).
- c) *Customer Service*: ChatGPT is widely utilized in customer support as a representative of businesses, facilitating automated conversations with customers. It effectively responds to frequently asked questions, provides information about services or products, and directs customers through troubleshooting guides. The model is capable of handling a large number of queries and its versatility allows it to adapt to various interactions in a professional manner, delivering tailored responses based on individual preferences and chat history (Rane, 2024, p. 81). In this way, the need for human intervention is minimized, thereby reducing the workload of professionals and enabling them to concentrate on more complex services.
- d) *Education*: ChatGPT serves as a proficient virtual tutor and study assistant across various educational settings. The model provides personalized tutoring and feedback tailored to students' knowledge levels and educational needs, creating customized lessons and material in an interactive environment (Baidoo-Anu & Owusu Ansah, 2023, pp. 55-56). The chatbot assists students in drafting essays, offering inspirational content and providing proofreading assistance. However, the utilization of ChatGPT should be limited to an assistant role, as the model may not always provide reliable information and poses risks of unintentional plagiarism and academic

fraud (Cotton et al., 2023, p. 230). Educators also benefit from the chatbot's ability to generate creative teaching material and automate grading, enhancing their productivity and efficiency.

- e) *Healthcare Assistance*: ChatGPT can function as a virtual medical assistant, aiding healthcare professionals, medical students, and patients. For healthcare professionals, the model facilitates the retrieval of medical information, aids in clinical decision-making, and assists in the summarization of clinical reports and patient history (Dave et al., 2023, p. 02). Additionally, ChatGPT serves as an interactive and multilingual learning tool for medical students, enhancing their comprehension of medical concepts (Sallam, 2023). The chatbot can also offer mental support to patients and guide them through their symptoms and treatment pathways.
- f) *Programming*: ChatGPT can be a valuable assistant to software developers by generating and optimizing code (Kasneci et al., 2023, p. 6). The chatbot is capable of automatizing repetitive coding tasks, identifying potential errors, and fixing bugs in lengthy codes. ChatGPT not only enhances coding efficiency but also reduces the workload and time demands placed on developers.
- g) *Legal Assistance*: ChatGPT is a valuable resource for legal professionals and individuals, providing legal assistance and advice tailored to various legal contexts. The model is able to reference case laws, recite legal documents and distinguish relevant laws. However, ChatGPT is suggested to be used judiciously as a consultative and informational tool, due to potential variations in applicability across different jurisdictions and the possibility of generating outdated information (Nazir & Wang, 2023, p. 8).

3. ChatGPT and Machine Translation

This research aims to conduct a comprehensive analysis of ChatGPT's performance in the domain of Machine Translation. The rapid evolution of LLMs and the advent of ChatGPT have profoundly

transformed the dynamics between translation technology and translation professionals. The model has proven to be a valuable assistant, streamlining workflows and raising important questions about the role of AI models in the industry. The AI-driven tool has garnered considerable attention from researchers in the field of translation studies and has been incorporated into the translation industry in various capacities. This study seeks to explore the model's capabilities, limitations, and optimal adaptation to determine how and whether ChatGPT can establish a secure and enduring role in the field.

To thoroughly understand ChatGPT's functionality as a translation engine, it is essential to examine the fundamental aspects of its translation capabilities. ChatGPT, among its diverse applications, possesses multilingual abilities, supporting at least 95 natural languages and offering natural user experience to over 4.5 billion people worldwide (Flensted, 2024). The model utilizes its extensive training data to generate Machine Translation outputs based on corresponding prompts. Machine Translation (MT) involves the utilization of computer algorithms to automatically translate text or speech from one language (source language) to another (target language) without any human intervention (Tabsharani, 2023).

Even though ChatGPT demonstrates promising translation capabilities, it is important to distinguish the model from commercial Neural Machine Translation (NMT)¹⁷ engines such as Google Translate¹⁸ and DeepL¹⁹. Both GPT models and NMT systems are based on transformer architecture; however, they exhibit significant differences in their architectural design. ChatGPT is a decoder-only model intended for text generation tasks, utilizing a single decoder stack to predict the next word in a sequence. On the other hand, NMT models employ an encoder-decoder architecture with attention

¹⁷ Neural Machine Translation (NMT) is a Machine Translation (MT) approach that utilizes neural networks to perform translation tasks. NMT models acquire rich representations of language structures through large amounts of bilingual data. NMT has significantly advanced the field of MT, offering superior fluency and accuracy in translation outputs compared to earlier approaches, such as Statistical Machine Translation (SMT) and Rule Based Machine Translation (RBMT).

¹⁸ Google Translate. (n.d.). Google. <https://translate.google.com/?sl=en&tl=el&op=translate>

¹⁹ DeepL Translate: The World's Most Accurate Translator. (n.d.). <https://www.deepl.com/en/translator>

mechanisms, specifically optimized for handling sequence-to-sequence tasks of varying lengths and structures (Hendy et al., 2023). Essentially, the GPT model behind ChatGPT is a general-purpose LLM trained to perform various language-based tasks, whereas NMT models are specifically tailored for translation, converting text from one language to another.

In addition, NMT models are typically trained on parallel corpora, which consist of text pairs from two different languages, while ChatGPT is primarily trained on general-domain monolingual corpora. This allows ChatGPT to be applied across a broader range of language pairs and translation tasks, facilitating easier adaptation to specific styles and registers, unlike NMT systems that are limited to preset language pairs and the encoded representations of the source language (Jiang et al., 2023). ChatGPT exhibits greater flexibility in adjusting to diverse contexts and is capable of generating more fluent and contextually aware translations, despite significant accuracy challenges. The chatbot can also refine its translation outputs, produce multiple translations of the same text, and adjust its translations to better align with their intended purpose.

NMT engines have revolutionized the field of MT by delivering high-quality translations that surpass the performance of earlier approaches, such as Statistical Machine Translation (SMT)²⁰ and Rule-based Machine Translation (RBMT)²¹. The superior performance of NMT models has established them as the preferred choice for practitioners in translation tasks. However, their dominance is now being challenged by the emergence of GPT models, particularly ChatGPT, due to its flexible and versatile character.

²⁰ Statistical Machine Translation (SMT) is a Machine Translation (MT) approach that relies on statistical models to generate translations. These models are trained on large volumes of bilingual data and learn to translate by analyzing the statistical relationships between original texts and their human translations (Statistical Machine Translation, 2023).

²¹ Rule-based Machine Translation (RBMT) was one of the earliest approaches to Machine Translation (MT). It relies on a comprehensive set of linguistic rules that describe the grammatical structures of both the input and output languages. RBMT systems utilize monolingual and bilingual dictionaries to map input terms to their corresponding output terms (Rule-based Machine Translation, 2023).

3.1 Related Work

Since the release of ChatGPT and its widespread adoption in professional contexts, a substantial body of published research has focused on evaluating the model as a Machine Translation (MT) tool. Recent studies have concentrated on assessing both the efficacy and limitations of ChatGPT in MT, as well as exploring its potential for optimal integration within the translation industry (Siu, 2023; Yamada, 2023). According to Jiang et al. (2023), ChatGPT has additionally attracted significant attention within Translation Quality Assessment (TQA)²² research, wherein the model's performance is systematically evaluated in comparison to NMT engines. The performance of ChatGPT in MT is studied across three primary axes: Comparative Machine Translation Evaluation, Prompt Engineering for Machine Translation, and Machine Translation Evaluation for High-resource and Low-resource Languages.

3.1.1 Comparative Machine Translation Evaluation

Initial evaluation studies of ChatGPT's performance in MT tasks have primarily involved comparative research methodologies to assess the model's efficacy and translation quality in comparison to leading commercial NMT engines that dominate the MT landscape. These comparative assessments, observed in both qualitative and quantitative research, have played an essential, albeit supplementary, role in validating hypotheses and evaluating strategic approaches for the optimization of ChatGPT for MT. In the majority of these evaluation studies, ChatGPT's translation quality scores have been systematically compared to those of well-established NMT systems; the most frequently utilized benchmark is Google Translate (Robinson et al., 2023; Peng et al., 2023; Moslem et al., 2023; Zhu et al.,

²² Translation Quality Assessment (TQA) is employed in both academic research and translation practice to assess translation quality, identify issues, and suggest improvements (Williams, 2009).

2023), followed by DeepL (Gao et al., 2023; Jiao et al., 2023; Yamada, 2023), Microsoft Translator²³ (Hendy et al., 2023), and Huawei Translate (Zhao et al., 2023).

Previous studies have primarily utilized neural network-based metrics and automatic evaluation metrics to assess the performance of ChatGPT in MT, with relatively few studies incorporating human evaluation (Hendy et al., 2023; Zhao et al., 2023). Within this evaluation framework, recent research findings indicate that ChatGPT has the potential to compete with, and in some cases outperform, conventional NMT systems across various language pairs, particularly when utilizing specific prompt techniques (Hendy et al., 2023). Nonetheless, several studies have identified a significant performance gap between ChatGPT and conventional MT systems. Even though ChatGPT exhibits competitive results in fluency and specific translation directions, it falls short in accuracy, multilingual competence, and domain-specific scenarios (Jiao et al., 2023). Therefore, a more comprehensive and objective evaluation of ChatGPT's MT capabilities necessitates the consideration of factors beyond comparative evaluation analyses. These additional factors, which will be thoroughly explored in subsequent sections, are imperative for a holistic understanding of ChatGPT's strengths and limitations in the MT domain.

At this point, it is essential to acknowledge the rapid evolution of ChatGPT and its underlying model architecture. As advancements in LLM technology continue, existing MT research findings may soon become outdated or insufficiently representative of the MT capabilities of newer GPT models. ChatGPT and its successors have swiftly achieved NMT capabilities and are expected to surpass commercial MT systems in the near future. This trajectory of continuous innovation and the advent of GPT-4 have set new benchmarks for MT. These advancements can be attributed to the model's advanced

²³ Microsoft Translator. (n.d.). <https://translator.microsoft.com/>

training data and its enhanced multilingual capacity, which facilitates higher quality of MT in a broader range of language pairs (Jiao et al., 2023).

3.1.2 Prompt Engineering for Machine Translation

ChatGPT is an intelligent and versatile conversational system designed to perform multiple NLP tasks, including translation. Unlike NMT models, which are specifically tailored for translation tasks, ChatGPT does not undergo specialized fine-tuning to enhance its translation capabilities. Since users and researchers cannot access or modify the internal components of ChatGPT for further specialization, the model is optimally treated as a black-box system that can be optimized solely through input data manipulation (Gao et al., 2023). In this framework, the natural language input data provided by users to elicit responses from the chatbot are referred to as “prompts²⁴.” The process of crafting suitable prompts to achieve desired outputs from a model is known as prompt engineering. Recent research underscores the critical role of prompts in harnessing the full potential of ChatGPT for MT, exploring prompting techniques to improve the quality of generated translations (Gao et al., 2023; Hendy et al., 2023; Jiao et al., 2023; Peng et al., 2023; Yamada, 2023; Wu & Hu, 2023; Zhang et al., 2023; Gu, 2023).

Hendy et al. (2023) first evaluated the performance of ChatGPT in MT, highlighting the impact of prompting strategies on translation output quality and the significance of the form, length and content of translation prompts. Jiao et al. (2023) initially designed three candidate simple prompts²⁵ to trigger the MT ability of ChatGPT, which included information about the source and target languages²⁶, as well as the direction of translation. These prompting structures were assessed in a multilingual context

²⁴ Prompts are different from tags. Tags are standardized labels employed to categorize data, informing and modifying systems. Unlike prompts, tags are not used directly in AI interactions; instead, they indirectly guide AI behavior.

²⁵ The candidate translation prompts proposed by Jiao et al. (2023) are: TP1: “Translate these sentences from [ST] to [TL].”, TP2: “Answer with no quotes. What do these sentences mean in [TL]?”, and TP3: “Please provide the [TL] translation for these sentences:”. Among the three prompts, TP3 performed best and was subsequently used by default.

²⁶ In this paper, the terms “source language” and “target language” will be referred to as SL and TL, respectively, for practical reasons.

alongside simple prompts inspired by the chatbot itself. Gao et al. (2023) and Peng et al. (2023) further developed these simple prompt structures, with the addition of subsidiary contextual information to boost the quality of the model's translations. Gao et al. (2023)²⁷ argued that "naive" prompts cannot fully unleash the MT capabilities of ChatGPT proposing prompt structures that contain translation task information, context domain information and Part-of-Speech (POS) tags, which outperformed standard prompts. Similarly, Peng et al. (2023)²⁸ suggested simple but more effective structures; Task-Specific Prompts (TSP) and Domain-Specific Prompts (DSP). Recent studies particularly attend to the robustness of MT under domain shift, highlighting the importance of domain-specific prompts. These prompts assist ChatGPT in generating more relevant translations and allow it to adapt to the particular terminology and nuances of the domain in question. In addition, Yamada (2023) investigates the feasibility of optimizing translation outputs for professional translators by integrating information about the purpose and target audience into prompts.

Multiple experimental prompting strategies have been investigated across various studies to evaluate the MT performance of ChatGPT after the employment of zero-shot and few-shot prompts. In this context, initial MT assessment research have utilized zero-shot prompt setups, wherein ChatGPT is assigned a task without prior explicit examples or training on similar tasks, relying solely on pre-existing knowledge (Hendy et al., 2023; Jiao et al., 2023). The aforementioned zero-shot prompts designed by Jiao et al. (2023) were extensively utilized by subsequent studies alongside few-shot prompting templates, in an attempt to achieve improved MT quality. In few-shot prompt settings, ChatGPT is

²⁷ Prompts proposed by Gao et al. (2023): TT (Translation Task): "This is a [SL] to [TL] translation, please provide the [TL] translation for these sentences:", CD (Context Domain): "Please provide the [TL] translation for these sentences taken from the [SPECIFIC DOMAIN]:"

²⁸ Prompts proposed by Peng et al. (2023): Task-Specific Prompt (TSP): ["role": "system", "content": "You are a machine translation system."], ["role": "user", "content": "Please provide the [TL] translation for the following sentence:"], Domain-Specific Prompt (DSP): ["role": "system", "content": "You are a machine translation system that translates sentences in the [DOM] domain."], ["role": "user", "content": "Please provide the [TL] translation for the following sentence:"]

provided with example translations that demonstrate the task, preparing and assisting the model to generate more accurate translations. Gao et al. (2023) and Robinson et al. (2023) experimented with one-shot and 5-shot templates, which outperformed baseline prompts and achieved better MT performance (see Table 2). Moreover, Peng et al. (2023) investigated zero-shot and few-shot templates to evaluate ChatGPT's in-context learning (ICL)²⁹ (Brown et al., 2020) and chain-of-thought (CoT)³⁰ abilities (Wei et al., 2022), respectively, for MT application.

Few-shot prompting techniques have been extensively employed in recent work, demonstrating improved outputs in MT tasks by effectively leveraging the translation capabilities of ChatGPT. However, previous studies argue that few-shot prompts offer only marginal improvements in translation quality compared to zero-shot prompts (Hendy et al., 2023; Moslem et al., 2023). In fact, specific prompting setups often find these improvements to be insignificant. Robinson et al. (2023) conducted a large-scale comparative study evaluating the quality of translations generated using both few-shot and zero-shot prompts within a multilingual and multidimensional framework. Their findings indicate that zero-shot prompting with ChatGPT generally surpasses few-shot templates in terms of time and cost efficiency. Although few-shot prompts achieve slightly higher evaluation scores, the incremental improvements do not justify the additional cost and time required in large-scale utilization (Robinson et al., 2023).

²⁹ Few-shot in-context learning (ICL) methods were introduced by Brown et al. (2020) enabling LLMs to generalize from their pre-existing knowledge and adapt to new tasks dynamically. Hendy et al. (2023) and Peng et al. (2023) argue that in-context learning strategies with additional labelled prompts can enhance the MT performance of ChatGPT.

³⁰ Wei et al. (2022) introduced the chain-of-thought (CoT) prompting technique to elicit the reasoning capabilities of LLMs. They designed prompts containing examples in which the reasoning process is demonstrated step-by-step, assisting the models to "think" through problem solving and NLP tasks. Peng et al. (2023) applied CoT prompts to MT tasks: "Please provide the [TL] translation for the following sentence step by step". In this study the CoT prompting strategy was deemed incompatible with MT tasks for ChatGPT, as it resulted in lower performance and encouraged a word-by-word translation approach.

Consequently, zero-shot prompts are recommended as they offer satisfactory MT results while being more affordable and practical for users (Gao et al., 2023).

Zero-shot Translation Prompt	Few-shot Translation Prompt (5-shot)
Input: This is an English to [TL] translation, please provide the [TL] translation for these sentences: [SL]: [source sentence] [TL]: Output: [TL]: [target sentence]	Input: This is an English to [TL] translation, please provide the [TL] translation for these sentences: [SL]: [source sentence] [TL]: [target sentence] [SL]: [source sentence] [TL]: [target sentence] [SL]: [source sentence] [TL]: [target sentence] [SL]: [source sentence] [TL]: [target sentence] [SL]: [source sentence] [TL]: [target sentence] Please provide the translation for the following sentence: [SL]: [source sentence] [TL]: Output: [TL]: [target sentence]

Table 2: Examples of sentence-level prompt templates (Jiao et al., 2023; Robinson et al., 2023; Gao et al., 2023)

Recent studies have introduced another prominent debate related to the design of prompts and the presentation of input information to fully leverage the MT capabilities of ChatGPT. This discussion pertains to whether a sentence-level or document-level approach yields higher quality translations. Most studies have concentrated on sentence-level MT with ChatGPT for practical evaluation reasons. Document translation with LLMs remains understudied and was initially examined by Wang et al. (2023) and Wu & Hu (2023), who investigated the effectiveness of ChatGPT and GPT-3.5 in preserving document-level coherence and consistency. Hendy et al. (2023) also explored the application of ChatGPT to document-level MT, anticipating enhanced translation results due to the model's access to the context and coherence of the entire document to generate more relevant translations. In specific few-shot prompt setups, document level MT outperformed sentence translation across metrics but struggled to keep sentence-level alignment intact.

3.1.3 Machine Translation Evaluation for High-resource and Low-resource Languages

ChatGPT is widely recognized for its multilingual capabilities, which have quickly been utilized in MT applications across diverse linguistic contexts. Naturally, the model's performance varies significantly among language pairs, giving way to an intriguing area of research. More specifically, a substantial branch of MT evaluation studies, which have conducted comparative multilingual assessments on ChatGPT, agrees that the model performs competitively with commercial translation engines in high-resource languages, but is especially disadvantaged for low-resource languages (Jiao et al., 2023; Hendy et al., 2023; Robinson et al., 2023; Zhu et al., 2023; Manakhimova et al., 2023). By "resource," we refer to the amount of training data, such as monolingual or parallel corpora, which are essential for developing ChatGPT's multilingual proficiency. High-resource languages are considered those with abundant linguistic data available for training, sourced from diverse domains and contexts. Previous studies and statistical analyses of training datasets, such as Wikipedia and Common Crawl, indicate that most high-resource languages are of European and Asian origin; the most prominent among these languages include English, Spanish, German, French, Chinese, and Japanese (Gao et al., 2023; Pelucchi, 2023; Robinson et al., 2023; Zhu et al., 2023). On the other hand, low-resource languages are underrepresented in digital resources, possessing limited amounts of linguistic data, and are predominantly from the Afro-Asiatic language family (Laumann, 2023). Robinson et al. (2023) highlights that the resource level of a language is the most critical determinant of ChatGPT's ability to translate to and from that language.

Hence, the primary cause of ChatGPT's suboptimal performance in multilingual MT is the limited number of high-resource languages and the model's English-centric design. By "English-centric," we refer to the predominance of the English language in the development and training of the model, indicating that ChatGPT is primarily optimized for tasks involving the English language. This English-

centric approach, which reflects various linguistic and cultural biases, extends to the field of MT research, where most evaluation studies focus on language pairs that include English.

Preliminary evaluation studies of ChatGPT as a MT tool have primarily involved small-scale comparative research, assessing the model's MT capabilities with high-resource language pairs centered around English, while giving limited attention to low-resource languages. Jiao et al. (2023) established initial MT benchmarks for 5 languages, including one low-resource language. Similarly, Hendy et al. (2023) provided MT evaluation data for 18 language pairs, comprising 6 high-resource and 4 low-resource languages from different language families. Both studies recognize ChatGPT as a competent MT tool but confirm a significant performance disparity between high-resource and low-resource languages. Furthermore, they both emphasize the critical role of the language family factor, indicating that MT between different language families is particularly challenging, especially for low-resource language pairs, thereby exacerbating the performance gap. Nonetheless, these evaluation data are not fully indicative of ChatGPT's MT capabilities due to the selective and limited scope of the research.

Subsequent large-scale and cross-lingual studies have provided a more comprehensive depiction of ChatGPT's MT performance. Zhu et al. (2023) evaluated popular LLMs, including ChatGPT and GPT-4, by assessing their performance across 102 languages, categorized by language family. Within this extensive multilingual framework, the study confirmed the model's success in high-resource scenarios while highlighting their significant underperformance in translating distant, low-resource languages compared to commercially available MT systems such as Google Translate. The study also underscored ChatGPT's relative struggles with non-English MT directions, indicating the potential impact of directionality in MT quality. In other words, ChatGPT exhibits superior results in English-directed translations compared to translations from English to other languages, particularly in low resource contexts. In a similar large-scale study, Robinson et al. (2023) evaluated ChatGPT across 204 languages,

including 168 low-resource languages. Their findings corroborated those of previous works, remarking that ChatGPT significantly lags behind in low-resource MT, particularly for low-resource language pairs and domain-specific contexts. The study suggested that NMT models remain the preferable choice for low-resource scenarios, noting that the rapidly advancing LLMs are gradually acquiring expanded multilingual capabilities and are approaching high MT quality scores.

Previous studies have investigated methods to enhance MT quality in low-resource scenarios through prompt engineering. Jiao et al. (2023) introduced an innovative strategy termed pivot prompting³¹, wherein ChatGPT is instructed to initially translate the source text into a high-resource intermediary language (pivot language) before translating it into the target language. This technique aims to improve ChatGPT's translation performance across a broader range of language pairs. The employment of a high-resource pivot language helps to effectively impart critical information to the target text, thereby mitigating the information loss that typically occurs in low-resource scenarios due to insufficient training data. Preliminary experimental results demonstrated that pivot prompting significantly enhances ChatGPT's translation performance, suggesting that it is a promising strategy for achieving high-quality multilingual MT.

Hendy et al. (2023) proposed an alternative method of obtaining high-quality translations for distant, underrepresented language pairs through the employment of hybrid strategies, which combine the strengths of both GPT and NMT models. Their study introduced the "Hybrid Max-Routing" approach, wherein an NMT model serves as the primary MT system, with a GPT model acting as a fallback option. The optimal translation is selected based on COMETkiwi scores. This technique consistently achieves higher evaluation scores, indicating that although NMT systems generally surpass GPT models in

³¹ The prompt proposed by Jiao et al. (2023) for the pivot prompting strategy is: "Please provide the [PIVOT LANGUAGE] translation first and then the [TL] translation for these sentences one by one:"

performance, GPTs exhibit superior efficacy in specific language pairs. These findings explore the potential for further utilization of hybrid techniques to enhance translation quality across a diverse array of language pairs.

3.2 The Case of the Greek Language

The primary objective of this paper is to conduct a small-scale evaluation of the MT capabilities of ChatGPT, with a particular emphasis on the English-Greek language pair. This research seeks to explore and implement optimization strategies to enhance the model's performance, thereby assessing its potential to generate high-quality translations in Greek. Considering the English-centric architecture of ChatGPT and the limited number of high-resource languages supported by the model, it is reasonable to assume that a less widely spoken language, such as Greek, is underrepresented.

OpenAI has not publicly disclosed detailed information regarding the volume of resource data utilized for individual languages supported by ChatGPT. Recent studies highlight that the Greek language struggles with limited digital representation, which is marked by a scarcity of linguistic resources, consequently leading to limited datasets and knowledge bases for NLP applications (Antonopoulos, 2023). This insufficiency poses significant challenges for Natural Language Understanding (NLU) in Greek, primarily due to the high degree of linguistic ambiguity inherent in the language (Papantoniou & Tzitzikas, 2020).

Additionally, there is a paucity of corresponding research on deep learning models tailored to the Greek language, with only a few initiatives aimed at developing resources and datasets for model training. The most frequently utilized corpora for Greek-specific training include Wikipedia, the European Parliament Proceedings Parallel Corpus (EuroParl)³², CommonCrawl³³, and low-quality datasets

³² EuroParl Parallel Corpus. (n.d.). <https://www.statmt.org/europarl/>

³³ Common Crawl - Open Repository of Web Crawl Data. (n.d.). <https://commoncrawl.org/>

consisting of machine translated texts (Evdaïmon et al., 2023; Giarelis et al., 2023). Leveraging this large-scale corpus, Evdaïmon et al. (2023) pretrained GreekBART, the first sequence-to-sequence model for the Greek language, intended for Natural Language Generation (NLG) tasks, including translation. However, this model, along with similar research efforts, remains in the preliminary stages of development, underscoring the need for more comprehensive Greek-specific training corpora and advanced modeling techniques.

The recent body of research demonstrates a notable absence of research focused on evaluating the performance of ChatGPT in the context of the Greek language, particularly in terms of its NLP and MT capabilities. Within the domain of MT evaluation, the Greek language has been examined solely through large-scale comparative evaluation studies as part of broader multilingual projects. Robinson et al. (2023) employed automatic evaluation metrics to compare the MT performance of ChatGPT with that of Google's NMT model. Their findings indicated that, for the Greek language, ChatGPT's translation quality was highly competitive, albeit approximately 3% lower on average compared to the NMT model.

Similarly, Zhu et al. (2023) included Greek in a comprehensive comparative study assessing the translation quality among ChatGPT, GPT-4, and Google Translate in both translation directions, thereby providing additional insights into the English-Greek language pair. Their findings suggest that ChatGPT is highly competitive with NMT for Greek, with an average quality score approximately 3.5% lower than NMT in both translation directions, as determined by automatic metrics. In this study, GPT-4 demonstrated marginally enhanced performance for the Greek language compared to ChatGPT, with an improvement of about 1.7% in both translation directions. Regarding directionality, the difference in performance between Greek-to-English and English-to-Greek translations was marginal, with English-directed translations showing a slight improvement of 3.5% compared to scenarios where Greek was the target language. These findings suggest that the factors mentioned above may affect MT performance in

high-resource contexts, but they are not as impactful for less represented languages like Greek. Additionally, the disparity in MT performance between ChatGPT and NMT, as well as between translation directions, may be negligible, primarily due to the limited and low-quality resources for training these models for Greek.

On the whole, both studies, through their data analyses, classified Greek as a medium-resource language, characterized by limited but fundamental resources to facilitate MT tasks. However, due to the scarcity of research data, we cannot draw accurate conclusions about ChatGPT's MT performance in Greek. This paper aims to address this gap by providing a more thorough evaluation of ChatGPT's capabilities in processing and translating the Greek language.

4. Research Methodology

The primary objective of this research is to explore the MT capabilities for the English-Greek language pair by implementing a variety of prompting strategies to gain comprehensive insights into the model's translation proficiency for Greek, identify frequent errors, and assess its potential viability and applications in the translation industry. It is essential to remark that, given the researchers' inability to access or modify the internal components of ChatGPT for further specialization, the model is optimally treated as a black-box system that can be optimized for specific tasks solely through input data manipulation and prompt engineering (Gao et al., 2023). In order to form our outcome expectations accordingly, we acknowledge that ChatGPT's training data for the Greek language are presumed to be limited and of unknown quality, thereby classifying Greek as a medium-resource language. This research adopts a small-scale approach, dictated by practical constraints, imparting an explorative and experimental nature to our objectives.

We carefully constructed a succinct experimental layout to evaluate the translation capabilities of ChatGPT 3.5. This research is designed based on widely accessible resources to facilitate effortless application by individuals and professionals across various Greek language contexts. Despite the release of GPT-4, which demonstrates expanded multilingual capacity and slightly enhanced performance compared to its predecessor in Greek MT assessments, we opted to evaluate ChatGPT 3.5 due to the marginal performance disparity and its open-access platform, which caters to a broader user spectrum (Ray, 2023). Additionally, GPT-4 incurs significantly higher costs, particularly for large-scale utilization in MT tasks and professional settings, rendering ChatGPT a more viable option for this study (Robinson et al., 2023).

For similar reasons, we have employed carefully designed zero-shot prompting strategies to assess and optimize ChatGPT's MT capabilities for the Greek language. Zero-shot baseline prompts were preferred over few-shot strategies due to their superior practicality and convenience for individual use. According to Robinson et al. (2023), ChatGPT's evaluation scores for English-to-Greek MT using both zero-shot and five-shot prompts, demonstrate null differences in translation quality. Furthermore, previous research by Hendy et al. (2023) discourages the employment of few-shot prompting techniques to acquire optimal translation quality in low-resource scenarios, highlighting that such methods are more complex, more costly, and provide minimal quality enhancement.

The experimental framework of this small-scale research is designed to evaluate the English-to-Greek translations of three texts originating from distinct domains. Given the slight performance disparity associated with the directionality of translation, our focus was directed towards one-way MT (Zhu et al., 2023). The translations generated by ChatGPT were produced using simple baseline prompts and subsequently evaluated through both human evaluation and automatic evaluation metrics to ensure a comprehensive analysis of translation quality.

4.1 Domains

In this research, three distinct paragraph-level texts from different domains were utilized to evaluate ChatGPT's MT capabilities and adaptability to domain-shift in MT for the Greek language. The domain typology employed in our research aligns with the text type classification framework outlined by Hendy et al. (2023). The selected dataset primarily encompasses texts sourced from online resources, with an average word count of 135 words per text. Listed below are the text domains employed in this study: Biomedical³⁴, News³⁵, and e-Commerce³⁶

- *Biomedical*: we have included a text from a medical journal within the biomedical and pharmaceutical domain. The provided text encompasses procedural details, medical terminology and technical descriptions, aiming to provide specialized information pertinent to medical education and research. The primary objective is to assess the performance of ChatGPT's MT in highly technical contexts and its ability to maintain technical accuracy and adhere to a formal register in translation, ensuring that the target text aligns with the original purpose of the text. Additionally, this evaluation allows us to indirectly assess the quality and diversity of the model's training data for a medium-resource language such as Greek.
- *News*: we have included a journalistic article characterized by an informative and political style that discusses recent political events. This text presents political implications, different perspectives, and cultural nuances while maintaining a formal and objective tone. Our goal is to assess ChatGPT's ability to transfer all available information clearly and objectively into the

³⁴ Patel et al., (2018). Histopathological Analysis of Hysterectomy Specimens in Tertiary Care Center: Two Year Study. Tropical Journal of Pathology and Microbiology. <https://doi.org/10.17511/jopm.2018.i01.06>

³⁵ Walker & Mason. (2024, March 29). Tory donor's knighthood is sign Sunak 'believes he's on way out', Labour says. The Guardian. <https://www.theguardian.com/politics/2024/mar/29/tory-donor-knighthood-sign-sunak-way-out-labour-mohamed-mansour>

³⁶ Libre Eau de parfum Women's Perfume. (n.d.). YSL. https://www.yslbeautyus.com/fragrance/womens-fragrances/libre/libre-eau-de-parfum/109YSL.html?fbclid=IwAR0B9BjPxAWDxSzRXj9_1UqfG_eznukYxsCsdP3BEg25X1MPXEILe0CfA8

target language, accurately translate associated terminology and denominations, and adjust to the cultural and linguistic framework of the target language and audience.

- *e-Commerce*: we have included an online advertisement for a product that is targeting a gender-specific audience within a marketing context. This text employs persuasive and figurative language to attract customers and align with the brand's marketing strategies. Our objective is to evaluate ChatGPT's versatility in handling informal language and adapting its MT capabilities to the text's persuasive style and task-specific character, while ensuring technical accuracy in the target text.

4.2 Zero-shot Experimental Layout

One of the primary goals of this paper is the investigation of ChatGPT's MT capabilities for the English-Greek language pair, with a particular focus on accessibility and efficacy. To gain a comprehensive and holistic understanding of ChatGPT's performance in Greek translation, we have employed simple instruction prompts for the initial and most significant part of our research. Specifically, we opt to initially trigger the translation ability of ChatGPT through baseline zero-shot prompts for paragraph-level translation, mirroring inputs that would be typically used by both individuals and professionals to initiate a MT conversation in the chatbot.

In our effort to simulate realistic scenarios, we have adopted a candidate prompt introduced by Jiao et al. (2023); "Please provide the [TL] translation for these sentences:" (see Figure 3). This prompting technique has been selected among other candidates due to its optimal application with Greek language and its frequent utilization and modification in previous studies (Gao et al., 2023; Peng et al., 2023). Notably, our preliminary test runs with similar prompts have indicated that the disparity in quality among the translation outputs they triggered was minimal, primarily differing in phrasing and

presentation. This suggests that even impulsive and unplanned prompts can elicit translations of comparable quality.

4.3 Automatic Evaluation

In the initial phase of evaluating the obtained translation outputs of ChatGPT, we have utilized reference-based automatic evaluation metrics. These metrics assign a score to the MT output based on its similarity to high-quality human translations, referred to as reference translations (Chatzikoumi, 2019). To facilitate this evaluation process, we have employed human translations of our own production for the target English texts. It is important to remark that reference translation-based metrics are well-regarded for their effectiveness in large-scale evaluation contexts, but they are also known to reproduce biases due to their inherent reliance on reference translations.

Please provide the Greek translation for these sentences: [A celebration of freedom, this daring, yet distinctly floral and feminine perfume is for those who do what they want and dare to be exactly who they are. Breaking boundaries and paving their own way, this women's perfume is inspired by the bold and the free.]



Certainly! Here is the Greek translation for the provided sentences:

Μια γιορτή της ελευθερίας, αυτό το τολμηρό, αλλά ξεκάθαρα λουλουδάτο και θηλυκό άρωμα είναι για εκείνους που κάνουν ό,τι θέλουν και τολμούν να είναι ακριβώς αυτοί που είναι. Σπάζοντας όρια και ανοίγοντας το δικό τους δρόμο, αυτό το γυναικείο άρωμα εμπνέεται από τους τολμηρούς και τους ελεύθερους.

Figure 3: Example of the zero-shot prompt (Jiao et al., 2023)

Since we have employed different categories of automated metrics in the evaluation process, it is essential to distinguish automatic metrics from neural-network based metrics. In contrast to traditional automatic metrics which solely rely on rule-based algorithms to compare the system output with reference translations, the architecture of neural metrics is more complex, facilitating text quality

assessment through machine learning architectures. Additionally, neural evaluation metrics exhibit a significantly higher correlation with human judgements compared to traditional metrics, which often present particular subjectivity due to their dependence on reference translations (Rei et al., 2023). The advanced architecture of neural metrics allows them to capture deeper semantic and contextual information, thereby providing evaluation scores that more closely align with human assessments. However, these metrics are computationally intensive and are often regarded as black-boxes, offering quality scores without transparency in their decision-making methodology. Conversely, traditional metrics yield evaluation scores that are easier to interpret due to their straightforward architecture (Chatzikoumi, 2019).

In light of practical constraints, we leveraged the capabilities of MATEO³⁷, an online open-source machine translation evaluation platform introduced by Vanroy et al. (2023) which facilitates the application of various automatic metrics without necessitating specialized expertise. The platform's user-friendly web interface enables efficient evaluation of machine-generated translations, thereby enhancing digital literacy among both experts and novices. In our research, we employed three automatic metrics: BLEU, BLEURT, and COMET.

- *BLEU*: The Bilinual Evaluation Understudy (BLEU) is an automated statistical metric that has been extensively utilized as the standard for MT evaluation across various studies (Guzmán et al., 2017). BLEU assesses MT quality by determining the degree of similarity between the machine-generated translation and the reference translation, focusing on the fraction of n-grams accurately transferred to the target text (Papineni et al., 2002). This metric quantifies this closeness, with higher scores indicating better MT performance. Additionally, BLEU incorporates a brevity penalty to penalize translations that diverge significantly in length from the reference

³⁷ Evaluate Machine Translations. (2023). MATEO. <https://mateo.ivdnt.org/Evaluate>

translation. Despite the metric's widespread adoption as a MT evaluation benchmark, it lacks semantic understanding, often undervaluing high-quality translations that deviate from the reference texts. In this study, the BLEU metric is employed to evaluate ChatGPT translations due to its prevalent application, thereby facilitating the compatibility of our results with those of similar studies.

- *BLEURT*: Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) represents a milestone in MT evaluation metrics. It is a neural network-based metric built upon the pretrained model BERT and fine-tuned specifically for quality evaluation tasks (Sellam et al., 2020). In contrast to the n-gram based BLEU metric, BLEURT is capable of capturing the contextual and semantic nuances of the text enabling it to predict human-like evaluation scores with improved accuracy. This metric's enhanced ability to reflect human judgment makes it particularly suitable for research contexts where precise alignment with human evaluation is critical, like our research.
- *COMET*: Crosslingual Optimized Metric for Evaluation of Translation (COMET) is another neural-based MT evaluation metric that leverages deep learning and pretrained language models to assess and score translation outputs (Rei et al., 2020). Similar to BLEURT, this metric excels in deep contextual understanding and is extensively employed in MT evaluation research. COMET is included in our evaluation layout as it consistently demonstrates improved correlations with human assessments.

4.4 Human Evaluation

In the second phase of this study's MT evaluation process, we have employed a multifaceted manual human evaluation methodology to assess the quality of ChatGPT's translation outputs. This approach complements automatic metrics to acquire deeper insights into ChatGPT's performance in

English-to-Greek translation. To facilitate human evaluation in a timely and cost-effective way, we have designed and distributed a questionnaire via Google Forms. The questionnaire's layout comprises three main sections per text domain, providing annotators with the source text, the machine-generated output, and sentence-level evaluation questions.

The human evaluation methodology utilized is a hybrid approach that integrates both directly expressed judgment (DEJ-based) and not directly expressed judgement (non-DEJ-based) evaluation methods (Chatzikoumi, 2019). DEJ-based methodologies involve human evaluators explicitly expressing their judgements about the quality of translations through direct feedback, such as ratings and rankings, while the evaluation process for non-DEJ-based methods is more analytic and task oriented. In this study, we have structured our evaluation setup into two distinct but interconnected parts: error classification tasks and DEJ-based assessment tasks. Annotators were instructed to thoroughly interact with the texts before providing quantitative quality ratings, thus facilitating a more descriptive and consistent evaluation framework.

A critical factor in guaranteeing the reliability of the evaluation process is the profile of the annotators, namely their linguistic proficiency and level of expertise. In our research, 93% of the judges were native Greek speakers with native or professional working proficiency in English. This bilingual capacity is crucial, enabling the majority of the evaluators to effectively assess the machine-generated outputs in the bilingual questionnaire format.

Our study has received evaluation responses from 29 bilingual annotators, comprising 59% professionals in the field of translation and 41% students or recent graduates from translation-related disciplines. Previous research underscores the critical role that professional translators play in MT assessments, highlighting their ability to deliver reliable judgments and promote interannotator agreement (Läubli et al., 2018). On the other hand, in various evaluation methodologies, inexperienced

annotators, such as students, may necessitate training with evaluation guidelines and examples to provide high-quality evaluations. To ensure the reliability of quality scores and broaden our pool of candidate annotators, we have organized our methodology to incorporate student evaluations as well. Consequently, we have devised a universally accessible evaluation framework designed to promote compatibility among annotators' scores regardless of their level of expertise, thereby reinforcing the robustness of human evaluation ratings.

Upon analyzing the evaluation scores, we have observed a high degree of compatibility in quality assessments between practitioners and translation students, with only minor inconsistencies (see Figure 4). Notably, professional translators exhibited a stricter evaluation of ChatGPT's translation outputs, assigning scores that were, on average, 6% lower compared to those assigned by translation students. This discrepancy can be attributed to the students' relative lack of experience, familiarity with quality standards and excessive exposure to MT outputs that professionals possess. Nonetheless, the minimal variation in quality scores suggests that the overall evaluations are sufficiently consistent, allowing for comparative evaluation of ChatGPT's MT capabilities.

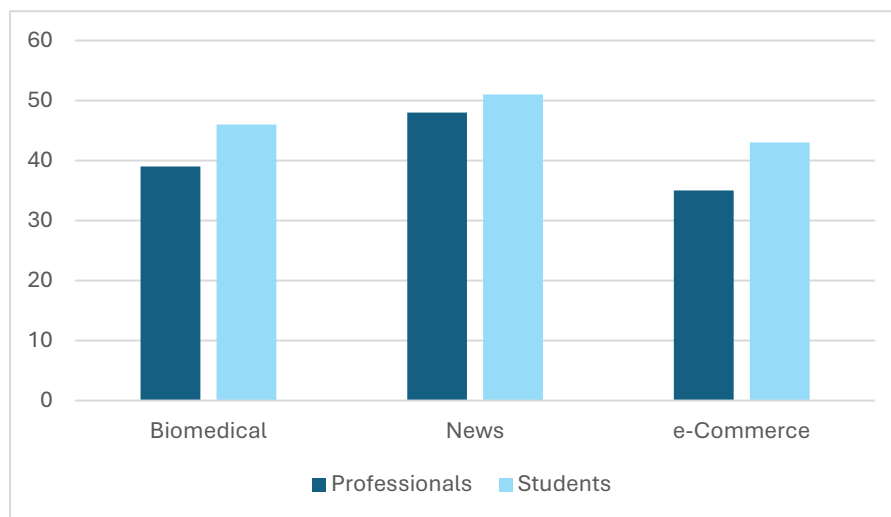


Figure 4: Comparison of evaluation scores between professional translators and translation students

4.4.1 Error Classification Task

In the initial phase of our human evaluation methodology, we employed a widely utilized metric – the error classification metric – to familiarize annotators with the texts they would subsequently assess. This methodology is intended to minimize evaluation score discrepancies, allowing annotators to engage with the texts in a more direct manner. Given that ChatGPT functions as a black-box system, we have utilized this method to acquire more information about the generated outputs. The three machine-translated texts were segmented into individual sentences, each paired with the corresponding source sentence. Annotators were tasked with identifying and classifying errors within eight predefined error categories for each segment of the translations.

Prior to designing the error classification task, we established a comprehensive error taxonomy, drawing inspiration from the error typologies outlined in Popović (2018). The error taxonomy employed in the evaluation process is a simplified version of previous taxonomies, designed to provide an inclusive and descriptive framework that would enhance interannotator agreement. The importance of a well-defined and personalized error typology cannot be overstated, as it is crucial for ensuring high-quality evaluations. The design of our error taxonomy was tailored to the needs of our research, including eight generalized error classes specifically designed to trigger desired quality assessments. These classes were carefully described as easily distinguishable, thereby facilitating effortless and time efficient evaluation and reducing the cognitive load on annotators. The error taxonomy we employed includes the following error classes: terminology errors/mistranslated text, morphological errors, syntactic errors, spelling errors, punctuation errors, errors in style, untranslated text, and unintelligible text (Popović, 2018).

Error Classification
<i>What kind of errors can you identify in the translation below?</i> <i>[Source Sentence]</i> [Target Sentence] ☐ Terminology errors (e.g., inaccurate medical terms) [Biomedical] Or Mistranslated text (e.g., inaccurate terms, inaccurate denominations) [News, e-Commerce] ☐ Morphological errors (e.g., casing, tense) ☐ Syntactic errors (e.g., word order) ☐ Spelling errors ☐ Punctuation errors ☐ Errors in style (e.g., higher or lower register) ☐ Untranslated text ☐ Unintelligible text ☐ I cannot identify any errors ☐ Other:

Table 3: Error Taxonomy

To efficiently guide annotators, we have provided additional information to specific error classes in order to limit ambiguity. Additionally, we have incorporated two supplementary options for cases where annotators could not identify erroneous words or phrasing, and for errors that did not fit the predefined categories. It is noteworthy that “terminology errors” and “mistranslation” often elicit ambiguous responses, particularly when both categories are present in the error typology scheme. This ambiguity is especially pronounced among annotators who lack professional training. To resolve this issue and disambiguate the two error classes, we restricted the option for “terminology errors” to the biomedical domain, where technical and specialized terminology is prevalent. This decision aimed at maintaining accuracy in phrasing while promoting high quality evaluations.

4.4.2 DEJ-based Evaluation Task

Following the error classification task, we have adopted an established human evaluation methodology introduced by Callison-Burch et al. (2007). They proposed a direct assessment metric where human evaluators rate translations in two separate five-point scales, focusing separately on accuracy and fluency. Accuracy measures the correctness and completeness of the information conveyed, while fluency refers to the naturalness and readability of the language used. This dual-scale approach facilitates nuanced evaluation, allowing for clearer distinction between content accuracy and linguistic quality (Callison-Burch et al., 2007). To enhance the precision of the quantitative assessments we integrated this direct assessment framework into a hybrid evaluation methodology. Namely, this metric follows a comprehensive analytical task, thereby mitigating potential inconsistencies in quality scores. In this research, we further refined this evaluation metric by incorporating modifications proposed by Chatzikoumi (2019), as illustrated in Table 4.

Accuracy	Fluency	Severity of Errors
<i>How much of the content from the source text does the translation convey?</i> ☐ Everything ☐ Most of the content ☐ Big part of the content ☐ Small part of the content ☐ None	<i>The language in the translation output is:</i> ☐ Perfect ☐ Good ☐ Not natural ☐ Ungrammatical ☐ Unintelligible	<i>How would you rate the severity of the errors?</i> 1 ⇔ 5 Null ⇔ Critical

Table 4: Directly expressed judgment task

In addition to the accuracy and fluency assessment, we have added an evaluation component focused on the severity of errors. Error severity, often overlooked in conventional human evaluation methodologies, refers to the level of critical errors made by the MT system and is crucial for a holistic understanding of translation quality (Sharou & Specia, 2022). Following the framework suggested by

Lommel et al. (2024), we adopted a modified version of the MQM error severity scale. This enhanced scale includes five categories: null, minor, moderate, major, and critical. This adaptation aligns with our previous assessment questions, ensuring a cohesive and comprehensive evaluation framework.

5. Key Findings

Upon conducting a comprehensive analysis of both human evaluation and automatic evaluation scores, we delve into the key findings that emerged from our study. Despite the limited scope of our research, which involved evaluating a relatively small dataset, the resulting outcomes do not fall behind large-scale studies, marking averagely 6% lower evaluation quality scores than Zhu et al. (2023) for the Greek language.

The primary objective of this research is to evaluate the MT capabilities of ChatGPT for the English-Greek language pair, thereby gaining a nuanced understanding of the model's resources and capacity concerning the Greek language. In this study, Greek is classified as a medium-resource language for ChatGPT, yielding moderate performance in cross-domain translation. This classification aligns with the status of an underrepresented language with lower accuracy in specific domains and slightly improved fluency scores, corroborating the outcomes of previous ChatGPT MT evaluation studies.

Our investigation reveals that, although Greek is not as robustly represented in the training data of ChatGPT as higher resource languages, the model still demonstrates a commendable ability to handle translation tasks. The identified limitations highlight the need for further enhancement of the model's proficiency in Greek, particularly in specialized domains where accuracy remains suboptimal. The findings contribute to the broader discourse on the performance of language models in medium-resource languages and underscore the ongoing challenge of improving linguistic accuracy and fluency in underrepresented languages like Greek.

5.1 Error Classification Analysis

Having carefully employed the outlined evaluation methodology, we have identified several noteworthy patterns and implications concerning the performance of ChatGPT. In the human evaluation phase, particularly in the error classification task, annotators explicitly identified multiple errors; the biomedical text elicited the highest number of remarks indicating significant challenges in this particular domain. On the other hand, texts from the News and e-Commerce domains exhibited a more balanced distribution of errors across different error categories. Through this comparative analysis we highlight domain-specific variations in error frequency and type, offering valuable insights into the model's strengths and limitations across diverse content areas.

In the Biomedical domain, our findings indicate that 70% of the annotators detected terminology-related errors, highlighting ChatGPT's significant limitations in providing reliable Greek translations in specialized and demanding scenarios. Through the employment of a medical text in the evaluation process, we opted to assess ChatGPT's adaptability to specialized domains, evaluating the model's flexibility in conveying information it cannot fully process. Despite delivering the general idea of the text in a medium-resource language, ChatGPT has mistranslated several concepts and terms. Additionally, the outputs were fraught with syntactic and morphological inaccuracies as noted by 66% and 40% of the annotators, respectively. These challenges led 34% of annotators to identify unintelligible phrasing, likely compounded by terminological errors. Furthermore, 26% of the annotators have highlighted stylistic inadequacies, deeming the translation inappropriate for the biomedical context. Overall, these challenges suggest that ChatGPT is still at an early stage in providing high-quality translations in Greek for specialized domains, where precision is of the utmost importance.

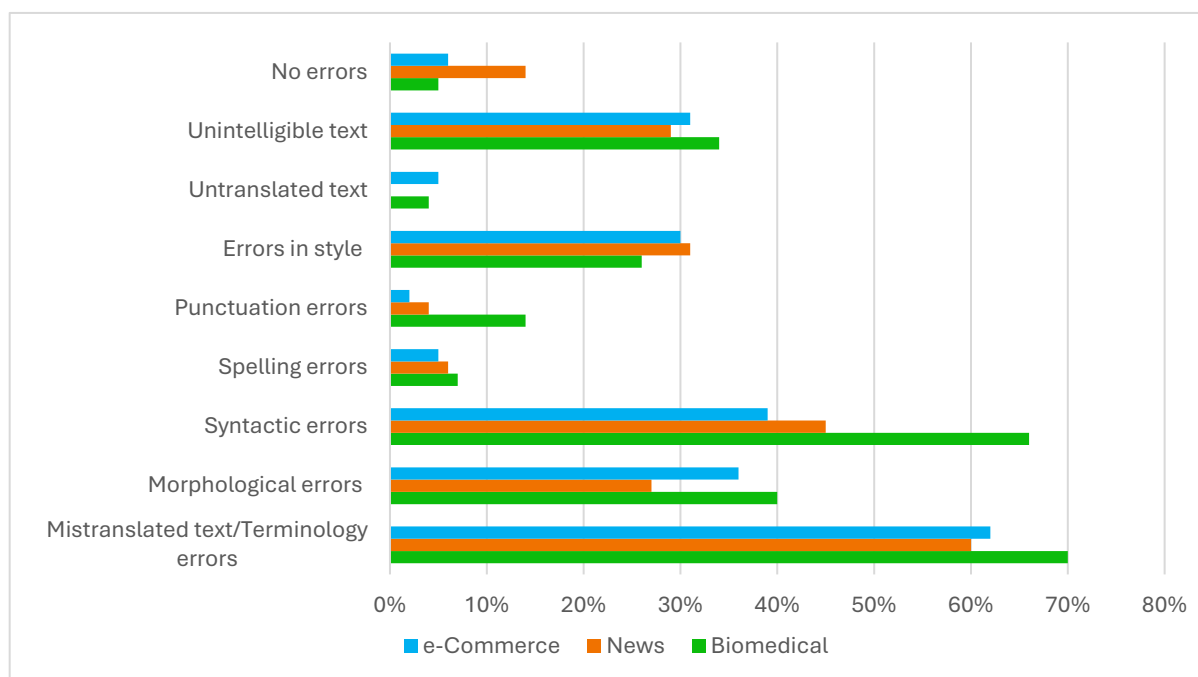


Figure 5: Error evaluation task scores

Similarly, in the News domain, the performance of ChatGPT is deemed moderate by the majority of evaluators. The text, rich in denominations and political references, has posed significant challenges for accurate transfer and transliteration into Greek, reflecting the model's limited training resources and knowledge base. Notably, 60% of annotators identified instances of mistranslated text, classifying these into the corresponding error class. Additionally, the model exhibits numerous syntactical and morphological errors, with 45% and 27% of evaluators selecting these categories, respectively. However, the incidence of such errors is lower compared to the Biomedical domain, indicating that ChatGPT performs relatively better in less rigid text types where some degree of flexibility is permissible. The machine generated output was also criticized in terms of style, with 31% of annotators pointing out an inappropriate transfer of the style of the journalistic article. Interestingly, the journalistic text is ranked first in the option for no identified errors, according to 14% of the judges, indicating it is the text with the fewest erroneous instances.

In the e-Commerce domain, the translation quality of the machine-generated output is similarly compromised by instances of mistranslations, as identified by 62% of the evaluators. The advertisement text features figurative expressions, gender-specific details, and an informal style, which the model has failed to transfer optimally to the Greek target text. This has resulted in 30% of the annotators identifying stylistic errors and 31% spotting unintelligible phrasing in various segments. Consistent with previous domains, the model exhibits a high degree of syntactical and morphological errors with 39% and 36% of the annotators selecting these error classes, respectively.

Overall, ChatGPT generated moderate-quality translations for each text domain, with significant errors in terminology, syntax, morphology, and style. These findings suggest that the model remains underdeveloped for providing accurate Greek translations in professional contexts, particularly for more technical texts. Nevertheless, ChatGPT has shown attempts to adjust to the stylistic requirements of each text type, confirming its versatile architecture. Notably, the evaluation for punctuation, spelling and untranslated text indicates that the machine-generated outputs are appropriate in these aspects. These observations underscore the need for further refinement and training of ChatGPT to enhance its proficiency in Greek translations especially for specialized and professional scenarios. These findings highlight the critical areas for improvement and set the stage for future advancements in the MT performance of ChatGPT.

5.2 Human and Automatic Evaluation Analysis

After an inclusive analysis comparing the DEJ-based evaluation scores to automatic metric scores, we have gained a more comprehensive understanding of ChatGPT's MT capabilities in the Greek language. To accurately present human evaluation responses, we have calculated overall quality scores by equally weighing accuracy and fluency. The overall scores were subsequently converted to a 100-

point scale to be normalized on a common scale with the automatic metric values, in order to easily correlate the final quality results with those of previous studies.

$$\text{Overall Quality Score} = \frac{\text{Accuracy Score} + \text{Fluency Score}}{2}$$

Our findings reveal that overall fluency scores slightly surpass accuracy scores, underscoring ChatGPT's proficiency in generating increasingly fluent, human-like natural language. However, these results also highlight the model's limitations in providing accurate information, especially in domain-specific contexts.

A domain-wise analysis indicated that the journalistic article received the highest scores in accuracy and fluency as it outperformed the biomedical text, which in turn surpassed the advertising text that received the lowest scores. Interestingly, the e-Commerce domain, which typically requires less specialized knowledge and permits greater flexibility, was rated lower than anticipated. This aberration can be attributed to stylistic errors and deviations from the intended purpose of the translation, as identified in the error classification task. The translation output of the journalistic article demonstrated significantly improved fluency and accuracy compared to other text types, likely due to its simpler domain-specific challenges and language use, which minimized corresponding errors. Despite the substantial challenges associated with medical terminology and technical accuracy, the biomedical text scored moderately, reflecting the intermediate difficulty of this particular domain.

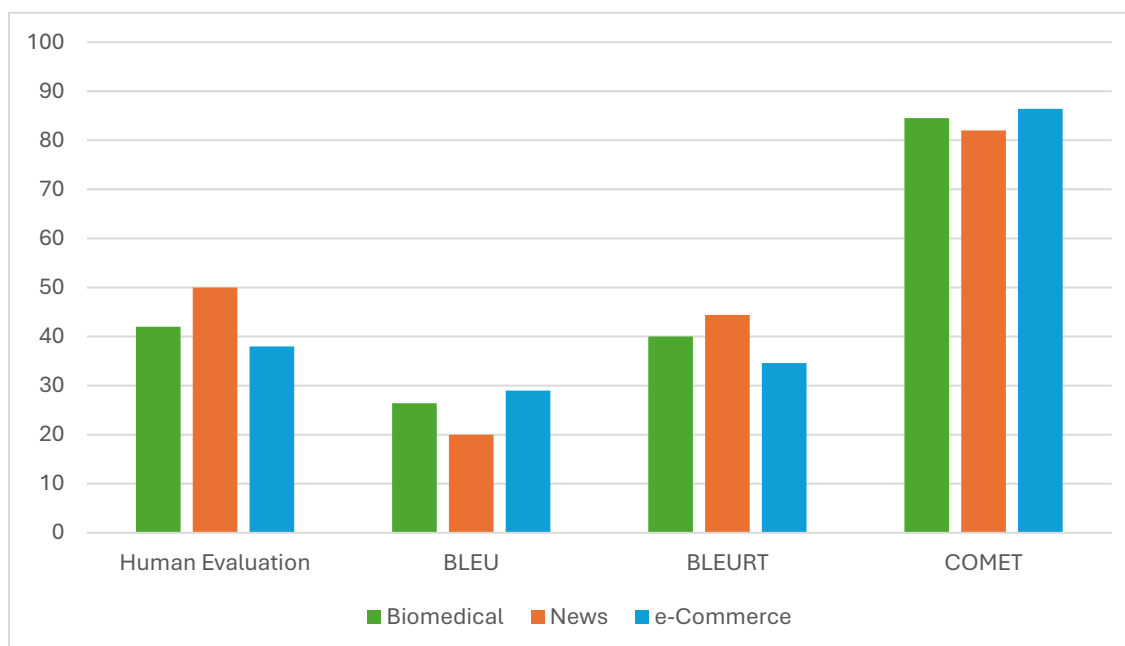


Figure 6: Human evaluation and automatic evaluation scores

Upon conducting an all-encompassing evaluation analysis, we have correlated human evaluation scores with the appointed automatic metrics (see Figure 6). This analysis underlines that the translation performance of ChatGPT for the English-Greek language pair is moderate compared to high-resource language-pairs. Prior to delving into details, it is imperative to note that the evaluation scores attributed to ChatGPT exhibit substantial consistency with those reported in large-scale evaluation studies for the same language pair (Robinson et al., 2023). This compatibility substantiates the robustness and reliability of our findings, reinforcing the validity of our comparative assessment.

In detail, BLEURT scores demonstrated a notable alignment with human judgments, assigning comparable values across machine-generated outputs per text type: e-Commerce < Biomedical < News. On the other hand, BLEU and COMET scores have displayed a divergent hierarchy of text-domain quality levels: News < Biomedical < e-Commerce. The relatively lower scores assigned by the BLEU metric indicate a significant deviation from the reference texts, implying that ChatGPT generated more flexible and creative translations. Conversely, neural-based metrics, BLEURT and COMET, exhibited a stronger

correlation with annotators' evaluations, reflecting a closer approximation to human judgment.

Specifically, COMET's higher scores are attributed to its more flexible and forgiving architecture and the metric's propensity to assign elevated scores is common and indicative of its ability to capture the contextual nuances of the MT outputs.

Another crucial component integrated into the human evaluation methodology is the assessment of error severity. Annotators were instructed to evaluate the criticality of the errors identified in the translation outputs per text domain, using a five-point severity scale. Upon meticulous analysis of the annotators' scores, our findings suggest that the majority of errors exhibit substantial criticality, highlighting ChatGPT's moderate performance in translation tasks (see Figure 7). A domain-specific analysis indicates a heightened level of error severity in the Biomedical and e-Commerce domains, contrasting with a lower incidence of critical errors in the News domain. These observations align with previous assessment scores, demonstrating that ChatGPT excessively generates critical errors across diverse text types in Greek MT.

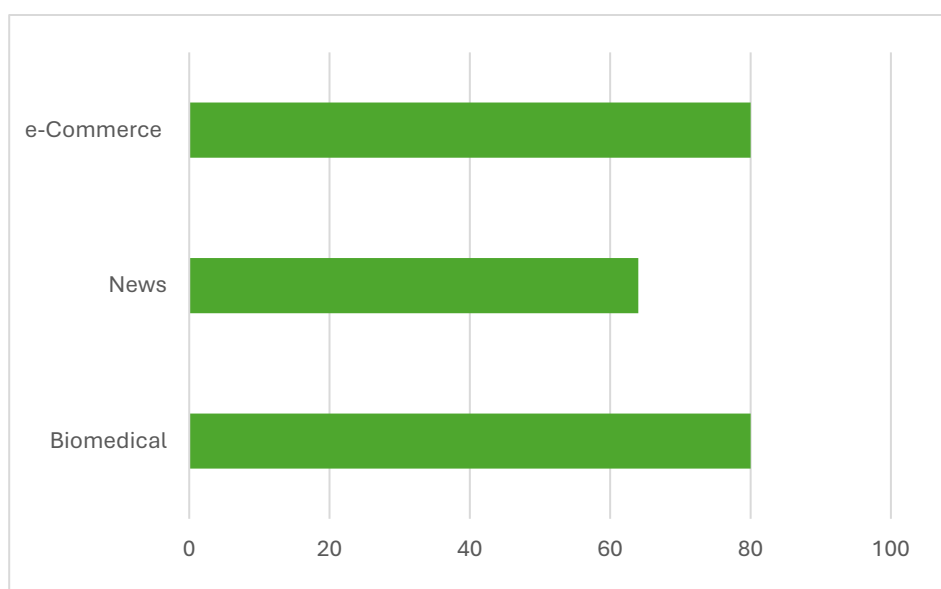


Figure 7: Error severity scores

6. Pivot Prompting

In our evaluation of ChatGPT's MT capabilities for the Greek language, we employed zero-shot techniques and baseline prompts. This approach aimed to trigger the model's translation ability by simply providing instruction and information about the target language and the source text. In this way, we gained comprehensive insights into ChatGPT's English-to-Greek translation capacity.

In this section, we briefly explore the potential of optimizing ChatGPT for Greek MT through a recently popular zero-shot prompt engineering strategy. Jiao et al. (2023) introduced an innovative technique termed pivot prompting, wherein ChatGPT is instructed to initially translate the source text into a high-resource intermediary language (pivot language) before translating it into the target language. The employment of a high-resource pivot language helps to effectively impart critical information to the target text, thereby mitigating the information loss that typically occurs in low-resource scenarios due to insufficient training data. Preliminary experimental results demonstrated that pivot prompting significantly enhances ChatGPT's translation performance, suggesting that it is a promising strategy for achieving high-quality multilingual MT.

Considering that Greek is an underrepresented language in ChatGPT, we opted to assess the pivot prompting technique through a small-scale experimental process. Specifically, we applied the prompt proposed by Jiao et al. (2023) to translate the three selected texts, using German as the pivot language due to its status as a high-resource language supported by ChatGPT. The intermediary translation into German is intended to preserve critical information that might otherwise be lost when translating directly from English to Greek, considering the relatively sparse training data available for Greek. The suggested prompt was adapted to our specific language data as follows: "Please provide the [German] translation first and then the [Greek] translation for these sentences one by one:".

Before analyzing our findings, it is essential to acknowledge the limitations inherent in our study. The scope of our research was constrained by a limited dataset and a narrow linguistic framework, which may affect the generalizability of our results. Consequently, the relative outcomes and their applicability to the Greek language may not accurately reflect the effectiveness of this technique in large-scale, realistic applications. Additionally, due to practical constraints, our evaluation of the machine-generated outputs following the implementation of pivot prompting relied solely on automatic evaluation metrics. This limitation may further affect the reliability and validity of our findings, as automatic metrics alone may not fully capture the nuanced aspects of translation quality that human annotators can provide.

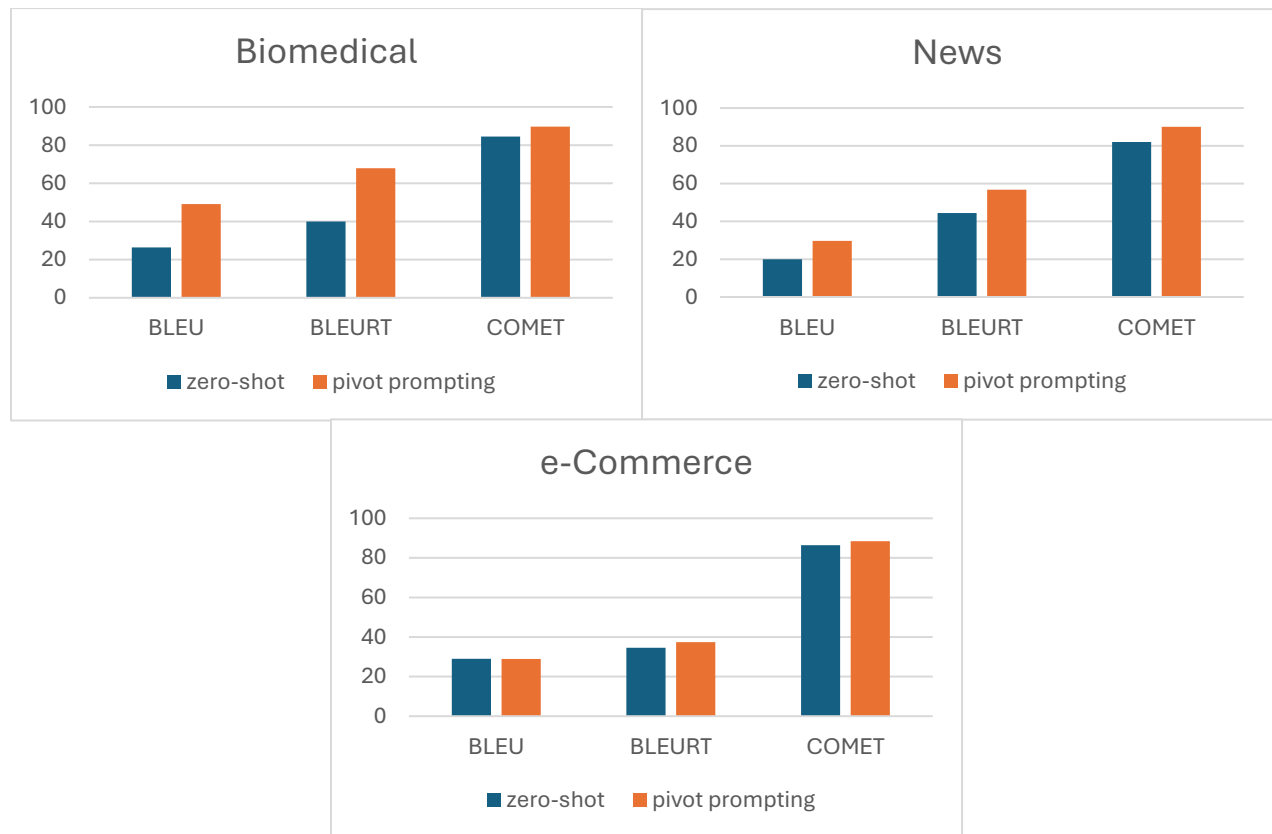


Figure 8: Pivot prompting quality scores per text domain using automatic metrics

Upon analyzing our score findings per text domain, we observed a significant and consistent improvement in ChatGPT's performance in all text types, highlighting the critical role of prompts in

eliciting more desirable responses. Notably, the most substantial increase in quality scores is identified in the Biomedical domain, with a 28% higher score attributed by the BLEURT metric, which aligns closely with human judgments as established in prior sections. The News domain also experienced a notable improvement in quality scores, with an average increase of 10% among automatic metrics. Conversely, the e-Commerce domain exhibited a marginal improvement, with only 1% general increase in quality scores (e-Commerce < News < Biomedical).

This variation in ChatGPT's performance underscores the impact of pivot prompting in enhancing translation quality for low-resource languages. By leveraging the training data of a high-resource intermediary language, ChatGPT can produce higher quality translations in specialized domains. Considering the significant improvement observed in the Biomedical domain we can assume that ChatGPT made substantial advancements in handling terminology errors through pivot prompting. Concurrently, the overall improvements may imply a reduction in morphological and syntactical errors as inferred from our error classification findings.

7. Conclusion

ChatGPT has already made significant contributions in the field of MT research and holds promising potential to evolve into a reliable and invaluable tool within the translation industry. However, the model's performance varies significantly across different linguistic and cultural contexts. Furthermore, its successful application in translation tasks is highly dependent on the complexity and specialization of the source texts. In this context, achieving high-quality machine translation is also influenced by various methodological factors, including prompting techniques and input manipulation. Overall, ChatGPT remains in a transitive and evolving stage, with its performance contingent on various factors to generate desirable translation outputs.

In the context of our language-specific and accessibility-promoting study, ChatGPT remains at an early stage of development, exhibiting generally moderate MT performance for Greek, a medium-resource language, across various domains. Our comprehensive evaluation, which involved both human annotators and automatic metrics, uncovered numerous critical errors in the English-Greek language pair. Notably, ChatGPT particularly underperformed in accurately rendering terminology and providing syntactically and morphologically correct Greek translations. This underperformance was particularly pronounced in specialized domains, where Greek translations frequently showed increased instances of hallucinations and unnatural language use. The evaluation scores further underscored this disparity; while the model displayed higher fluency scores, its accuracy scores were significantly lower, suggesting a substantial gap in training resources for the Greek language and emphasizing the urgent need for more comprehensive Greek-specific training corpora and advanced modeling techniques.

These findings highlight critical areas for improvement and set the stage for future advancements in ChatGPT's MT performance. Through this research, we highlight the imperative necessity for further refinement and training of the model, to enhance its proficiency in Greek translations, particularly for specialized professional scenarios. Additionally, the insights gained from this study not only illuminate the current limitations of ChatGPT but also pave the way for targeted efforts to optimize its capabilities, ultimately contributing to the broader goal of advancing MT technology for low-resource languages.

In our subsequent endeavor to optimize ChatGPT for Greek MT, we employed a recently introduced prompting strategy designed for low-resource languages, termed pivot prompting. Our findings demonstrated a significant improvement in ChatGPT's performance, particularly in specialized texts. This underscores the critical role of prompts in harnessing the full potential of ChatGPT for MT in

low-resource scenarios, highlighting the promising, albeit preliminary, evaluation of this prompting technique.

In light of these findings, it is imperative to assess ChatGPT's potential for large-scale professional application. The introduction of ChatGPT into the translation industry has raised concerns about the potential displacement of human expertise and workforce by AI technology, sparking significant debates about the chatbot's reliability and suitability for professional purposes. ChatGPT has transformed the dynamics between translation technology and translation professionals, prompting questions about its optimal application, aiming to balance technological advancements with the preservation of human expertise.

While the AI-chatbot has been incorporated into the translation industry in various capacities worldwide, its widespread adoption for the Greek language remains uncertain. Considering our findings, we conclude that ChatGPT is not yet suitable for large-scale autonomous incorporation by translation practitioners. Its current developmental stage renders it more appropriate for subsidiary and inspirational tasks, proving optimal for limited applications where it can assist professionals in time-consuming activities. This observation underscores ChatGPT's early stage of development and its unreliability in domains requiring high accuracy and specialized knowledge, highlighting the considerable distance it must travel before becoming a dependable tool for Greek language contexts. A significant concern is the model's inability to be accountable for professional decisions, take initiative in critical situations, and maintain transparency in professional settings. Nevertheless, our research reveals that the key to harnessing ChatGPT's MT capabilities lies in effective prompting techniques, which have shown promising improvements in performance.

With the advent of GPT-4 and its expanded multilingual capabilities, coupled with the rapid evolution of GPT models, ChatGPT is poised to secure a robust and enduring role within the translation

industry. This progression heralds a transformative era in human-machine collaboration, suggesting a future where technological advancements complement and enhance human expertise rather than displace it.

References

- Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Abiodun, O. I. (2023). A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information*, 14(8), 462. <https://doi.org/10.3390/info14080462>
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Curēus*. <https://doi.org/10.7759/cureus.35179>
- Antonopoulos, V. (2023, December 13). A Greek LLM after Fine-tuning Llama2 7b. Medium. <https://medium.com/11tensors/a-greek-llm-after-fine-tuning-llama2-7b-678a8eca1ab3>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4337484>
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q., V., Xu, Y., & Fung, P. (2023, February 8). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv.org*. <https://arxiv.org/abs/2302.04023>
- Bergmann, D. (2023, November 10). What is Reinforcement Learning from Human Feedback (RLHF)?. IBM. <https://www.ibm.com/topics/rlhf>
- Broadhead, G. (2023, August 26). A Brief Guide to LLM Numbers: Parameter Count vs. Training Size. *Medium*. <https://medium.com/@greg.broadhead/a-brief-guide-to-llm-numbers-parameter-count-vs-training-size-894a81c9258>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh,

- A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020, May 28). Language Models are Few-Shot Learners. arXiv.org. <https://arxiv.org/abs/2005.14165>
- Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007, June 1). (Meta-) Evaluation of Machine Translation. ACL Anthology. <https://aclanthology.org/W07-0718/>
- Chan, A. (2022). GPT-3 and InstructGPT: Technological Dystopianism, Utopianism, and “Contextual” Perspectives in AI Ethics and Industry. *AI And Ethics*, 3(1), 53–64.
<https://doi.org/10.1007/s43681-022-00148-6>
- ChatGPT. (n.d.). OpenAI. <https://openai.com/chatgpt/>
- Chatzikoumi, E. (2019, September 11). How to Evaluate Machine Translation: A Review of Automated and Human Metrics. Semantic Scholar. <https://www.semanticscholar.org/paper/How-to-evaluate-machine-translation%3A-A-review-of-Chatzikoumi/7c50274e3226cf00e780afe266dd07dace24761a>
- Chowdhury, A., & Rahman, S. (2023). A Brief Review of ChatGPT: Limitations, Challenges and Ethical-Social Implications. Chongqing University of Posts and Telecommunications.
<https://doi.org/10.5281/zenodo.7629888>
- Common Crawl - Open Repository of Web Crawl Data. (n.d.). <https://commoncrawl.org/>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228-239. <https://doi.org/10.1080/14703297.2023.2190148>
- Dale, R. (2020). GPT-3: What’s it Good for? *Natural Language Engineering*, 27(1), 113–118.
<https://doi.org/10.1017/s1351324920000601>
- DALL·E: Creating Images from Text. (2021, January 5). OpenAI. <https://openai.com/index/dall-e/>
- DALL·E 2. (n.d.). OpenAI. <https://openai.com/index/dall-e-2/>
- DALL·E 3. (n.d.). OpenAI. <https://openai.com/index/dall-e-3/>

- Dave, T., Athaluri, S. A., & Singh, S. (2023). ChatGPT in Medicine: An Overview of its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1169595>
- DeepL Translate: The World's Most Accurate Translator. (n.d.). <https://www.deepl.com/en/translator>
- Duarte, F. (2024, April 30). Number of ChatGPT Users (May 2024). *Exploding Topics*.
<https://explodingtopics.com/blog/chatgpt-users>
- EuroParl Parallel Corpus. (n.d.). <https://www.statmt.org/europarl/>
- Evaluate Machine Translations. (2023). MATEO. <https://mateo.ivdnt.org/Evaluate>
- Evdaimon, I., Abdine, H., Xypolopoulos, C., & Stamou, G. (2023). GreekBART: the First Pretrained Greek Sequence-to-Sequence Model. ResearchGate.
https://www.researchgate.net/publication/369760229_GreekBART_The_First_Pretrained_Greek_Sequence-to-Sequence_Model
- Flensted, T. (2024, April 24). How Many Languages does ChatGPT Support? The Complete ChatGPT Language List. <https://seo.ai/blog/how-many-languages-does-chatgpt-support>
- Gao, Y., Wang, R., & Hou, F. (2023, April 5). How to Design Translation Prompts for ChatGPT: An Empirical Study. arXiv.org. <https://arxiv.org/abs/2304.02182>
- Gemini. (2024, May 14). Google DeepMind. <https://deepmind.google/technologies/gemini/>
- Giarelis, N., Mastrokostas, C., & Karacapilidis, N. (2023, November 13). GreekT5: a Series of Greek Sequence-to-Sequence Models for News Summarization. arXiv.org.
<https://arxiv.org/abs/2311.07767>
- Gillis, S. (2023, August 15). Natural Language Understanding (NLU). TechTarget.
<https://www.techtarget.com/searchenterpriseai/definition/natural-language-understanding-NLU>

Giray, L., Jacob, J., & Gumalin, D. L. (2024). Strengths, Weaknesses, Opportunities, and Threats of Using ChatGPT in Scientific Research. *International Journal of Technology in Education*, 7(1), 40–58.

<https://doi.org/10.46328/ijte.618>

Google Forms. (n.d.). Google. <https://www.google.com/forms/about/>

Google Translate. (n.d.). Google. <https://translate.google.com/?sl=en&tl=el&op=translate>

GPT-4. (n.d.). OpenAI. <https://openai.com/index/gpt-4/>

Gu, W. (2023, March 27). Linguistically Informed ChatGPT Prompts to Enhance Japanese Chinese Machine Translation: A Case Study on Attributive Clauses. *arXiv.org*.

<https://arxiv.org/abs/2303.15587>

Guinness, H. (2024, May 15). What is GPT-4o? OpenAI's New Multimodal AI model. *zapier*.

https://zapier.com/blog/gpt-4o/?utm_source=google&utm_medium=cpc&utm_campaign=gaw-row-nua-evr-search_nb_desktop_blog_prospecting_developing1_developing2-ads&utm_term=&utm_content=9067713&utm_ads_campaign_id=19622168382&utm_ads_ads_et_id=157760468828&utm_ads_ad_id=659058345731&gad_source=1&gclid=Cj0KCQjw0_WyBhDMARIsAL1Vz8vb7x-8fu7xs8GZ_e_ZX39DCLLGuc419QfFkE57IHbBBvwVgi_RLHlaAkpCEALw_wcB

Guleria, A., Krishan, K., Sharma, V., & Kanchan, T. (2023). ChatGPT: Ethical Concerns and Challenges in Academics and Research. *Journal of Infection in Developing Countries*, 17(09), 1292–1299.

<https://doi.org/10.3855/jidc.18738>

Guzmán, F., Joty, S. R., Márquez, L., & Nakov, P. (2017, October 5). Machine Translation Evaluation with Neural Networks. *arXiv.org*. <https://arxiv.org/abs/1710.02095>

Hashemi-Pour, C., & Carew, J. M. (2023, August 16). Reinforcement Learning. *TechTarget*.

<https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning>

Hello GPT-4o. (2024, May 13). OpenAI. <https://openai.com/index/hello-gpt-4o/>

- Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023, February 18). How Good are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv.org. <https://arxiv.org/abs/2302.09210>
- Introducing ChatGPT. (2022, November 30). OpenAI. <https://openai.com/index/chatgpt/>
- Introducing Whisper. (2022, September 21). OpenAI. <https://openai.com/index/whisper/>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Jiang, Z., Lv, Q., & Zhang, Z. (2023, December 17). Distinguishing Translations by Human, NMT, and ChatGPT: A Linguistic and Statistical approach. arXiv.org. <https://arxiv.org/abs/2312.10750>
- Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023, January 20). Is ChatGPT a Good Translator? A Preliminary Study. *Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine*. arXiv.org. <https://arxiv.org/abs/2301.08745>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. A Position Paper. <https://doi.org/10.35542/osf.io/5er8f>
- Khan, S. A. (2023, June 15). Ethical Implications of ChatGPT: Misinformation, Manipulation and Biases. YourStory. <https://yourstory.com/2023/05/ethical-implications-of-chatgpt-language-processing-misinformation>
- Läubli, S., Sennrich, R., & Volk, M. (2018, August 21). Has Machine Translation Achieved Human Parity? a Case for Document-level Evaluation. arXiv.org. <https://arxiv.org/abs/1808.07048>

Laumann, F. (2023, December 18). Low-resource Language: What Does it Mean? - NeuralSpace.

Medium. <https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5>

Lecler, A., Duron, L., & Soyer, P. (2023). Revolutionizing Radiology with GPT-based Models: Current Applications, Future Possibilities and Limitations of ChatGPT. Diagnostic and Interventional Imaging, 104(6), 269–274. <https://doi.org/10.1016/j.diii.2023.02.003>

Libre Eau de parfum Women's Perfume — Perfume — YSL Beauty. (n.d.). YSL.

https://www.yslbeautyus.com/fragrance/womens-fragrances/libre/libre-eau-de-parfum/109YSL.html?fbclid=IwAR0B9BjPxAWDxSzRXj9_1UqfG_eznukYxsCsdP3BEg25X1MPXEILe0CfA8

Lisowski, E. (2024, February 27). How do I Use the OpenAI API and Find the API key? Step by Step Guide for 2024. *addepto*. <https://addepto.com/blog/what-is-an-openai-api-and-how-to-use-it/#:~:text=The%20OpenAI%20API%2C%20specifically%2C%20serves,into%20various%20applications%20and%20services.>

Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Sparano, R. M., Foresi, M., Innis, J., Han, L., & Nenadic, G. (2024, May 27). The Multi-Range Theory of Translation Quality Measurement: MQM Scoring Models and Statistical Quality Control. *arXiv.org*. <https://arxiv.org/abs/2405.16969>

Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S., & Möller, S. (2023). Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT? In Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.wmt-1.23>

Microsoft Copilot: Your Everyday AI Companion. (n.d.). Microsoft. <https://copilot.microsoft.com/>

Microsoft Translator. (n.d.). <https://translator.microsoft.com/>

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023, January 30). Adaptive Machine Translation with Large Language Models. arXiv.org. <https://arxiv.org/abs/2301.13294>

Mucci, T., & Stryker, C. (2024, April 18). Getting Ready for Artificial General Intelligence with Examples. *IBM Blog*. <https://www.ibm.com/blog/artificial-general-intelligence-examples/>

Naumova, E. N. (2023). A Mistake-find Exercise: a Teacher's Tool to Engage with Information Innovations, ChatGPT, and their Analogs. *Journal of Public Health Policy*, 44(2), 173–178. <https://doi.org/10.1057/s41271-023-00400-1>

Nazir, A., & Wang, Z. (2023). A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta-radiology*, 1(2), 100022. <https://doi.org/10.1016/j.metrad.2023.100022>

O'Connor, R. (2023, September 29). How DALL-E 2 actually works. *AssemblyAI*. <https://www.assemblyai.com/blog/how-dall-e-2-actually-works/#what-is-dall-e-3>

OpenAI. (n.d.). <https://openai.com/>

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2023). GPT-4 Technical Report. In arXiv.org. <https://arxiv.org/abs/2303.08774>

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askill, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, March 4). Training Language Models to Follow Instructions with Human Feedback. *arXiv.org*. <https://arxiv.org/abs/2203.02155>

Papantoniou, K., & Tzitzikas, Y. (2020). NLP for the Greek Language: A Brief Survey.

<https://www.semanticscholar.org/paper/NLP-for-the-Greek-Language%3A-A-Brief-Survey-Papantoniou-Tzitzikas/f919148839e06103fbaf07ff4e5530e8a62c1300>

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. <https://www.semanticscholar.org/paper/Bleu%3A-a-Method-for-Automatic-Evaluation-of-Machine-Papineni-Roukos/d7da009f457917aa381619facfa5ffae9329a6e9>

Patel, A. S., & Shah, K. J. (2018). Histopathological Analysis of Hysterectomy Specimens in Tertiary Care Center: Two Year Study. Pathology Update: Tropical Journal of Pathology and Microbiology, 4(1), 34–39. <https://doi.org/10.17511/jopm.2018.i01.06>

Pelucchi, M. (2023). Exploring ChatGPT's Accuracy and Confidence in High-Resource Languages [Bachelor's Thesis, University of Groningen]. https://fse.studenttheses.ub.rug.nl/30857/1/Thesis_MartinoPelucchi.pdf

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023, March 24). Towards Making the Most of ChatGPT for Machine Translation. arXiv.org. <https://arxiv.org/abs/2303.13780>

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In Translation Quality Assessment (pp. 129–158). https://doi.org/10.1007/978-3-319-91241-7_7

Proximal Policy Optimization. (2017, July 20). OpenAI. <https://openai.com/index/openai-baselines-ppo/>

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December 6). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv.org. <https://arxiv.org/abs/2212.04356>

- Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation. *Journal of Applied Artificial Intelligence*, 5(1), 69–93. <https://doi.org/10.48185/jaai.v5i1.1052>
- Ray, P. P. (2023). ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. *Internet of Things and Cyber-physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Rei, R., Guerreiro, N. M., Treviso, M., Coheur, L., Lavie, A., & Martins, A. F. T. (2023, May 19). The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics. arXiv.org. <https://arxiv.org/abs/2305.11806>
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <https://arxiv.org/abs/2009.09025>
- Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023, September 14). ChatGPT MT: Competitive for High- (but not Low-) Resource Languages. arXiv.org. <https://arxiv.org/abs/2309.07423>
- Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI models: A Preliminary Review. *Future Internet*, 15(6), 192. <https://doi.org/10.3390/fi15060192>
- Rule-based Machine Translation. (2023, October 30). Machine Translate. <https://machinetranslate.org/rule-based-machine-translation>
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>

- Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Vol. 5). <https://arxiv.org/abs/2004.04696>
- Sharou, K. A., & Specia, L. (2022). A Taxonomy and Study of Critical Errors in Machine translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. <https://aclanthology.org/2022.eamt-1.20/>
- Siu, S. C. (2023). ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. Social Science Research Network. <https://doi.org/10.2139/ssrn.4448091>
- Statistical Machine Translation. (2023, October 30). Machine Translate. <https://machinetranslate.org/statistical-machine-translation>
- Tabsharani, F. (2023, August 2). Machine Translation. Enterprise AI. <https://www.techtarget.com/searchenterpriseai/definition/machine-translation>
- Vanroy, B., Tezcan, A., & Macken, L. (2023). MATEO: MACHine Translation Evaluation Online. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation. https://www.researchgate.net/publication/371736176_MATEO_MACHine_Translation_Evaluation_Online
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is All You Need. arXiv.org. <https://arxiv.org/abs/1706.03762>
- Walker, P., & Mason, R. (2024, March 29). Tory Donor's Knighthood is Sign Sunak 'believes he's on way out', Labour says. The Guardian. <https://www.theguardian.com/politics/2024/mar/29/tory-donor-knighthood-sign-sunak-way-out-labour-mohamed-mansour>

- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023, April 5). Document-Level Machine Translation with Large Language Models. arXiv.org. <https://arxiv.org/abs/2304.02210>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022, January 28). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv.org. <https://arxiv.org/abs/2201.11903>
- Williams, M. (2009). Translation Quality Assessment. ResearchGate. https://www.researchgate.net/publication/50863650_Translation_Quality_Assessment
- Wu, Y., & Hu, G. (2023). Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. <https://doi.org/10.18653/v1/2023.wmt-1.15>
- Yamada, M. (2023). Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In arXiv.org. <https://arxiv.org/abs/2308.01391>
- Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., & Wang, L. (2022). An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 3081–3089. <https://doi.org/10.1609/aaai.v36i3.20215>
- Zhang, B., Haddow, B., & Birch, A. (2023, January 17). Prompting Large Language Model for Machine Translation: A Case Study. arXiv.org. <https://arxiv.org/abs/2301.07069>
- Zhao, Y., Zhang, M., Chen, X., Deng, Y., Geng, A., Liu, L., Liu, X., Li, W., Jiang, Y., Yang, H., Han, Y., Tao, S., Xie, N., Li, X., Ma, M., & Zhang, Z. (2023). Human Evaluation for Translation Quality of ChatGPT: A Preliminary Study. https://doi.org/10.26615/issn.2683-0078.2023_023
- Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023, February 18). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv.org. <https://arxiv.org/abs/2302.09419>

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023, April 10). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. arXiv.org.
<https://arxiv.org/abs/2304.04675>

Appendices

Appendix A: Zero-shot Experimental Layout

Text Type	ST (EN)	TL (EL)	Reference
Biomedical	The uterus is a vital reproductive and hormone responsive organ is subjected to many non-neoplastic and neoplastic conditions. Many treatment options are available including medical and conservative surgical but hysterectomy still the second most frequently performed major surgical procedure in females worldwide next to cesarean section [1]. It is the treatment of choice for many indications which include dysfunctional uterine bleeding, fibroids, gynecological cancers and obstetric disorders [2]. Hysterectomy is done by the vaginal or abdominal route, or with the help of laparoscope. (...) The Hysterectomy specimens received were immediately transferred into 10% formalin saline in the ratio of 1:10. After 24 hours fixation, the specimen was examined grossly for size, wall thickness and any mass present and necessary sections were taken from uterus that includes endometrium, myometrium and serosa from fundus, body and lower – uterine segment.	Η μήτρα είναι ένα ζωτικό αναπαραγωγικό και ευαίσθητο στις ορμόνες όργανο που υφίσταται πολλές μη νεοπλαστικές και νεοπλαστικές καταστάσεις. Υπάρχουν πολλές επιλογές θεραπείας διαθέσιμες, συμπεριλαμβανομένης της ιατρικής και της συντηρητικής χειρουργικής, αλλά η υστερεκτομή παραμένει το δεύτερο συχνότερα εκτελούμενο μείζον χειρουργικό επείγον πρόσθετο παρεμβατικό επείγον σε γυναίκες σε παγκόσμιο επίπεδο, μετά την καισαρική τομή. Είναι η θεραπεία της επιλογής για πολλές ενδείξεις που περιλαμβάνουν τη λειτουργική μήτρας αιμορραγία, τους μυώματα, τους γυναικολογικούς καρκίνους και τις τοκετικές διαταραχές. Η υστερεκτομή πραγματοποιείται με βαγινή ή κοιλιακή οδό ή με τη βοήθεια λαπαροσκοπίου. (...) Τα δείγματα υστερεκτομής που λήφθηκαν μεταφέρθηκαν αμέσως σε φορμαλδεΐδη 10% σε φυσιολογικό αλάτι σε αναλογία 1:10. Μετά από 24 ώρες σταθεροποίησης, το δείγμα εξετάστηκε μακροσκοπικά για μέγεθος, πάχος τοίχου και οποιαδήποτε μάζα	Η μήτρα είναι ένα ζωτικής σημασίας αναπαραγωγικό και ορμονικά ανταποκρινόμενο όργανο το οποίο υπόκειται σε πολλές μη νεοπλασματικές και νεοπλασματικές συνθήκες. Διατίθενται πολλές επιλογές θεραπείας συμπεριλαμβανομένων ιατρικών και συντηρητικών χειρουργικών θεραπειών, αλλά η υστερεκτομή παραμένει η δεύτερη συχνότερη βαρεία χειρουργική επέμβαση στις γυναίκες παγκοσμίως μετά την καισαρική τομή [1]. Είναι η θεραπεία που επιλέγεται σε περιπτώσεις διάφορων ενδείξεων όπως η δυσλειτουργική αιμορραγία της μήτρας, ινομυώματα, γυναικολογικοί καρκίνοι και μαιευτικές διαταραχές [2]. Η υστερεκτομή πραγματοποιείται μέσω της κοιλιακής ή κοιλιακής οδού, ή με τη βοήθεια λαπαροσκοπίου. (...) Τα δείγματα υστερεκτομής που ελήφθησαν μεταφέρθηκαν άμεσα σε διάλυμα φορμόλης 10% σε αναλογία 1:10. Έπειτα από 24 ώρες μονιμοποίησης, κάθε δείγμα εξετάστηκε μακροσκοπικά ως προς το μέγεθος, το πάχος

		παρούσα και πραγματοποιήθηκαν οι αναγκαίες τομές από τη μήτρα που περιλαμβάνουν το ενδομήτριο, το μυομήτριο και τη σερόζα από τον φόνο, το σώμα και το κάτω-ετερικό τμήμα της.	των τοιχωμάτων και κάθε υπάρχουσα μάζα, ενώ παρήχθησαν απαραίτητα τμήματα της μήτρας όπως το ενδομήτριο, το μυομήτριο και ορογόνο χιτώνας από τον πυθμένα, το σώμα και το κατώτερο τμήμα της μήτρας.
News	Mohamed Mansour, a senior treasurer of the Tory party for just over a year, was among surprise recipients of honours unexpectedly announced late on Thursday, who also included a series of Conservative MPs. While the citation for Mansour said the knighthood was for business, charity and political service, Anneliese Dodds, the Labour chair, said it appeared clearly tied to his donation to the Tories. "This seems to be either the action of an incredibly arrogant prime minister who just doesn't care what the public thinks, or someone who's effectively demob happy, and not thinking he's going to stay in that role as prime minister," she told Sky News. "Really, I think the public are going to be quite appalled by this news today".	Ο Μοχάμεντ Μανσούρ, ανώτερος ταμίας του κόμματος Tory για λίγο παραπάνω από έναν χρόνο, ήταν μεταξύ των απροσδόκητων αντιδίκων των τιμών που ανακοινώθηκαν απροσδόκητα την Πέμπτη καθυστερημένα, οι οποίες περιλάμβαναν επίσης μια σειρά από βουλευτές των Συντηρητικών. Ενώ η αναφορά για τον Μανσούρ έλεγε ότι ο ιππότης ήταν για επιχειρήσεις, φιλανθρωπία και πολιτική υπηρεσία, η Άννελιζ Ντοντς, η πρόεδρος του Εργατικού Κόμματος, είπε ότι φαινόταν σαφώς δεμένο με τη δωρεά του στους Τόρις. «Αυτό φαίνεται είτε να είναι η πράξη ενός απίστευτα αλαζονικού πρωθυπουργού που απλώς δεν νοιάζεται για αυτό που νομίζει ο κόσμος, είτε κάποιος που είναι ουσιαστικά χαρούμενος για την αποχώρηση και δεν νομίζει ότι θα παραμείνει σε αυτόν τον ρόλο ως πρωθυπουργός», δήλωσε στο Sky News. «Πραγματικά, πιστεύω ότι ο κόσμος θα είναι αρκετά απογοητευμένος από αυτά τα νέα σήμερα».	Ο Μοχάμεντ Μανσούρ, ανώτερος ταμίας του κόμματος των Συντηρητικών για λίγο περισσότερο από έναν χρόνο, ήταν ανάμεσα στους αποδέκτες «έκπληξη» τιμητικών διακρίσεων που ανακοινώθηκαν αιφνιδίως αργά την Πέμπτη, μεταξύ των οποίων ήταν ακόμη αρκετοί Συντηρητικοί βουλευτές. Ενώ η παραπομπή για τον Μανσούρ ανέφερε ότι ο τίτλος του ιππότη του απονεμόταν για επιχειρηματικές, φιλανθρωπικές και πολιτικές υπηρεσίες, η Ανελίζ Ντόντς, η πρόεδρος του Εργατικού κόμματος, δήλωσε ότι φαινόταν ξεκάθαρα η σύνδεσή του με τη δωρεά του στους Συντηρητικούς. «Φαίνεται ότι πρόκειται για μια κίνηση είτε ενός εξαιρετικά αλαζονικού πρωθυπουργού που απλά δεν ενδιαφέρεται για τη γνώμη του κοινού είτε ενός πραγματικά ευχαριστημένου "απόστρατου" που δεν πιστεύει ότι θα παραμείνει στον ρόλο του πρωθυπουργού» είπε στο Sky News. «Αλήθεια, πιστεύω ότι το κοινό θα σοκαριστεί ιδιαίτερα με σημερινά νέα».
e-Commerce	Libre Eau de Parfum is a bold, floral perfume for women. A statement feminine fragrance for those who live by their own rules, is also a statement of value and sustainability,	Το Libre Eau de Parfum είναι ένα τολμηρό, λουλουδάτο άρωμα για γυναίκες. Ένα αρωματικό δήλωσης θηλυκό άρωμα για εκείνες που ζουν με τους δικούς τους	Το Libre Eau de Parfum είναι ένα γυναικείο τολμηρό, λουλουδάτο άρωμα. Μία δήλωση θηλυκής ευωδίας για εκείνες που ζουν με τους δικούς τους κανόνες, αλλά και μια

refillable at home with a 100ml refill bottle. Lavender essence from France combines with the sensuality of Moroccan orange blossom and daring notes of musk accord to create a unique long-lasting perfume. (...)

A celebration of freedom, this daring, yet distinctly floral and feminine perfume is for those who do what they want and dare to be exactly who they are. Breaking boundaries and paving their own way, this women's perfume is inspired by the bold and the free. An oversize golden cassandra logotype freely bent & nailed into the glass like a jewel. Sexy golden chains and an asymmetric black lacquered cap tops it off. Your forever Libre bottle, now sustainably refillable at home.

κανόνες, είναι επίσης μια δήλωση αξίας και βιωσιμότητας, με δυνατότητα αναπλήρωσης στο σπίτι με ένα μπουκάλι αναπλήρωσης 100ml. Η εσάνς λεβάντας από τη Γαλλία συνδυάζεται με την αισθησιακότητα του μαροκινού πορτοκαλιού και τις τολμηρές νότες του αρώματος μόσχου για να δημιουργήσει ένα μοναδικό και διαρκές άρωμα. (...)

Ένας εορτασμός της ελευθερίας, αυτό το τολμηρό, αλλά διακριτικά λουλουδάτο και θηλυκό άρωμα είναι για εκείνες που κάνουν ό,τι θέλουν και τολμούν να είναι ακριβώς ό,τι είναι. Σπάζοντας τα όρια και ανοίγοντας τον δικό τους δρόμο, αυτό το άρωμα για γυναίκες είναι εμπνευσμένο από το τολμηρό και το ελεύθερο. Ένα μεγάλο χρυσό λογότυπο cassandra που καμπυλώνεται ελεύθερα και καρφώνεται στο γυαλί σαν κοσμήματα. Σέξι χρυσές αλυσίδες και ένα ασύμμετρο μαύρο καπάκι το ολοκληρώνουν. Το μπουκάλι Libre, που είναι πλέον βιώσιμα αναπληρώσιμο στο σπίτι, για πάντα.

δήλωση για την αξία και τη βιωσιμότητα με την συσκευασία των 100ml που μπορεί ξαναγεμίσει στο σπίτι. Το αιθέριο έλαιο λεβάντας από τη Γαλλία συνδυάζεται με το αισθησιακό Μαροκινό άνθος πορτοκαλιάς, ενώ οι τολμηρές νότες μόσχου λαμβάνουν μέρος στη δημιουργία ενός μοναδικού ανθεκτικού αρώματος. (...) Μια γιορτή ελευθερίας. Αυτό το τολμηρό αλλά ιδιαίτερα λουλουδάτο και θηλυκό άρωμα είναι για εκείνες που κάνουν ότι επιθυμούν και τολμούν να είναι ο εαυτός τους. Ξεπερνώντας τα όρια και ανοίγοντας νέους δρόμους, αυτό το άρωμα είναι εμπνευσμένο από τις τολμηρές και τις ελεύθερες. Ένα πολύ μεγάλο, χρυσό λογότυπο του cassandra ελεύθερα λυγισμένο αγκαλιάζει το γυαλί σαν κόσμημα. Οι αισθησιακές χρυσές αλυσίδες και το ασύμμετρο μαύρο γυαλιστερό καπάκι του το απογειώνουν. Το μπουκάλι Libre που θα έχεις για πάντα, μπορείς τώρα να το ξαναγεμίσεις στο σπίτι.

Appendix B: Error Classification Scores

	Biomedical	News	e-Commerce
Terminology errors (e.g., inaccurate medical terms)	70%	60%	62%
Morphological errors (e.g., casing, tense)	40%	27%	36%
Syntactic errors (e.g., word order)	66%	45%	39%
Spelling errors	7%	6%	5%
Punctuation errors	14%	4%	2%
Errors in style (e.g., higher or lower register)	26%	31%	30%
Untranslated text	4%	0%	5%
Unintelligible text	34%	29%	31%
I cannot identify any errors	5%	14%	6%

Appendix C: Human Evaluation and Automatic Evaluation Scores

Zero-shot	Human Evaluation	BLEU	BLEURT	COMET
Biomedical	42	26	40	84
News	50	20	44	82
e-Commerce	38	29	35	86

Appendix D: Pivot Prompting Evaluation Scores

Pivot Prompting	BLEU	BLEURT	COMET
Biomedical	49	68	90
News	30	57	90
e-Commerce	29	37	88