Comparative Evaluation of ChatGPT and DeepSeek Across Key NLP Tasks: Strengths, Weaknesses, and Domain-Specific Performance

Wael Etaiwi, Bushra Alhijawi

Princess Sumaya University for Technology, Amman, Jordan

Abstract

The increasing use of large language models (LLMs) in natural language processing (NLP) tasks has sparked significant interest in evaluating their effectiveness across diverse applications. While models like ChatGPT and DeepSeek have shown strong results in many NLP domains, a comprehensive evaluation is needed to understand their strengths, weaknesses, and domain-specific abilities. This is critical as these models are applied to various tasks, from sentiment analysis to more nuanced tasks like textual entailment and translation. This study aims to evaluate ChatGPT and DeepSeek across five key NLP tasks: sentiment analysis, topic classification, text summarization, machine translation, and textual entailment. A structured experimental protocol is used to ensure fairness and minimize variability. Both models are tested with identical, neutral prompts and evaluated on two benchmark datasets per task, covering domains like news, reviews, and formal/informal texts. The results show that DeepSeek excels in classification stability and logical reasoning, while ChatGPT performs better in tasks requiring nuanced understanding and flexibility. These findings provide valuable insights for selecting the appropriate LLM based on task requirements.

Keywords: ChatGPT, DeepSeek, Large language models, machine translation, NLP, sentiment analysis, text summarization, textual entailment, topic classification

Email addresses: w.etaiwi@psut.edu.jo (Wael Etaiwi), b.alhijawi@psut.edu.jo (Bushra Alhijawi)

1. Introduction

Artificial Intelligence (AI) technology has significantly transformed various domains, including healthcare, finance, education, and scientific research, by enabling machines to perform complex cognitive tasks [44]. AI-driven solutions have progressed from rule-based and statistical models to deep learning approaches, demonstrating remarkable computer vision, speech processing, and natural language processing (NLP) capabilities [8].

A breakthrough in AI research has been the development of generative AI (Gen AI) [15]. Gen AI is a class of machine learning models designed to create new content rather than solely analyze or classify existing data [15]. Unlike traditional AI models that rely on explicit rules and predefined patterns, Gen AI models utilize deep neural networks to generate text, images, audio, and code with contextually appropriate outputs [15]. These models have applications in machine translation, content generation, code synthesis, and automated reasoning. Gen AI has demonstrated remarkable capabilities in storytelling [45], dialogue generation [30], content summarization [38], and question answering [27]. Large Language Models (LLMs) enable these capabilities by earning from vast text corpora and then being fine-tuned for specific tasks. LLMs have become the foundation of modern NLP applications due to their ability to process and generate human-like text with contextual understanding [26]. These models leverage self-attention mechanisms to capture long-range dependencies in text, enabling them to perform a wide range of linguistic tasks [19]. Recent LLM advancements have resulted in highly sophisticated models, such as ChatGPT [1] and DeepSeek [29], demonstrating state-of-the-art performance across multiple NLP benchmarks. ChatGPT, developed by OpenAI, is a transformer-based model designed for various language generation tasks, including dialogue systems, text summarization, code generation, and question answering [1]. It has been fine-tuned with reinforcement learning from human feedback to enhance its conversational abilities, coherence, and contextual understanding. ChatGPT has gained widespread adoption due to its ability to generate human-like responses, making it a leading choice for chatbot applications, virtual assistants, and creative writing tasks [5]. In contrast, DeepSeek, developed by DeepSeek AI, is an advanced LLM designed to focus on multilingual processing, domain-specific applications, and knowledge-intensive tasks [28]. It leverages a highly optimized training pipeline and extensive datasets to improve factual accuracy, logical reasoning, and robustness across diverse NLP benchmarks. While both models are built on the transformer architecture, their training methodologies, optimization strategies, and application focus differ, necessitating a detailed comparative analysis to evaluate their strengths, weaknesses, and suitability for various NLP tasks.

Many studies evaluated the performance of ChatGPT and DeepSeek in different applications. However, to the best of our knowledge, few studies have focused on evaluating and comparing ChatGPT and DeepSeek on NLP tasks using benchmark datasets. Fu et al. [17] investigated whether ChatGPT can effectively evaluate the quality of planning documents, specifically in the context of climate change plans. By comparing ChatGPT's evaluations with those of human coders, the research reveals that ChatGPT's results align reasonably well with traditional methods, although it struggles with planning-specific jargon. Thelwall [42] evaluated ChatGPT 4.0's ability to assess journal article quality using UK Research Excellence Framework 2021 guidelines, applied to 51 of the author's articles. Also, the author compared ChatGPT results with self-evaluations. Caramancion [7] evaluated the capability of ChatGPT 3.5 and 4.0, Google's Bard/LaMDA, and Microsoft's Bing AI in verifying the truthfulness of news items using black box testing. Johnson et al. [21] assessed the accuracy and completeness of ChatGPT in responding to 284 medical questions created by 33 physicians across 17 specialties. They analyzed its performance using Likert scale ratings and statistical comparisons, finding generally high accuracy with limitations in complex cases, highlighting the need for further model refinement. Antaki et al. [3] assessed ChatGPT's accuracy in answering ophthalmology exam questions, showing improved performance with the Plus version and greater consistency across topics. Dunder et al. [13] evaluated the capability of ChatGPT in solving programming tasks using Kattis, revealing proficiency in simple problems but challenges with complex ones. ChatGPT successfully solved 19 out of 127 problems. Elyoseph et al. [14] evaluated ChatGPT's capacity to identify and describe emotions using an objective emotional awareness scale, demonstrating superior performance with progressive improvement over time.

The results indicate potential applications in cognitive training and psychiatric assessment. Khlaif et al. [22] assessed ChatGPT's role in scientific research, showcasing its ability to produce high-quality content but identifying challenges in research methodology and literature review. Deandres et al. [11] investigated ChatGPT's potential in face biometrics, including tasks like face verification and soft-biometrics estimation, revealing its ability to enhance the explainability of automatic decisions. The results demonstrate the effectiveness of LLMs such as ChatGPT in improving transparency and robustness in human-related biometric systems. Table 1 provides an overview of the experimental evaluation of ChatGPT across several tasks.

In that context, this study aims to evaluate the performance of ChatGPT and DeepSeek across five core NLP applications: sentiment analysis, topic classification, text summarization, machine translation, and textual entailment in the English language. For each application, two benchmark datasets are selected for comprehensive evaluation. The performance of both models is then compared using an offline evaluation methodology. To summarize the main contribution of this research paper:

- Evaluating the performance of ChatGPT on sentiment analysis, topic classification, text summarization, machine translation, and textual entailment tasks in the English language.
- Assessing the performance of DeepSeek on sentiment analysis, topic classification, text summarization, machine translation, and textual entailment tasks in the English language.
- Comparing the performance of ChatGPT and DeepSeek based on offline evaluation methodology.

The remainder of this paper is organized as follows. The next section provides an overview of ChatGPT and DeepSeek. Section 3 outlines the methodology employed in this study to perform an offline evaluation of both models. Section 4 presents, compares, and discusses the results obtained from the two LLMs. Finally, Section 5 concludes the paper and suggests potential directions for future research.

Table 1: Summary of the Experimental Evaluation of ChatGPT on Various Tasks

Article	Year	Objective	Description
[17]	2024	ChatGPT vs Human	- Explored if ChatGPT can evaluate planning documents, focusing on climate change plans. - Compared ChatGPT's evaluation results with those from human coders.
[42]	2024	Evaluation	- Assessed whether ChatGPT 4.0 can automate research evaluations of journal articles, focusing on its accuracy. - Tested ChatGPT 4.0's ability to evaluate journal articles using the UK Research Excellence Framework 2021 guidelines. - Compared ChatGPT-4's evaluations with the author's self-assessments of 51 articles.
[13]	2024	Evaluation	- Evaluated ChatGPT's ability to generate code solutions for programming tasks in introductory computer science courses. - Tested ChatGPT on 127 randomly selected programming problems from Kattis, an automated grading tool.
[11]	2024	ChatGPT vs Other	- Explored ChatGPT 4.0's capability in face biometrics tasks, including face verification, soft-biometrics estimation, and result explainability. - Experiments using public benchmarks were conducted to assess ChatGPT's performance in face biometrics and compare it with state-of-the-art methods.
[7]	2023	ChatGPT vs Other	- Assessed the proficiency of popular LLMs in verifying the truthfulness of news items using black box testing. - Evaluated ChatGPT 3.5 & 4.0, Google's Bard/LaMDA, and Microsoft's Bing AI on 100 fact-checked news items.
[21]	2023	Evaluation	- Evaluated the accuracy and completeness of ChatGPT's responses to medical queries across various specialties. - Assessed ChatGPT's performance on 284 medical questions created by 33 physicians from 17 specialties.
[3]	2023	Evaluation	- Evaluated ChatGPT's accuracy in answering ophthalmology-related multiple-choice questions. - Tested ChatGPT on two question banks (BCSC and OphthoQuestions) used for the OKAP exam.
[14]	2023	ChatGPT vs Human	- Investigated ChatGPT's ability to identify and describe emotions using the levels of emotional awareness scale. - Analyzed ChatGPT's responses to 20 emotional scenarios and compared its performance to humans.
[22]	2023	Evaluation	- Examined ChatGPT's potential in research by assessing the quality of AI-generated articles, data analysis, and literature reviews. - Generated four articles and 50 abstracts using ChatGPT, evaluated by 23 reviewers, and analyzed using ANOVA and thematic analysis.

2. ChatGPT and DeepSeek Overview

Recent LLM advancements have led to the development of sophisticated systems such as OpenAI's ChatGPT and DeepSeek's models, significantly enhancing NLP capabilities. These models exhibit impressive performance across various tasks, including question-answering, text generation, and summarization.

ChatGPT, developed by OpenAI[1], is based on the GPT-4 architecture, a dense transformer model trained on extensive internet text. It has been fine-tuned using Reinforcement Learning with Human Feedback, improving alignment with user intent and safety constraints [34]. This fine-tuning process enables ChatGPT to excel in general-purpose reasoning, complex problem-solving, and creative tasks such as code generation [9] and academic writing [35]. OpenAI has introduced multimodal capabilities in ChatGPT, enabling it to process text and images and enhancing its versatility in generating lifelike images and coherent text. Additionally, OpenAI has launched subscription services like DALL-E and ChatGPT Plus, offering users access to advanced features and models, including GPT-4.5, which provides up-to-date information retrieval and supports file and image uploads [18].

DeepSeek[2] has adopted a Sparse Mixture of Experts architecture, activating only a subset of model parameters during inference to optimize computational efficiency [31]. The latest model, DeepSeek-V3, released in March 2025, boasts 671 billion parameters with 37 billion activated per token, demonstrating significant improvements in reasoning and coding capabilities compared to its predecessors. DeepSeek's models have shown efficiency in running on consumer-grade hardware, with DeepSeek-V3 achieving 20 tokens per second on a Mac Studio, challenging traditional notions about AI model deployment requirements [20]. Furthermore, DeepSeek has narrowed the AI development gap with leading U.S. companies, with experts suggesting a convergence within three months in certain areas [33].

[1] https://openai.com/

[2] https://www.deepseek.com/

3. Methodology

This section outlines the methodology followed in evaluating the performance of ChatGPT and DeepSeek across five NLP tasks: sentiment analysis, topic classification, text summarization, machine translation, and textual entailment. The evaluation is conducted using an offline evaluation methodology, ensuring a standardized comparison. The methodology includes details on experimental setup (Section 3.1), datasets (Section 3.2), response collection (Section 3.3), and evaluation measures (Section 3.4).

3.1. Experimental Setup and Protocol

A well-defined experimental protocol is established to ensure a systematic and unbiased evaluation of ChatGPT and DeepSeek. This protocol governs the querying process, response collection, and evaluation framework, minimizing sources of variability and ensuring a fair comparison. The key components of the experimental design are detailed as follows:

- Prompt standardization: A set of predefined, structured prompts is designed for each NLP task to ensure consistency in model inputs. The prompts are crafted to be neutral, unambiguous, and representative of real-world use cases, minimizing biases that may affect model responses. The following prompt standardization principles are applied to eliminate inconsistencies in input structure:

    - Uniformity across models: Identical prompts are used for ChatGPT and DeepSeek across all tasks to ensure a fair comparison. These prompts are designed to be clear, concise, and unambiguous, adhering to best practices in LLM prompt engineering. Additionally, no extra context or system instructions are provided beyond what is essential for task completion.
    - Example-based prompting (Few-Shot Learning): For textual entailment and machine translation tasks, few-shot examples are tested to assess the impact of context on response accuracy. Example-based prompting is applied consistently across models to ensure fairness and prevent bias.

- Dataset splitting and consistency: To ensure a diverse and comprehensive evaluation, two benchmark datasets are selected for each NLP task, with both models evaluated on identical dataset splits to ensure fairness. The LLMs are assessed on balanced class distributions to prevent bias. Additionally, text examples from various domains, including news, reviews, and formal and informal texts, are incorporated to evaluate model generalization.

The considered structured protocol ensures robust and reliable performance evaluation, minimizing potential biases or inconsistencies.

3.2. NLP Tasks and Benchmark Datasets

Two benchmark datasets are selected to assess the performance of ChatGPT and DeepSeek across five core NLP tasks: sentiment analysis, topic classification, text summarization, machine translation, and textual entailment in the English language. These datasets, which are chosen to provide a comprehensive evaluation of the models' capabilities, are summarized in Table 2.

Table 2: Summary of Benchmark Datasets Used in the Experiments

Task	Dataset Name	Domain	Count of Records	Count of Testing Samples
Sentiment analysis	Multilingual Sentiment Datasets [4]	Scientific Documents	3036	150
	IMDB [32]	Movie Reviews	25000	100
Topic classification	AG News Classification Dataset [12]	News Articles	1 million	80
	Web of Science Dataset [24, 25]	Scientific Articles	64688	100
Text Summarization	CNN / Daily Mail Summarization Dataset [39]	News Articles	300000	50
	Gigaword [36]	News Articles	4 millions	50
Machine Translation	AraBench [37]	Diverse textual genres	173000	50
	ArzEn-MultiGenre [2]	Novels, subtitles, and songs	25557	50
Textual entailment	Scitail [23]	-	27026	50
	FraCaS [10]	-	350	75

3.2.1. Sentiment Analysis Datasets

Sentiment analysis involves analyzing textual data, such as comments, reviews, and opinions, to determine the sentiment expressed toward a specific subject, such as a product, event, or individual [6]. The performance of ChatGPT and DeepSeek in sentiment classification is evaluated using the IMDB and Multilingual Sentiment datasets. The IMDB dataset [3] is designed for binary sentiment classification. This dataset consists of 25,000 movie reviews for training and an additional 25,000 reviews for testing [32]. The Multilingual Sentiment dataset [4] combines several datasets to capture sentiment across multiple languages, including various Asian languages. The English portion of this dataset, sourced from the SemEval dataset [4], includes 1,840 training, 871 testing, and 325 validation records. Each record is classified as positive, neutral, or negative. Random samples of 50 instances per class are selected as testing samples for the experiments to ensure a consistent and fair evaluation.

3.2.2. Topic Classification Datasets

Topic classification is a text-mining technique that categorizes documents into predefined categories or labels based on their content [41]. In this study, ChatGPT and DeepSeek's performance in topic classification is evaluated using two benchmark datasets: the AG News Classification dataset [12] and the Web of Science dataset [24, 25]. The AG News dataset [5] consists of over one million news articles collected from more than 2,000 sources over a year, categorized into four main classes: Business, Sci/Tech, Sports, and World. For the AG News dataset, 20 testing samples are randomly selected per class. The Web of Science topic classification dataset [6] serves as a standardized benchmark for text classification, focused explicitly on topic categorization in scientific literature. This dataset includes research article abstracts assigned to seven predefined subject areas: Computer Science (CS), Electrical and Computer Engineering (ECE), Psychology, Mechanical and Aerospace Engineering (MAE), Civil Engineering, Medical Science, and Biochemistry. It is available in three versions: WOS-11967, WOS-46985, and WOS-5736, containing 11,967, 46,985, and 5,736 abstracts, respectively. For the Web of Science dataset, 100 testing samples are randomly selected, with

[3] https://github.com/Ankit152/IMDB-sentiment-analysis

[4] https://github.com/tyqiangz/multilingual-sentiment-datasets

[5] $http : //groups.di.unipi.it/gulli/AG_{c}orpus_{o}f_{n}ews_{a}rticles.html$

[6] https://data.mendeley.com/datasets/9rw3vkcfy4/6

varying instances per class.

3.2.3. Text Summarization Datasets

Text summarization is the automated process of generating a concise and coherent summary from a given text [40]. It can be classified into two main types: extractive and abstractive summarization. Extractive summarization involves selecting key sentences or paragraphs directly from the source text using scoring algorithms to identify the most relevant content. In contrast, abstractive summarization generates new sentences based on a semantic understanding of the document, producing summaries that may not include the original text verbatim but retain its essential meaning.

In this study, two benchmark datasets are used to evaluate the performance of ChatGPT and DeepSeek in text summarization: the CNN/Daily Mail Summarization Dataset [39] and the Gigaword Dataset [36]. The CNN/DailyMail dataset [7] is an English-language corpus containing over 300,000 news articles from CNN and the Daily Mail. It is widely used for evaluating text summarization models. The Gigaword text summarization dataset [8] is a large-scale dataset derived from the English Gigaword corpus, consisting of millions of news articles from respected sources like The New York Times, Associated Press, and Xinhua News Agency. This dataset is extensively used for both text summarization and headline generation tasks. For the evaluation, 100 random instances from each dataset are selected as testing samples.

3.2.4. Machine Translation Dataset

Machine Translation is the automatic conversion of text from one language to another using computational models and techniques [43]. It is widely used in applications like online translation services (e.g., Google Translate and DeepL) and real-time speech translation.

To evaluate the performance of ChatGPT and DeepSeek in machine translation, this study employs two benchmark datasets: AraBench [37] and the ArzEn-MultiGenre [2]

[7] https://cs.nyu.edu/ kcho/DMQA/

[8] https://catalog.ldc.upenn.edu/LDC2012T21

datasets. The AraBench dataset [9] is a dialectal Arabic-to-English machine translation corpus. It contains approximately 173,000 sentences categorized into 4 coarse-grained, 15 fine-grained, and 25 city-level classifications. The dataset includes a variety of textual genres such as media, chat, religion, and travel, each with varying levels of dialectal complexity. The ArzEn-MultiGenre dataset [10] is a manually translated parallel corpus consisting of 25,557 sentence pairs in Egyptian Arabic and English. It covers three genres: novels, subtitles, and songs. For the evaluation process, 100 random samples are selected as testing data from each dataset to ensure a consistent and reliable performance comparison between the two models.

3.2.5. Textual Entailment Dataset

Textual entailment refers to the semantic relationship between two text fragments, where one fragment logically follows or is implied by the other [16]. This relationship indicates that the meaning of one text can be inferred from the other, establishing a form of semantic implication. Automating textual entailment recognition is crucial for enhancing various NLP tasks, including information retrieval, extraction, question answering, text summarization, and machine translation.

To evaluate the performance of ChatGPT and DeepSeek in textual entailment recognition, this study utilizes SciTail [23] and FraCaS dataset [10]. The SciTail dataset [11] is derived from multiple-choice science exams and web sentences. It contains 27,026 examples, with 10,101 labeled as ”entails” and 16,925 labeled as ”neutral.” Each example consists of a question, and the correct answer choice is transformed into a hypothesis, with relevant text retrieved from a vast corpus of web sentences as the premise. The $FraCaS$ dataset [12] is an inference test suite created to evaluate the inferential capabilities of various NLP systems. It consists of approximately 350 labeled examples, with annotations to assess the recognition of entailment, contradiction, and neutrality. For evaluation, 50 samples per class are randomly selected as testing data, ensuring a fair

[9] https://alt.qcri.org/resources1/mt/arabench/

[10] https://data.mendeley.com/datasets/6k97jty9xg/5

[11] https://leaderboard.allenai.org/scitail/submissions/get-started

[12] https://nlp.stanford.edu/ wcmac/downloads/fracas.xml

and consistent performance comparison between the models.

3.3. LLM Querying and Response Collection

To evaluate the performance of ChatGPT and DeepSeek, we interact with both models directly through their user interfaces. This approach ensures that the study reflects real-world usage scenarios where users manually input queries and receive responses without API-based automation. Querying follows a structured, consistent methodology to ensure fairness and comparability across tasks. The response collection procedure includes the following steps:

- Consistent input formatting: The same prompts are used for both models to ensure a fair comparison. They are carefully designed to be clear, unambiguous, and aligned with best practices in prompt engineering. No additional instructions or system messages are provided beyond those necessary for the task. Table 3 provides examples of the prompts employed for each NLP task.

- Manual query execution: Each query is entered manually into ChatGPT and DeepSeek to simulate real-world user interactions. The generated responses are collected and stored for further analysis.

- Data storage and documentation: All responses are systematically documented in a structured format to facilitate the offline evaluation process. Figure 1 shows a sample of responses stored for the topic classification task.

3.4. Evaluation Measures

A set of established evaluation measures is employed to comprehensively evaluate the performance of ChatGPT and DeepSeek across five distinct NLP tasks. The selected metrics are tailored to the specific nature of each task to ensure a robust and meaningful comparison. For classification tasks, including sentiment analysis, topic classification, and textual entailment, precision, recall, F1-score, and accuracy are used to assess the models' performance. These metrics provide a balanced evaluation of how well the models identify correct instances (precision), relevant instances (recall), and their overall performance (F1-score and accuracy). For generative tasks, such as text

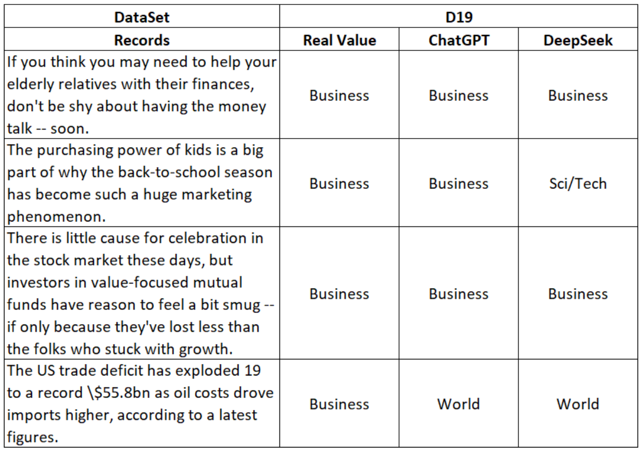
Table 3: Sample Queries for NLP Tasks

Task	Sample Query
Sentiment Analysis	Determine the sentiment of the following review, it should be either positive, neutral, or negative: 'The movie was fantastic! The storyline was engaging, and the acting was top-notch.'
Topic Classification	Classify the following news article into one of four topics: World, Sports, Business, Sci/Tech: 'The central bank announced a new policy to regulate inflation and stabilize the economy.'
Text Summarization	Produce text summary of the following text: 'Harry Potter star Daniel Radcliffe gains access to a reported Â£20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. ...'
Machine Translation	Translate the following sentence from English to Arabic in the Qatar dialect: 'Artificial intelligence is transforming the future of technology.'
Textual Entailment	Given the sentences in the following format sent_1, sent_2, determine if sent_1 entails sent_2 or not: 'The sun is shining brightly., It is nighttime.'

summarization and machine translation, BERTScore is employed. BERTScore measures the quality of generated text by comparing it to reference texts, using contextual embeddings to capture semantic similarity, making it suitable for evaluating the quality and fluency of the model-generated outputs.

3.4.1. Classification Tasks

For Classification tasks, including sentiment analysis, topic classification, and textual entailment, standard classification evaluation measures are adopted:



DataSet	D19		
Records	Real Value	ChatGPT	DeepSeek
If you think you may need to help your elderly relatives with their finances, don't be shy about having the money talk -- soon.	Business	Business	Business
The purchasing power of kids is a big part of why the back-to-school season has become such a huge marketing phenomenon.	Business	Business	Sci/Tech
There is little cause for celebration in the stock market these days, but investors in value-focused mutual funds have reason to feel a bit smug -- if only because they've lost less than the folks who stuck with growth.	Business	Business	Business
The US trade deficit has exploded 19 to a record \$55.8bn as oil costs drove imports higher, according to a latest figures.	Business	World	World

Figure 1: Sample of Collected Responses Stored.

- Precision. Precision measures the proportion of correctly classified positive instances among all instances predicted as positive. A higher precision indicates fewer false positive classifications, which is critical when the cost of false positives is high. It is formally defined as:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

  where TP represents true positives, and FP denotes false positives.

- Recall. Recall quantifies the model's ability to identify all relevant positive instances correctly. A higher recall means that fewer relevant instances are missed by the model, which is important when it is essential not to overlook any positive cases. It is defined as:

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

  where FN represents false negatives.

- F1-Score. The F1-score is the harmonic mean of precision and recall, offering a balanced measure of both metrics. The F1-score is particularly useful when dealing with cases where precision and recall must be optimized simultaneously. It is computed as:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{3}$$

- Accuracy. Accuracy represents the proportion of correctly classified instances in all classes. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

where TN represents true negatives.

3.4.2. Generative Tasks

For evaluating text generation tasks (Text Summarization and Machine Translation), the BERTScore evaluation measure is used, which provides a more semantically meaningful assessment compared to traditional n-gram-based metrics:

BERT Score evaluation measure is used to evaluate text generation tasks, including text summarization and machine translation. BERT Score is particularly effective for assessing semantic adequacy, as it captures deeper contextual meaning rather than surface-level lexical similarity. Traditional measures like BLEU and ROUGE often focus on n-gram overlap, which may fail to account for semantic variations like paraphrasing or synonym usage. In contrast, BERTScore leverages the power of pretrained transformer models to understand the meaning behind words in context, making it highly suitable for assessing tasks like text summarization and machine translation. These tasks often involve rewording or rephrasing information, where word choice, syntactic variations, and semantic equivalence are crucial to a meaningful evaluation. BERTScore's ability to evaluate this semantic richness ensures a more accurate and relevant measure of the quality of the generated text. The measure is based on token-wise cosine similarity and is defined as:

$$BERTScore = \frac{1}{N} \sum_{i=1}^{N} \max_{j} cosine(E(x_i)E(y_i)), \tag{5}$$

where $E(x_i)$ and $E(y_i)$ represent contextual embeddings of tokens from the generated and reference texts, respectively.

4. Experimental Results and Discussion

This section presents the experimental results of ChatGPT and DeepSeek across five NLP tasks. The performance of LLMs is evaluated using two benchmark datasets per task. The collected responses are then analyzed to compare the effectiveness of both models.

4.1. Sentiment Analysis Task Results

The sentiment classification performance of ChatGPT and DeepSeek is evaluated using the IMDB and Multilingual Sentiment datasets. Table 4 and Table 6 present the confusion matrices for both models, while Table 5 and Table 7 report their precision, recall, F1-Score, and accuracy. Overall, DeepSeek demonstrates superior sentiment analysis performance compared to ChatGPT.

Table 4: Confusion Matrix of LLMs - Sentiment Analysis using IMDB dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Positive	Negative
Positive	39	11
Negative	1	49

(b)

	Positive	Negative
Positive	50	0
Negative	1	49

Table 5: Performance of LLMs - Sentiment Analysis using IMDB Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Positive	97.5%	78.0%	86.7%	87.9%
	Negative	81.4%	98.0%	88.9%	
DeepSeek	Positive	98.0%	100.0%	99.0%	99.0%
	Negative	100.0%	98.0%	99.0%	

Table 4 and Table 5 present the results collected from both LLMs using the IMDB dataset. The sentiment class in the IMDB dataset is a binary, where the review is either positive or negative. Notably, DeepSeek achieves a higher overall accuracy (99.0%) compared to ChatGPT (87.9%). The recall for positive sentiment is 100% for DeepSeek, indicating that it classified all positive samples correctly, whereas ChatGPT achieves 78.0%. Similarly, DeepSeek attains higher F1-Scores across both sentiment classes, reinforcing its superior performance in binary sentiment classification. These results prove that DeepSeek outperforms ChatGPT in sentiment analysis for this dataset, particularly in detecting positive sentiment, which is crucial in applications such as customer feedback analysis and opinion mining.

Table 6: Confusion Matrix of LLMs - Sentiment Analysis using Multilingual Sentiment Dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Positive	Neutral	Negative
Positive	34	13	3
Neutral	4	28	18
Negative	0	15	35

(b)

	Positive	Neutral	Negative
Positive	44	6	0
Neutral	5	36	9
Negative	2	14	34

Table 6 and Table 7 present the results collected from both LLMs using the Multilingual Sentiment dataset. This dataset is used to evaluate the LLMs' ability to handle more complex sentiment classification, which introduces an additional ”neutral” sentiment class. The confusion matrices in Table 6 highlight key performance differences between ChatGPT and DeepSeek. ChatGPT shows a modest performance with neutral sentiment, misclassifying a significant number of neutral samples as either positive or negative (eighteen as negative and four as positive). On the other hand, DeepSeek demonstrates a more stable classification performance, showing fewer

Table 7: Performance of LLMs - Sentiment Analysis using Multilingual Sentiment Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Positive	85.0%	68.0%	75.6%	64.7%
	Neutral	50.0%	56.0%	52.8%	
	Negative	62.5%	70.0%	66.0%	
DeepSeek	Positive	86.3%	88.0%	87.1%	76.0%
	Neutral	64.3%	72.0%	67.9%	
	Negative	79.1%	68.0%	73.1%	

misclassifications in the neutral category, contributing to its higher overall accuracy. While DeepSeek achieves an accuracy of 76.0%, ChatGPT lags at 64.7%. ChatGPT's F1-Score for neutral sentiment is significantly lower (52.8%) compared to DeepSeek (67.9%), highlighting its difficulty in handling this additional sentiment category. Furthermore, DeepSeek achieves consistently higher precision, recall, and F-score values across all sentiment classes.

4.2. Topic Classification Task Results

ChatGPT and DeepSeek are assessed in topic classification using the AG News Classification and Web of Science datasets. Table 8 and Table 10 show the confusion matrices for both LLMs. Table 9 and Table 11 summarize their precision, recall, F1-score, and accuracy. The results indicate that DeepSeek achieves higher performance than ChatGPT in topic classification.

Table 8 and Table 9 present the results collected from ChatGPT and DeepSeek using the AG News Classification dataset. This dataset comprises four categories: Business, Sci/Tech, Sport, and World. The confusion matrices in Table 6 highlight key classification patterns and challenges for each model. ChatGPT exhibits better performance in classifying Sci/Tech news articles than DeepSeek. In contrast, DeepSeek shows fewer misclassifications in the Business and Sport classes compared to ChatGPT. DeepSeek achieved a higher accuracy (81.3%) than ChatGPT (80%). ChatGPT shows high precision for Business and Sci/Tech (100%), while it exhibits a low precision for World class (62.1%). On average, ChatGPT's F1-Score is lower (80.03%) compared to

Table 8: Confusion Matrix of LLMs - Topic Classification using AG News Classification Dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Business	Sci/Tech	Sport	World
Business	11	0	0	9
Sci/Tech	0	17	3	0
Sport	0	0	18	2
World	0	0	2	18

(b)

	Business	Sci/Tech	Sport	World
Business	13	2	0	5
Sci/Tech	3	14	3	0
Sport	0	0	20	0
World	0	0	2	18

Table 9: Performance of LLMs - Topic Classification using AG News Classification Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Business	100.0%	55.0%	71.0%	80.0%
	Sci/Tech	100.0%	85.0%	91.9%	
	Sport	78.3%	90.0%	83.7%	
	World	62.1%	90.0%	73.5%	
DeepSeek	Business	81.3%	65.0%	72.2%	81.3%
	Sci/Tech	87.5%	70.0%	77.8%	
	Sport	80.0%	100.0%	88.9%	
	World	78.3%	90.0%	83.7%	

DeepSeek (80.65%). ChatGPT generally had better precision, making it more reliable when misclassifications need to be minimized. On the other hand, DeepSeek had a better recall, indicating it retrieved more correct instances per category.

Table 10 and Table 11 present the results of topic classification using the Web of

Table 10: Confusion Matrix of LLMs - Topic Classification using Web of Science Dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Biochemistry	Civil	CS	ECE	MAE	Medical	Psychology
Biochemistry	25	0	0	4	0	7	1
Civil	1	2	1	1	2	0	0
CS	0	0	15	1	1	1	0
ECE	0	1	0	3	1	0	1
MAE	2	1	0	1	1	0	0
Medical	11	0	0	1	0	9	3
Psychology	1	0	0	0	0	1	1

(b)

	Biochemistry	Civil	CS	ECE	MAE	Medical	Psychology
Biochemistry	19	1	1	1	0	13	2
Civil	0	4	1	2	0	0	0
CS	1	1	12	2	0	2	0
ECE	0	0	1	2	1	2	0
MAE	2	3	0	0	0	0	0
Medical	7	0	0	0	0	15	2
Psychology	1	0	0	0	0	2	0

Science dataset for both ChatGPT and DeepSeek. This dataset includes seven diverse scientific disciplines, posing a challenge for general-purpose LLMs. Table 10 highlights key classification patterns and challenges for each model. ChatGPT performs better in classifying CS than DeepSeek, correctly identifying 15 out of 18 instances. Similarly, Biochemistry is identified with moderate accuracy (25 correct predictions out of 37). On the other hand, ChatGPT demonstrates a higher misclassification rate in the Civil class compared to DeepSeek. A critical observation is DeepSeek's complete failure to classify Psychology and MAE, as all samples from these categories are misclassified into others, leading to an F1-score of 0.0%. Furthermore, Medical sam-

Table 11: Performance of LLMs - Topic Classification using Web of Science Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Biochemistry	62.5%	67.6%	64.9%	56.0%
	Civil	50.0%	28.6%	36.4%	
	CS	93.8%	83.3%	88.2%	
	ECE	27.3%	50.0%	35.3%	
	MAE	20.0%	20.0%	20.0%	
	Medical	50.0%	37.5%	42.9%	
	Psychology	16.7%	33.3%	22.2%	
DeepSeek	Biochemistry	63.3%	51.4%	56.7%	52.0%
	Civil	44.4%	57.1%	50.0%	
	CS	80.0%	66.7%	72.7%	
	ECE	28.6%	33.3%	30.8%	
	MAE	0.0%	0.0%	0.0%	
	Medical	44.1%	62.5%	51.7%	
	Psychology	0.0%	0.0%	0.0%	

ples are predominantly classified correctly (15 out of 18), contributing to higher recall in this category than ChatGPT. Overall, ChatGPT achieves a higher overall accuracy (56.0%) than DeepSeek (52.0%). For CS and Biochemistry, ChatGPT demonstrates a better F1-Score (88.2% and 64.9%, respectively) than DeepSeek (72.7% and 56.7%, respectively). However, ChatGPT faces challenges in classifying Civil (recall of 28.6%) and Medical (recall of 37.5%) scientific articles. In contrast, DeepSeek exhibits better recall in Medical (62.5%) and Civil (57.1%). Both LLMs generally struggle with niche fields like MAE and Psychology, suggesting domain-specific adaptation is needed for improved classification.

4.3. Text Summarization Task Results

ChatGPT and DeepSeek are assessed in text summarization task using Gigaword and CNN/Daily Mail datasets. The evaluation was carried out using the BERT Score, which includes three measures: precision, recall, and F1-Score. Table 12 and Table 13 summarize the results of these evaluations.

Table 12: Performance of LLMs - Text Summarization using Gigaword Dataset.

LLM	BERT Score		
	Precision	Recall	F1-Score
ChatGPT	71.00%	72.28%	71.59%
DeepSeek	70.62%	71.66%	71.11%

Table 12 presents the performance results of ChatGPT and DeepSeek for the Gigaword dataset. The Gigaword dataset consists of news articles and corresponding summaries, challenging both models to generate concise yet accurate summaries. ChatGPT slightly outperformed DeepSeek in all three BERT Score metrics. Specifically, ChatGPT achieved a precision of 71.00%, a recall of 72.28%, and an F1-Score of 71.59%. In comparison, DeepSeek obtained a precision of 70.62%, a recall of 71.66%, and an F1-Score of 71.11%. The results indicate that ChatGPT performs marginally better in precision and F1-Score. Therefore, ChatGPT better captures relevant content from the original text and includes it in the summarized output. Although the differences are minor, the slight edge in performance for ChatGPT could be attributed to its ability to balance content inclusion (recall) and conciseness (precision) more effectively.

Table 13: Performance of LLMs - Text Summarization using CNN/Daily Mail Summarization Dataset.

LLM	BERT Score		
	Precision	Recall	F1-Score
ChatGPT	68.80%	73.59%	71.09%
DeepSeek	68.44%	74.11%	71.13%

Table 13 presents the performance results of ChatGPT and DeepSeek for the CNN/Daily Mail dataset. The CNN/Daily Mail dataset is a widely used text summarization benchmark involving news articles and summaries requiring high levels of abstraction and content condensation. DeepSeek marginally outperformed ChatGPT in F1-Score. DeepSeek achieves a precision of 68.44%, a recall of 74.11%, and an F1-Score of 71.13%. In contrast, ChatGPT shows a precision of 68.80%, a recall of 73.59%, and an F1-Score of 71.09%. While ChatGPT demonstrates slightly higher precision, DeepSeek exhibits a marginally better recall and F1-Score. These results indicate that DeepSeek may be more effective at generating summaries with broader content coverage. On the other hand, ChatGPT excels in producing more concise summaries and performing better in terms of output brevity.

4.4. Machine Translation Task Results

The performance of ChatGPT and DeepSeek is evaluated on the Machine Translation task, which involved translating from English to Arabic. Two datasets are used for this evaluation: ArzEn-MultiGenre and AraBench. The LLMs are assessed using the BERT Score, which includes three measures: precision, recall, and F1-Score. Table 14 and Table 15 summarize the results for the respective datasets.

Table 14: Performance of LLMs - Machine Translation using ArzEn-MultiGenre Dataset.

LLM	BERT Score		
	Precision	Recall	F1-Score
ChatGPT	78.23%	78.61%	78.39%
DeepSeek	77.80%	77.32%	77.53%

Table 14 presents the performance results of ChatGPT and DeepSeek for the ArzEn-MultiGenre dataset. This dataset consists of Egyptian Arabic song lyrics, novels, and subtitles paired with English translations. ChatGPT outperformed DeepSeek in all three BERT Score metrics. Specifically, ChatGPT achieved a precision of 78.23%, recall of 78.61%, and an F1-Score of 78.39%. In comparison, DeepSeek attained a precision of 77.80%, recall of 77.32%, and an F1-Score of 77.53%. The results indicate that ChatGPT performs slightly better than DeepSeek across all BERT Score metrics. Hence, ChatGPT may more effectively capture relevant translation content in the generated text, leading to a higher recall value.

Table 15: Performance of LLMs - Machine Translation using AraBench Dataset.

LLM	BERT Score		
	Precision	Recall	F1-Score
ChatGPT	78.34%	78.37%	78.33%
DeepSeek	78.79%	78.60%	78.67%

Table 15 presents the performance results of ChatGPT and DeepSeek for the AraBench dataset, which includes Arabic-English translation pairs across various dialects. The evaluation focuses on Qatari and Jordanian Arabic. DeepSeek showed slight superiority in all three BERT Score metrics, achieving a precision of 78.79%, a recall of 78.60%, and an F1-Score of 78.67%. Conversely, ChatGPT achieved a precision of 78.34%, a recall of 78.37%, and an F1-Score of 78.33%. The results indicate that both LLMs perform similarly well in translating between English and Arabic, mainly when dealing with dialects such as Qatari and Jordanian Arabic.

4.5. Textual Entailment Task Results

The textual entailment performance of ChatGPT and DeepSeek is evaluated using the Scitail and FraCaS datasets. Table 16 and Table 18 present the confusion matrices for both LLMs. Table 17 and Table 19 report their precision, recall, F1-Score, and accuracy. Overall, DeepSeek outperforms ChatGPT in textual entailment tasks.

Table 16: Confusion Matrix of LLMs - Textual Entailment using Scitail Dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Entail	Neutral
Entail	25	0
Neutral	20	5

(b)

	Entail	Neutral
Entail	25	0
Neutral	18	7

Table 16 and Table 17 present the results collected from both LLMs using the Scitail dataset. The entailment class in the Scitail dataset is binary, where each text pair is classified as entailment or neutral. The confusion matrices in Table 16 highlight key performance differences between ChatGPT and DeepSeek. ChatGPT and DeepSeek

Table 17: Performance of LLMs - Textual Entailment using Scitail Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Entail	55.6%	100.0%	71.4%	60.0%
	Neutral	100.0%	20.0%	33.3%	
DeepSeek	Entail	58.1%	100.0%	73.5%	64.0%
	Neutral	100.0%	28.0%	43.8%	

correctly identify 25 instances as entailments (True Positives), with no misclassifications as neutral. For the neutral class, ChatGPT makes fewer misclassifications (5 out of 25) compared to DeepSeek. DeepSeek achieves an accuracy of 64.0%, while ChatGPT's accuracy is 60.0%. In terms of the entail class, DeepSeek outperforms ChatGPT with a higher precision (58.1%) and F1-Score (73.5%) compared to ChatGPT's precision of 55.6% and F1-Score of 71.4%. For the neutral class, ChatGPT has a lower recall (20.0%) and F1-Score (33.3%) compared to DeepSeek, which achieves a recall of 28.0% and F1-Score of 43.8%.

Table 18: Confusion Matrix of LLMs - Textual Entailment using FraCaS Dataset. (a) ChatGPT. (b) DeepSeek.

(a)

	Entail	Neutral	Contradict
Entail	16	5	4
Neutral	10	13	2
Contradict	1	5	19

(b)

	Entail	Neutral	Contradict
Entail	20	4	1
Neutral	11	10	4
Contradict	2	1	22

Table 18 and Table 19 present the results collected from ChatGPT and DeepSeek

Table 19: Performance of LLMs - Textual Entailment using FraCaS Dataset.

LLM		Precision	Recall	F1-Score	Accuracy
ChatGPT	Entail	59.3%	64.0%	61.5%	64.0%
	Neutral	56.5%	52.0%	54.2%	
	Contradict	76.0%	76.0%	76.0%	
DeepSeek	Entail	60.6%	80.0%	69.0%	69.3%
	Neutral	66.7%	40.0%	50.0%	
	Contradict	81.5%	88.0%	84.6%	

using the FraCaS dataset. This dataset evaluates the LLMs' ability to handle more complex entailment classification, introducing an additional ”Contradicts” class. Table 18 presents the confusion matrices for both models on the FraCaS dataset. DeepSeek correctly identifies 20 out of 25 instances as entailments, while ChatGPT correctly identifies 16. Similarly, DeepSeek correctly specifies 22 out of 25 contradictions, while ChatGPT correctly identifies 19. ChatGPT misclassifies 12 instances for the neural class and correctly identifies 13 instances. DeepSeek misclassifies 14 entailments as neutral and correctly identifies 10 neutral relations. The overall accuracy of DeepSeek is 69.3%, outperforming ChatGPT (64.0%). For entailment, ChatGPT achieves a precision of 59.3% and a recall of 64.0%. ChatGPT performs less effectively for the neutral class, with a precision of 56.5% and a recall of 52.0%. However, ChatGPT achieves a high precision and recall of 76.0% for contradiction. DeepSeek shows improved performance on the entailment class with a precision of 60.6% and a recall of 80.0%. However, it underperforms for neutral relations with a low recall of 40.0%. DeepSeek excels in contradiction prediction, achieving an 81.5% precision and an 88.0% recall. ChatGPT demonstrates a higher F1-Score (54.2%) than DeepSeek (50.0%) for the neutral class. In contrast, DeepSeek exhibits a better performance in terms of F1-Score than ChatGPT in terms of entailments and instances of contradiction.

Textual entailment tasks assess the ability of ChatGPT and DeepSeek to determine whether a hypothesis logically follows from a given premise. Evaluations using the SciTail (Table 17) and FraCaS (Table 19) datasets reveal that DeepSeek outperforms Chat-

GPT in overall accuracy (64.18% vs. 62.00%). On SciTail, a binary entailment classification task, DeepSeek achieves an accuracy of 64.0% compared to ChatGPT's 60.0%, showing stronger performance in correctly identifying entailment relationships. On FraCaS, which introduces an additional "contradiction" class, DeepSeek demonstrates better classification balance, particularly excelling in detecting contradictions (81.5% precision vs. 76.0% for ChatGPT). Hence, DeepSeek is superior in distinguishing contradictions due to better context comprehension. Both models exhibit weaknesses in handling highly nuanced entailment cases, indicating potential for improvement in logical reasoning tasks. Overall, DeepSeek shows better performance, with an accuracy improvement of 7.50%.

4.6. Discussion and Comparative Insights

This section provides a comprehensive discussion of the performance of ChatGPT and DeepSeek across the five evaluated tasks: sentiment analysis, topic classification, text summarization, machine translation, and textual entailment. The analysis highlights key trends, strengths, and challenges each LLM faces, offering comparative insights that reveal their respective capabilities. Table 20 and Table 21 summarize the average performance per task and the percentage improvement achieved by DeepSeek. The mean results show that ChatGPT outperforms DeepSeek in three out of five NLP tasks.

ChatGPT and DeepSeek exhibit distinct performance patterns when dealing with sentiment classification, particularly in a neutral sentiment class. DeepSeek outperforms ChatGPT in overall accuracy, achieving 76.0% compared to ChatGPT's 64.7% (Table 7). The confusion matrices reveal that ChatGPT struggles with neutral sentiment, frequently misclassifying neutral samples as positive or negative. In contrast, DeepSeek maintains a more stable classification, demonstrating fewer errors in identifying neutral sentiment, contributing to its superior performance. DeepSeek is generally more balanced across all sentiment classes, leading to fewer extreme misclassifications. Overall, DeepSeek outperforms ChatGPT significantly in all metrics, with a 14.68% improvement in accuracy.

Topic classification remains challenging for both LLMs, particularly in distinguish-

Table 20: Summary of ChatGPT and DeepSeek Results on Text Classification Tasks

Task	LLM	Precision	Recall	F1-Score	Accuracy
Sentiment Analysis	ChatGPT	75.28%	74.00%	74.00%	76.30%
	DeepSeek	85.54%	85.20%	85.22%	87.50%
Improvements		13.63%	15.14%	15.17%	14.68%
Topic Classification	ChatGPT	60.06%	58.21%	57.27%	68.00%
	DeepSeek	53.41%	54.18%	53.14%	66.65%
Improvements		-11.08%	-6.92%	-7.22%	-1.99%
Textual Entailment	ChatGPT	69.48%	62.40%	59.28%	62.00%
	DeepSeek	73.38%	67.2%	64.18%	66.65%
Improvements		5.62%	7.7%	8.27%	7.5%

Table 21: Summary of ChatGPT and DeepSeek Results on Text Generation Tasks

Task	LLM	BERT Score		
		Precision	Recall	F1-Score
Text Summarization	ChatGPT	69.90%	72.94%	71.34%
	DeepSeek	69.53%	72.89%	71.12%
Improvements		-0.53%	-0.07%	-0.31%
Machine Translation	ChatGPT	78.29%	78.49%	78.36%
	DeepSeek	78.30%	77.96%	78.10%
Improvements		0.0128%	-0.68%	-0.33%

ing closely related categories. Both models exhibited difficulties with overlapping topics, such as Medical and Psychology, especially in the Web of Science dataset (Table 11). Similarly, in the AG News dataset (Table 9), ChatGPT faces difficulties differentiating between Business and World News due to semantic similarities. Hence, these findings suggest a need for fine-tuning or incorporating additional context-aware mechanisms. Overall, ChatGPT demonstrates superior precision and accuracy across most categories, making it more suitable for technical fields like CS and Biochemistry. However, DeepSeek performs better for domains requiring high recall, such as Medical and

Civil Engineering. Both models require further improvement in niche areas like MAE and Psychology. Additionally, they struggle with ambiguous topics, leading to higher misclassification rates. In this task, ChatGPT outperforms DeepSeek, achieving an accuracy of 68.00% compared to 66.65%, reflecting a 1.99% improvement.

The text summarization task assesses how well each model generates concise and informative summaries. Performance assessments on the Gigaword (Table 12) and CNN/Daily Mail (Table 13) datasets reveal a mixed trend. On the Gigaword dataset, ChatGPT marginally outperforms DeepSeek across all BERT Score metrics, achieving a precision of 71.00% and an F1-Score of 71.59%, compared to 70.62% precision and 71.11% F1-Score for DeepSeek. Conversely, on CNN/Daily Mail dataset, DeepSeek slightly outperforms ChatGPT in F1-Score (71.13% vs. 71.09%), but ChatGPT achieves slightly higher precision. The results indicate that ChatGPT generates more concise summaries, contributing to higher precision. DeepSeek demonstrates superior recall, suggesting a broader content capture. However, the differences between the LLMs are minor, indicating both are effective summarization tools, with trade-offs depending on whether conciseness (ChatGPT) or content coverage (DeepSeek) is prioritized. Overall, ChatGPT outperforms DeepSeek by 0.31% in this task.

The machine translation task evaluates how well ChatGPT and DeepSeek handle English-to-Arabic translations using the ArzEn-MultiGenre (Table 14) and AraBench (Table 15) datasets. The results reveal that performance varies based on the dialect of Arabic used. On ArzEn-MultiGenre, which includes Egyptian Arabic song lyrics, novels, and subtitles, ChatGPT outperforms DeepSeek across all BERT Score metrics, achieving an F1-Score of 78.39% vs. 77.53% for DeepSeek. On AraBench, which includes Jordanian and Qatari dialects, DeepSeek slightly outperforms ChatGPT, achieving an F1-Score of 78.67% vs. 78.33% for ChatGPT. Despite these minor variations, both models show similar overall performance, with slight differences based on the dataset and the Arabic dialect. In general, ChatGPT demonstrates stronger performance in machine translation tasks compared to DeepSeek, with an 0.3318% improvement.

To recap, DeepSeek demonstrates a stronger performance in structured tasks, such as sentiment analysis and textual entailment, where its classification stability is likely a contributing factor. The model consistently exhibits reliable results in these tasks, indicating its robustness in handling structured data with clear categories. On the other hand, ChatGPT excels in more subjective and nuanced tasks, such as topic classification, summarization, and certain translation cases. These findings indicate that ChatGPT may possess a more refined understanding of context, enabling it to perform well in tasks that require a deeper interpretation of meaning and subtleties.

Both models perform similarly regarding text summarization and translation, with the trade-off between precision and recall being a key differentiating factor. ChatGPT tends to prioritize conciseness and precision, while DeepSeek may capture more comprehensive content, showing better recall. These findings indicate that the choice between the models may depend on the specific requirements of the task, such as whether precision or recall is more critical.

Despite these individual strengths, neither model consistently outperforms the other across all tasks. Their applicability depends on the nature of the task, suggesting that one model may be more suitable than the other depending on specific requirements. Moreover, improvements are needed in handling tasks that involve neutrality and contradictions. Specifically, ChatGPT shows room for improvement in sentiment analysis and textual entailment tasks, where it struggles with neutral classifications and contradictions.

5. Conclusion

This study evaluated the performance of ChatGPT and DeepSeek across multiple NLP tasks, including sentiment analysis, topic classification, text summarization, machine translation, and textual entailment. The methodology employed in this study ensured a systematic and unbiased evaluation of both LLMs. A well-defined experimental protocol governed the querying process, response collection, and evaluation framework, minimizing variability sources and ensuring comparison fairness. Identical prompts are used for both LLMs and are designed to be neutral, clear, and representative of real-world use cases. For tasks like textual entailment and machine translation, few-shot examples are incorporated to evaluate the models' contextual responses. Two benchmark datasets are selected for each NLP task, ensuring a comprehensive evaluation across diverse domains, including news, reviews, and formal/informal texts.

The results indicate that both LLMs exhibit strengths and weaknesses depending on the nature of the task. DeepSeek generally excels in structured tasks, such as sentiment analysis and textual entailment, where its classification stability and logical reasoning abilities are particularly evident. In contrast, ChatGPT demonstrates superior performance in more subjective and nuanced tasks, including topic classification, summarization, and translation, where its refined contextual understanding is particularly valuable. Both models show similar performance in summarization and translation, with trade-offs between precision and recall, making their suitability task-dependent. DeepSeek favors recall, ensuring comprehensive content capture, whereas ChatGPT prioritizes precision, leading to more concise outputs. However, both LLMs have areas for improvement. Specifically, handling contradictions, neutrality, and domain-specific classifications remains challenging. These weaknesses point to the need for further advancements in logical reasoning and nuanced content understanding for both models.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, .

[2] Al-Sabbagh, R. (2024). Arzen-multigenre: An aligned parallel dataset of egyptian arabic song lyrics, novels, and subtitles, with english translations. *Data in Brief*, *54*, 110271. URL: `http://dx.doi.org/10.1016/j.dib.2024.110271`. doi:10.1016/j.dib.2024.110271.

[3] Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the performance of chatgpt in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmology science*, *3*, 100324.

[4] Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations

from scientific publications. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 546–555). Vancouver, Canada: Association for Computational Linguistics. URL: `https://aclanthology.org/S17-2091/`. doi:10.18653/v1/S17-2091.

[5] Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 274–279). IEEE.

[6] Biltawi, M., Etaiwi, W., Tedmori, S., Hudaib, A., & Awajan, A. (2016). Sentiment classification techniques for arabic language: A survey. In *2016 7th International Conference on Information and Communication Systems (ICICS)* (pp. 339–346). doi:10.1109/IACS.2016.7476075.

[7] Caramancion, K. M. (2023). News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. In *2023 IEEE Future Networks World Forum (FNWF)* (pp. 1–6). IEEE.

[8] Chen, Y., Wang, H., Yu, K., & Zhou, R. (2024). Artificial intelligence methods in natural language processing: A comprehensive review. *Highlights in Science Engineering and Technology*, *85*, 545–550.

[9] Coello, C. E. A., Alimam, M. N., & Kouatly, R. (2024). Effectiveness of chatgpt in coding: A comparative analysis of popular large language models. *Digital*, *4*, 114–125. URL: `http://dx.doi.org/10.3390/digital4010005`. doi:10.3390/digital4010005.

[10] Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M. et al. (1996). *Using the framework*. Technical Report Technical Report LRE 62-051 D-16, The FraCaS Consortium.

[11] Deandres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., & Ortega-Garcia, J. (2024). How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability. *IEEE Access*, .

[12] Del Corso, G. M., Gullí, A., & Romani, F. (2005). Ranking a stream of news. In *Proceedings of the 14th International Conference on World Wide Web* WWW '05 (p. 97–106). New York, NY, USA: Association for Computing Machinery. URL: `https://doi.org/10.1145/1060745.1060764`. doi:10.1145/1060745.1060764.

[13] Dunder, N., Lundborg, S., Wong, J., & Viberg, O. (2024). Kattis vs chatgpt: Assessment and evaluation of programming tasks in the age of artificial intelligence. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 821–827).

[14] Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in psychology*, *14*, 1199058.

[15] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, *66*, 111–126.

[16] Fidelangeli, A., Galli, F., Loreggia, A., Pisano, G., Rovatti, R., Santin, P., & Sartor, G. (2025). The summarization of italian tax-law decisions: The case of the prodigit project. *IEEE Access*, *13*, 38833–38855. URL: `http://dx.doi.org/10.1109/ACCESS.2025.3545419`. doi:10.1109/access.2025.3545419.

[17] Fu, X., Wang, R., & Li, C. (2024). Can chatgpt evaluate plans? *Journal of the American Planning Association*, *90*, 525–536.

[18] González, R., Poenaru, D., Woo, R., Trappey, A. F., Carter, S., Darcy, D., Encisco, E., Gulack, B., Miniati, D., Tombash, E., & Huang, E. Y. (2024). Chatgpt: What every pediatric surgeon should know about its potential uses and pitfalls. *Journal of Pediatric Surgery*, *59*,

941–947. URL: http://dx.doi.org/10.1016/j.jpedsurg.2024.01.007. doi:10.1016/j.jpedsurg.2024.01.007.

[19] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, (p. 122666).

[20] Jiang, Q., Gao, Z., & Karniadakis, G. E. (2025). Deepseek vs. chatgpt vs. claude: A comparative study for scientific computing and scientific machine learning tasks. *Theoretical and Applied Mechanics Letters*, (p. 100583). URL: http://dx.doi.org/10.1016/j.taml.2025.100583. doi:10.1016/j.taml.2025.100583.

[21] Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E. et al. (2023). Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, (pp. rs–3).

[22] Khlaif, Z. N., Mousa, A., Hattab, M. K., Itmazi, J., Hassan, A. A., Sanmugam, M., & Ayyoub, A. (2023). The potential and concerns of using ai in scientific research: Chatgpt performance evaluation. *JMIR Medical Education*, *9*, e47049.

[23] Khot, T., Sabharwal, A., & Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*. URL: http://dx.doi.org/10.1609/aaai.v32i1.12022. doi:10.1609/aaai.v32i1.12022.

[24] Kowsari, K. (2018). Web of science dataset. URL: https://data.mendeley.com/datasets/9rw3vkcfy4/6. doi:10.17632/9RW3VKCFY4.6.

[25] Kowsari, K., Brown, D. E., Heidarysafa, M., Jafari Meimandi, K., , Gerber, M. S., & Barnes, L. E. (2017). Hdltex: Hierarchical deep learning for text classification.

In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE.

[26] Kumar, P. (2024). Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*, 260.

[27] Li, Z., Fan, S., Gu, Y., Li, X., Duan, Z., Dong, B., Liu, N., & Wang, J. (2024). Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 18608–18616). volume 38.

[28] Liao, H. (2025). Deepseek large-scale model: technical analysis and development prospect. *Journal of Computer Science and Electrical Engineering*, *7*, 33–37.

[29] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, .

[30] Liu, C., Xie, Z., Zhao, S., Zhou, J., Xu, T., Li, M., & Chen, E. (2024). Speak from heart: an emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (pp. 533–542).

[31] Liu, X., Zhu, Y., Pang, T., Xue, K., Zhang, X., & Fan, C. (2024). Medical document embedding enhancement with heterogeneous mixture-of-experts. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2238–2243). doi:10.1109/BIBM62325.2024.10822374.

[32] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. URL: `http://www.aclweb.org/anthology/P11-1015`.

[33] Mo, L., & Wu, K. (2025, Access Date: 29/3/2025). Deepseek narrows china-us ai gap to three months, 01.ai founder lee kai-fu says. *Reuters*, .

[34] 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., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* NIPS '22. Red Hook, NY, USA: Curran Associates Inc.

[35] Rezaei, M., Salehi, H., & Tabatabaei, O. (2024). Uses and misuses of chatgpt as an ai-language model in academic writing. In *2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)* (p. 256–260). IEEE. URL: `http://dx.doi.org/10.1109/QICAR61538.2024.10496607`. doi:10.1109/qicar61538.2024.10496607.

[36] Rush, A. M. (2024). Gigaword dataset. URL: `https://service.tib.eu/ldmservice/dataset/5e1639e1-c107-48c0-80a8-c593c6b8ad5b`. doi:10.57702/XAWJFKMB.

[37] Sajjad, H., Abdelali, A., Durrani, N., & Dalvi, F. (2020). Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics* (p. 5094–5107). International Committee on Computational Linguistics. URL: `http://dx.doi.org/10.18653/v1/2020.coling-main.447`. doi:10.18653/v1/2020.coling-main.447.

[38] van Schaik, T. A., & Pugh, B. (2024). A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2832–2836).

[39] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers) (pp. 1073–1083). Vancouver, Canada: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/P17-1099`. doi:10.18653/v1/P17-1099.

[40] Silva, G., Ferreira, R., Lins, R. D., Cabral, L., Oliveira, H., Simske, S. J., & Riss, M. (2015). Automatic text document summarization based on machine learning. In *Proceedings of the 2015 ACM Symposium on Document Engineering* (pp. 191–194). ACM.

[41] Sunagar, P., Kanavalli, A., Nayak, S. S., Mahan, S. R., Prasad, S., & Prasad, S. (2021). News topic classification using machine learning techniques. In V. Bindhu, J. M. R. S. Tavares, A.-A. A. Boulogeorgos, & C. Vuppalapati (Eds.), *International Conference on Communication, Computing and Electronics Systems* (pp. 461–474). Singapore: Springer Singapore.

[42] Thelwall, M. (2024). Can chatgpt evaluate research quality? *Journal of Data and Information Science*, *9*, 1–21.

[43] Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143–153. URL: `http://dx.doi.org/10.1016/j.eng.2021.03.023`. doi:10.1016/j.eng.2021.03.023.

[44] Wang, X., Tang, Z., Guo, J., Meng, T., Wang, C., Wang, T., & Jia, W. (2025). Empowering edge intelligence: A comprehensive survey on on-device ai models. *ACM Computing Surveys*, .

[45] Yoo, T., & Cheong, Y.-G. (2024). Leveraging llm-constructed graphs for effective goal-driven storytelling. In *CEUR Workshop Proceedings* (pp. 83–95). CEUR-WS volume 3818.