UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Evaluation of AI text generators

Sanaz Khoobi

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2024
2025

Faculty of Business Economics

Master of Management

Master's thesis

Evaluation of AI text generators

Sanaz Khoobi

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Data Science

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

Abstract

This thesis explores the evaluation of AI-generated text by comparing the performance of commonly used language models across a variety of natural language generation (NLG) tasks. The study thoroughly investigates their abilities in semantic comprehension, summarization, grammar sensitivity, fact-checking, and decision-making. To analyze the quality and reliability of the generated outputs, an integrated method is used, combining human evaluation with automated measurements. The results show significant differences in model performance, showing GPT-4's overall superiority in human-likeness and factual correctness, as well as identifying certain weaknesses in vision and coherence under task complexity. This study adds to the expanding discussion about benchmarking LLMs and offers insights into their appropriate deployment in real-world circumstances.

Keywords:

Large Language Models, Natural Language Generation, Text Evaluation, GPT-4, Human Evaluation, Automated Metrics, Semantic Analysis, AI Text Generation

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Dr. Koen Vanhoof, for his invaluable guidance, patience, and continuous support throughout this thesis. His insightful feedback and encouragement have been crucial to the successful completion of this research.

I would also like to express my appreciation to Professor Saeid Amiri, whose linguistic skills and wise assistance helped shape this study.

Furthermore, I am grateful to the academic staff and my fellow students at Hasselt University for their support and collaboration during this journey.

Finally, I wish to thank my family and close friends for their unwavering encouragement and emotional support, without which this work would not have been possible.

Table of Contents

List of Tables

1. Introduction

1.1 Background

Large language models (LLMs) have developed greatly in recent years, which has caused a change in our views on natural language processing. Tools such as GPT-4 (OpenAI, 2023), Claude 3.7 (Anthropic, 2025), Gemini 1.5 (Google, 2025), and Command R (Cohere, 2025), among many others, have effectively supported humans in producing fluent, human-like text, as well as identifying complicated language patterns and combining a range of tasks from content summarization to decision making.

The basic architecture of modern LLMs, specifically the Transformer model described by Vaswani et al. (2017), makes use of paying attention processes to model dependence over a long time in text. This innovation led to in successes in fields such as machine translation (Bahdanau et al., 2015), conversation systems (Zhang et al., 2020), and abstractive summarization.
As LLMs grow in popularity in key sectors such as the medical field, education, media, and finance (Bommasani et al., 2021), it is essential to evaluate their impact on other measures. Although these models create very fluent writing, questions remain about their actual reliability, the meaning consistency, and logic capacity (Ji et al., 2023; Maynez et al., 2020).

1.2 Problem Statement

Although their capabilities are amazing, LLMs are known for producing outputs that contain hallucinations (true mistakes that have no root in reality), grammatical errors (errors made that keep the original meaning), and contextual gaps (Ji et al., 2023; Krishna et al., 2023). Traditional evaluation metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) primarily focus on surface-level text similarity. While useful for measuring overlap with reference texts, these tools often fail to capture deeper dimensions of output quality—such as logical reasoning, factual integrity, and resilience to input noise.

Furthermore, while standard evaluations—such as SuperGLUE (Wang et al., 2019) and TruthfulQA (Lin, Hilton, & Evans, 2022)—are useful, they rarely reflect real-world conditions in which user inputs may be grammatically flawed, semantically complex, or contain overlapping intents (e.g., requests for both "process" and "decision" explanations). Consequently, there is a growing need for task-specific, human-cantered, and context-aware evaluations of LLMs that align more closely with practical, real-life usage scenarios. These typical techniques also fail to capture other human-centric features like semantic alignment, factual accuracy, hallucination severity, and task relevance.

1.3 Research Motivation and Objectives

This study is motivated by the growing disconnect between theoretical evaluations of large language models (LLMs) and their performance in real-world contexts. As LLMs become increasingly integrated into business processes for summarization, decision-making, and content generation (Shinn et al., 2023; Bommasani et al., 2021), concerns arise regarding the quality, reliability, and reasoning strength of their outputs.

In particular, there is a lack of research examining how input quality (e.g., clean vs. noisy data) impacts performance, and whether models can handle nuanced distinctions such as between processes and decisions. Moreover, most evaluation benchmarks fail to reflect the real-world complexity of human prompts, which are often grammatically flawed or semantically hybrid.

To address these gaps, this study sets the following objectives:

- Objective 1: Evaluate the practical performance of top-tier LLMs (GPT-4, Claude 3.7, Gemini, Command R) across diverse NLP tasks such as summarization, question answering, grammatical correction, decision analysis, and fact-checking.
- Objective 2: Investigate how varying input quality (clean vs. noisy) affects model reliability and fluency.
- Objective 3: Assess the models' ability to distinguish between complex semantic structures like processes versus decisions.
- Objective 4: • Objective 4: Benchmark AI-generated outputs against human-written references using both fluency-based and semantic-level criteria such as accuracy, task relevance, and hallucination detection.

1.4 Research Questions

Based on the gaps that were found in the literature, the study tackles the following basic research questions:

1. Does source quality (clean vs. noisy text) affect the language structure and logical quality of AI-generated conclusions?
2. Can LLMs accurately read hybrid semantic constructs (e.g., process vs. decision) in real-world texts?
3. How do AI-generated final products compare to human-written outputs in terms of accuracy, harmony, and specific to a task performance (summarization, question answering, grammatical correction, fact-checking)?

1.5 Research Contributions

This study offers several important contributions to the NLP and AI evaluation literature:

- Real-world-focused Evaluation:
  Unlike previous researches, which focused on artificial or measured data, this study evaluated LLMs using texts that reflect real-world situations, such as grammatical errors and mixed linguistic purpose.
- Cross-task Contrast:
  The thesis compares AI-generated outputs against human sources within a number of actions offering a deeper examination of model performance.
- Semantic Reasoning Assessment:
  The study emphasizes the limitations of current LLMs' semantic understanding by creating examines that test their ability to distinguish between processes and decisions.
- Error Tracking and Resiliency Tests:
  The study looks at the most common forms of errors made by LLMs and assesses their capacity to overcome noise and actual error.

These efforts stand a foundation for stronger useful assessment frameworks that are better aligned with real-world AI deployment requirements.

1.6 Methodological Approach

To achieve the research aims, this study takes a systematic, hybrid empirical strategy, which involves the following steps:

- Model Selection: Four innovative LLMs are chosen for evaluation—GPT-4 (OpenAI, 2023), Claude 3.7 (Anthropic, 2025), Gemini (Google, 2025), and Command R (Cohere, 2025)—based on availability, popularity, and technical diversity.

- Six NLP tasks are aimed at modelling practical application cases, such as noisy vs. clean summarization, linguistic grouping, grammatical correction, structured decision-making, and fact verification. Each activity requires the creation or adaptation of both artificial and real-world datasets.
- Evaluation Strategy: Human evaluation followed a 5-point Likert scale (Amidei, Piwek, & Willis, 2019) and included dimensions such as semantic match, task relevance, hallucination severity, and factual accuracy—alongside standard automated metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020).
- Comparative Analysis: The final results of each model are compared to human-written references to identify both quantitative and qualitative differences, strengths, and limitations.

These tasks were selected to reflect high-frequency NLP use cases encountered in business, education, and information services.

This methodological approach aims to balance experimental control with real-world applicability, allowing generalization of findings regarding LLM behaviour in practical conditions.

2. Literature Review

2.1 NLP and the Evolution of Large Language Models

Natural Language Processing (NLP) is a multidisciplinary field that merges linguistics, computer science, and artificial intelligence. It enables computers to understand and manipulate human language, and it has underpinned progress in systems like chatbots, translation engines, and automated summarizers (Jurafsky & Martin, 2021). A key turning point in NLP has been the rise of Large Language Models (LLMs), particularly those built using transformer architectures first proposed by Vaswani et al. (2017).

By training on massive datasets, these models learn complex patterns in grammar, semantics, and contextual dependencies. Notable examples include GPT (Radford et al., 2018; Brown et al., 2020), BERT (Devlin et al., 2018), T5 (Raffel et al., 2020), and more recent offerings like Claude (Anthropic, 2025), Gemini (Google, 2025), and Command R (Cohere, 2025). While their ability to produce fluent and cohesive text is impressive, questions persist regarding their consistency, factuality, and reasoning capabilities.

2.2 Evaluating AI Text: Metrics and Methods

Assessing AI-generated language is a complex challenge. Because language is inherently nuanced, many different outputs may all be acceptable for a given prompt. Evaluation techniques therefore span from subjective human judgments to standardized automated metrics.

Human review remains the top standard, and this study used tested, blind evaluation utilizing qualitative assessments such as semantic match, hallucination severity, and task relevance.

It enables the assessment of subtle traits like fluency and factual accuracy (Belz & Kow, 2010; Amidei et al., 2019). However, human reviews are resource-intensive, prompting a growing reliance on rubric-based scoring tools that evaluate features like coherence or relevance using scales.

Among automated metrics, BLEU (Papineni et al., 2002) has long been a staple. It checks how much a model's output overlaps with a reference text. Yet, this metric emphasizes surface-level similarity and struggles to credit valid paraphrasing or semantic alignment (Callison-Burch et al., 2006; Reiter, 2018).

ROUGE (Lin, 2004) extends evaluation to recall, making it more useful for summarization. But like BLEU, it still overlooks sentence structure, grammar, and deeper logical relations (Graham, 2015).

BERTScore (Zhang et al., 2020) marks a significant advancement by using contextual embeddings to assess similarity, even when word choices differ. This approach has been found to better match human ratings in tasks like summarization (Sellam et al., 2020), though it too has drawbacks—such as dependency on the specific underlying model and difficulty comparing across domains.

2.3 Persistent Challenges in LLM Outputs

Even as LLMs grow more capable, several recurring flaws limit their reliability in sensitive domains.

Hallucinations—outputs that appear fluent and convincing questions remain about but are factually incorrect—remain common. Ji et al. (2023) highlight how even top-performing models can fabricate details, a concern in areas like healthcare or legal services where precision is vital.

Factual inconsistency is another risk. When summarizing long or complex texts, models may distort the original content or overlook key facts (Maynez et al., 2020). This inconsistency undermines trust, especially when outputs are used in decision-making.

Semantic misalignment, as described by Kryscinski et al. (2020), refers to grammatically correct responses that fail to capture the intended meaning. Such errors often go undetected by standard metrics, which focus more on surface-level text properties.

2.4 Benchmarking and Research Gaps

Benchmark datasets such as TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2021), and SuperGLUE (Wang et al., 2019) have been widely used in the literature to evaluate LLMs on reasoning and factuality. While this study does not use traditional benchmarks, it does use task-driven difficulties based on real-world settings, such as noisy input, hybrid semantics, and factual inaccuracies.

Preliminary technical reports and early benchmarks suggest that models like GPT-4 (OpenAI, 2023) and Claude 3.7 (Anthropic, 2025) demonstrate stronger factual consistency across tasks, whereas Gemini (Google, 2025) and Command R (Cohere, 2025) show more variability when dealing with noisy or ambiguous user input.

2.5 Summary of Gaps and Study Relevance

Despite their remarkable abilities, LLMs are not yet robust or accountable enough for high-stakes use. Miss deeper concerns of meaning and context, which this study aims to capture using human assessment and demanding input variations.

This review identifies several key gaps:

1. Limited integration of human and semantic-based evaluation.
2. Inadequate testing on noisy, grammatically flawed, or deceptive input.
3. Poor detection of semantic structure and logical progression.
4. Lack of reliable comparisons between AI and human-generated responses in complex settings.

The present study addresses these needs by using a blended evaluation method that combines expert human assessment with automatic scoring across challenging input types. Additionally, this thesis also includes model-wise evaluations of multiple NLP tasks under different input conditions to investigate real model limits.

3. Research Methodology

3.1 Study Design and Model Selection

This study uses a systematic empirical methodology to investigate the real-world performance of top large language models (LLMs) across a variety of NLP tasks. Rather than using constructed benchmarks, the evaluation focuses on realistic input conditions—such as grammatical noise and semantic ambiguity—to see how models manage practical linguistic barriers. Qualitative (human-centered) and quantitative (metric-based) evaluation methodologies are used to provide a comprehensive perspective of model capabilities.

3.2 Task Design and Data Collection

To assess LLM performance in practical application cases, the study included six fundamental NLP tasks meant to replicate frequent issues in language interpretation and creation, including summarization, classification, grammatical correction, and fact verification. The task designs aimed at capturing characteristics such as fluency, reasoning, semantic interpretation, and robustness to noisy input.

These activities were supplemented by a mix of adapted academic texts and synthetically produced examples that represented both clean and unclear input types. Word length, topic difficulty, and grammatical structure were all standardized in the inputs. Chapter 4 provides complete definitions of the six tasks, as well as extensive examples, datasets, and source references.

Data Preparation and Input Standardization

The tasks were complemented by a combination of real-world academic texts and synthetic examples intended to mimic actual NLP issues. While inputs were balanced for length, complexity, and language qualities (such as noise and ambiguity), the precise sources and examples utilized in each task are detailed in Chapter 4.

3.3 Evaluation Methodology

A hybrid evaluation methodology was used to analyse results based on linguistic quality, factual correctness, and task performance (Belz & Reiter, 2006; Zhang et al., 2020).

A. Human Evaluation

Human assessments were carried out using a 5-point Likert scale. The main criteria used for scoring were:

- Linguistic fluency and coherence: The logical flow and readability of the output.
- Clarity and correctness: Whether the response clearly and accurately reflected the intended meaning of the task prompt.
- Factual consistency: The degree to which the generated content matched or preserved factual elements of the input.

The evaluation process involved one academic expert in computational linguistics and AI ethics, and 13 graduate students or research collaborators. While no formal inter-rater agreement metric (e.g., Cohen’s Kappa) was applied, all evaluators were given shared rubrics and participated in a pilot calibration phase to minimize subjectivity and encourage consistency in scoring. Importantly, raters were blind to the model identities during assessment to avoid bias.

While the scoring focused on overall quality assessments, raters were asked to implicitly consider elements such as semantic match, hallucination risk, and importance using an assessment form provided during validation. Each task case was evaluated once by a single rater from the set group; the sample size and one-time nature of the evaluation are discussed in Chapter 4.

B. Automated Evaluation Metrics

To complement human scoring, automated evaluation was conducted using three established metrics:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Used for measuring word overlap in summarization tasks, including ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004).
- BLEU (Bilingual Evaluation Understudy): Measured the precision of n-gram matches between generated and reference texts, particularly for translation or paraphrasing tasks (Papineni et al., 2002).
- BERTScore: Captured semantic similarity at the sentence level using contextual embeddings from pretrained BERT-like models, helpful in comparing paraphrased but semantically equivalent responses (Zhang et al., 2020).

All metric scores were computed using Python-based tools including Hugging Face Transformers, SacreBLEU, and rouge-score (Wolf et al., 2020) and results were validated against human assessments to identify metric-human misalignment, particularly in semantically hybrid scenarios.

3.4 Experimental Controls and Validity Measures

To ensure methodological rigor, several control mechanisms were implemented:

- Prompt Consistency: All models were given identical prompts and instructions to eliminate input variation bias (Belz & Reiter, 2006).
- Model Isolation: Each model was tested in independent sessions, preventing leakage or prompt memory effects.
- Balanced Dataset: Texts were selected from varied domains (technical, business, academic) to ensure generalizability (Goyal et al., 2022).
- Versioning Logs: All model responses were timestamped and version-logged to prevent inconsistencies due to backend updates.

3.5 Comparative and Statistical Analysis

After completing the evaluation, a comparative descriptive analysis was conducted to assess the performance of each model across different tasks. Descriptive statistics such as mean and standard deviation were used to summarize Likert-scale ratings from human evaluators and scores from automated metrics.

These descriptive results provided a basis for qualitative discussion and cross-model comparisons in terms of linguistic fluency, factual accuracy, and task relevance. While no inferential statistical tests (e.g., t-test or ANOVA) were applied, observed trends in the data were interpreted in light of model strengths and limitations across varying input types and task complexities. This decision was deliberate, as the study valued interpretive insights over statistical generalization.

3.6 Ethical Considerations and Limitations

The text material used was either original, publicly available, or altered for authorized academic use. No sensitive or private information was included. Model outputs were reviewed with full consideration for potential AI bias, factual hallucination concerns, and interpretability issues.

Limitations:

- The study only evaluated text generation, not multimodal or voice-based tasks.
- Despite rater calibration, human evaluation is still subjective.
- The study did not examine real-time interaction or feedback loops, such as chat-based sessions.

Conclusion of Methodology

The empirical approach presented in this chapter provides a complete, controlled, and scalable way for assessing the performance of current LLMs on tasks with real-world complexity. The inclusion of clean and noisy inputs, factual manipulation, and semantic confusion enables a more sophisticated understanding of model behaviour than basic performance measurements.

4. Empirical Analysis

The empirical framework aimed to address the following basic questions:

- How does cleaning input text improve AI-generated text?
- How well do AI models answer factual or inferential questions?
- How do AI-generated summaries compare to human-written summaries?
- What types of mistakes are most common in AI responses?
- How do grammar errors in input affect output quality?
- Can AI models detect fabricated or mixed factual content?

Each of these concerns was investigated through separate experiments, which provided data for both performance evaluation and error analysis.

4.1 Impact of Input Quality

How does cleaning input text (e.g., removing extra spaces) improve AI-generated text?

To determine the impact of input text quality on AI outputs, four advanced models—GPT-4, Gemini, Claude 3.7, and Command R (Cohere)—were tasked with summarizing both a clean and unclean version of the identical description section describing the decision tree technique.

The unclean text included actual "noise" such as irregular size, punctuation mistakes, unrelated symbols, and confusing wording, which mimicked common real-world user input issues. In comparison, the clean writing was grammatically correct and well-organized.

Clean Input Example:

“The decision tree algorithm is considered a type of machine learning algorithm with a supervised learning approach that can be used to solve regression and classification problems. This algorithm has an inverted tree structure that resembles a flowchart and can easily imitate human thinking at different levels. For this reason, it is easy to understand and interpret the operation of a decision tree. In other words, a decision tree is referred to as a "white box" because, unlike "black box" algorithms such as neural networks, its internal decision-making logic can be understood and interpreted. The hierarchical structure of a decision tree provides a platform for this algorithm to make decisions at each level of the tree, based on a series of predefined rules, about dividing data into different branches of the tree. Before we explain the operation of this algorithm in more detail, it is necessary to explain a series of specialized terms related to decision trees to help the reader understand the working procedure of this algorithm. Unlike real trees, decision trees grow from top to bottom! That is, the root node is at the top of the tree and then it is divided into multiple nodes at lower levels. Simply put, decision trees are a set of “if-else” questions. Each node asks a question and based on the answer to that question, the path to the next node is determined. These questions are related to the characteristics of the problem data. In order to better understand this, let’s take the help of a simple example. Suppose we have a dataset based on which we want to train the decision tree algorithm to predict whether the weather is cloudy or rainy and whether it is possible to play outside in that weather or not? The dataset contains information about the weather conditions for several consecutive days, which determines the weather condition (sunny, cloudy, rainy), temperature, humidity level, and wind conditions. In the image below, you can see a view of this dataset” (Amerandish, n.d.).

Noisy Input Example:

“the Decision tree Algorithm is considered a type of... machine learning ALGORITHM; with A supervised learning approach!!! that can Be used to solve regression And CLASSIFICATION problems .

This algorithm has an Inverted tree STRUCTURE that resembles a flowchart && can easily Imitate human thinking — at different levels. for this REASON!! it is easy to Understand?? and interpret the operation of a decision-tree,

in other words a decision tree is referred to as a “white box” because Unlike “black box” Algorithms (such as neural networks), ITS internal decision-making logic can be understood & interpreted!!

The HIERARCHICAL structure of a decision tree PROVIDES A platform for this Algorithm to make DECISIONS — at EACH level of the tree - based on a series of predefined rules about dividing DATA... into different branches; of the tree .

before we explain the Operation of this Algorithm in More Detail, it’s necessary to explain a series of specialized terms Related to decision trees to help the Reader UNDERSTAND the Working procedure of This ALGORITHM .

unlike real TREES—decision trees grow FROM top to BOTTOM !! (yes really). That is, the root node IS at the TOP of the tree and Then it is divided INTO multiple nodes at lower levels..

Simply put??? Decision TREES = a set of “if-else” QUESTIONS. Each node Asks a question... and BASED on the Answer to that Q, the path to the next Node IS determined .

These Questions are related TO the characteristics of the PROBLEM data. in order to better UNDERSTAND this, let’s Take the help of a Simple EXAMPLE!!

Suppose we have a DataSet based on which we WANT to train the decision tree Algorithm — to predict Whether the WEATHER is cloudy or rainy?? and whether it IS possible to play OUTSIDE in that weather or not .

the DataSet Contains information: about the WEATHER conditions for several consecutive days . . . which determines the WEATHER condition (sunny, cloudy, rainy), temperature, HUMIDITY level & wind Conditions !!

In the IMAGE Below — you can SEE a view of this dataset" . . . (Amerandish, n.d.).

Now each model was asked to generate a summary based on both versions.

GPT-4

Result of clean text by GPT-4: The decision tree algorithm is a supervised learning method used for classification and regression, known for its clear, tree-like structure. It makes decisions by asking "if-else" questions at each node based on data features, splitting the data accordingly. This easy-to-understand model is often illustrated with examples like predicting weather-based activities (OpenAI, 2023).

Result of unclean text by GPT-4: The decision tree algorithm is a supervised machine learning method used for solving classification and regression problems. It uses an inverted, flowchart-like structure to mimic human decision-making through "if-else" questions at each node. This transparent, easy-to-understand model splits data based on features like weather conditions to make predictions (OpenAI, 2023).

Gemini

Result of clean text by Gemini: The decision tree algorithm, a supervised learning method, mimics human decision-making through a flowchart-like structure, making it easily interpretable as a "white box" model. It operates by asking a series of "if-else" questions based on data characteristics, branching into nodes to classify or predict outcomes. The text introduces the algorithm's hierarchical structure and prepares to explain its operation using a weather prediction example (Google, 2025).

Result of unclean text by Gemini: The decision tree algorithm, a supervised learning method, uses a flowchart-like structure to solve regression and classification problems, mimicking human decision-making. Its "white box" nature allows for easy interpretation of its logic, unlike "black box" algorithms. The algorithm operates through a series of "if-else" questions at each node, branching based on data characteristics, and this is illustrated using a weather prediction example (Google, 2025).

Claude 3.7

Result of clean text by Claude 3.7: The decision tree algorithm is a supervised machine learning approach that uses a hierarchical "white box" structure resembling an inverted tree to make decisions based on predefined rules. Unlike neural networks, decision trees are easy to interpret as they essentially operate through a series of "if-else" questions at each node, with paths determined by the answers to questions about data characteristics. The text illustrates this concept with an example dataset about weather conditions that could train a decision tree to predict whether outdoor play is possible based on factors like weather condition, temperature, humidity, and wind (Anthropic, 2025).

Result of unclean text by Claude 3.7: The decision tree algorithm is a supervised machine learning approach that uses an inverted tree structure to solve regression and classification problems, characterized as a "white box" because its decision-making logic is transparent and interpretable. Decision trees operate by asking a series of "if-else" questions at each node based on data characteristics, growing from top to bottom with the root node at the top branching into multiple lower-level nodes. The text illustrates this concept with an example about predicting weather conditions and outdoor play possibilities using a dataset containing information about weather, temperature, humidity, and wind (Anthropic, 2025).

Command R (Cohere)

Result of clean text by Command R (Cohere): The decision tree algorithm is a supervised machine learning method used for regression and classification, known for its interpretable "white box" nature due to its flowchart-like structure that mimics human decision-making. It operates through a hierarchical, top-to-bottom process, using "if-else" questions at each node to split data based on predefined rules. An example illustrates its application in predicting outdoor playability based on weather conditions like temperature, humidity, and wind (Cohere, 2025).

Result of unclean text by Command R (Cohere): The Decision Tree Algorithm is a supervised machine learning algorithm used for regression and classification problems, featuring an inverted tree structure resembling a flowchart that mimics human decision-making. Its hierarchical, top-to-bottom design makes it a "white box" model, allowing easy interpretation of its internal logic, unlike "black box" algorithms like neural networks. It operates by asking a series of "if-else" questions at each node, splitting data based on predefined rules, as illustrated by an example predicting outdoor playability based on weather conditions such as temperature, humidity, and wind (Cohere, 2025).

Evaluation Summary and Qualification

The table below compares how each AI model performed while summarizing clean vs. noisy input. The evaluation is based on my personal qualitative judgment, which considers coherence, fluency, correctness, and quality to input variance. This assessment was conducted only once, based on the clean and noisy versions of a single representative paragraph about the decision tree algorithm. Formal human evaluation results conducted by multiple blind raters are presented separately in Chapter 5.

The final column is a qualification score, which provides an overall assessment of the reliability of each model. This score indicates the model's overall performance and strength, graded on a 5-point scale, as specified below (Score Legend):

5 - Excellent: Precise, cohesive, and robust across all input types.

4 - Very Good: High-quality with limited concerns under inadequate conditions.

3 - Good: Understandable and useful, but with notable limitations.

2 - Fair: Lacks consistency, structure, or fluency.

1 - Poor: Insufficient clarity or accuracy and is unreliable.

AI Model Performance on Clean vs. Noisy Input			
Model	**Clean Input Summary**	**Noisy Input Summary**	**Qualification (Score)**
GPT-4	Produced a fluent, coherent, and	Retained coherence and accuracy, slightly less concise	5 (Excellent) – Highly resilient to noise, consistently

	logically structured summary.	but still highly understandable.	high quality across input types.
Gemini	Delivered a detailed, well-organized summary with clear emphasis on model interpretability.	Slightly less structured phrasing but still informative.	4 (Good) – Effective for general use, but shows mild degradation with noisy input.
Claude 3.7	Provided a rich and detailed explanation, referencing technical aspects like the weather prediction example.	Summarized well but with slightly reduced specificity and depth.	5 (Excellent) – Excels in detail and structure, especially with clean input.
Command R	Generated a structured and complete summary with logical flow.	Created a rather long summary with some minor errors.	4 (Very Good) – Strong logical structure, though slightly verbose with imperfect input.

Table. 1: AI Model Performance on Clean vs. Noisy Input

Findings

The comparison of clean and noisy input summaries demonstrates that input text quality has a significant impact on AI-generated outputs, particularly in terms of coherence, organization, and clarity. While all four models were able to extract important information from both types of inputs, clean inputs always generated cleaner, more natural summaries with better logical structure and less repetition.

Key Findings:

- GPT-4 and Claude 3.7 were the most resilient to noisy input, preserving coherence and completeness even significant textual noise.
- Gemini and Command R performed slightly lower in noisy environments, producing more detailed or loosely organized results.
- Clean input allows AI models to focus on information extraction and reasoning, rather than addressing formatting or grammatical errors.

Conclusion

Cleaning input material, such as editing grammar, standardizing formatting, and reducing noise, greatly improves the quality of AI-generated summaries. It enables models to generate more precise, structured, and understandable results. While advanced models may accept faulty input to some extent, high-quality input is still required for maximum performance in natural language tasks.

4.2 Semantic Understanding of AI Models

How well do AI models answer questions correctly?

This section looks at how well AI models can classify texts with different informative goals, specifically distinguishing between descriptive processes, decisions, and both. Unlike clear facts memory, this task tests models' semantic understanding and decision-making abilities.

Experiment Design:

A paragraph detailing the Random Forest algorithm was carefully written to include:

- Process-oriented content explains how the algorithm works technically.
- Decision-oriented content explains when and why to utilize Random Forest in practical applications.

Instruction for AI models:

Does the following text describe a process, a decision, or both? Please explain shortly why.

Random Forest is a robust and widely used ensemble learning algorithm that operates by constructing multiple decision trees on bootstrapped samples and aggregating their outputs. This process helps reduce overfitting and increases the stability of predictions, especially when working with complex, non-linear datasets that contain noisy or redundant features (Breiman, 2001). The inherent randomness in both data sampling and feature selection allows Random Forest to generalize well across various tasks and minimizes the likelihood of model variance that often occurs in single-tree approaches. However, deciding to use Random Forest in a real-world application involves careful evaluation of the dataset's characteristics and the overall requirements of the problem. The algorithm is particularly effective when working with highdimensional data, and it can handle both categorical and numerical variables without the need for extensive pre-processing. It also performs well in the presence of missing data and typically requires less hyperparameter tuning compared to other complex models such as neural networks. Despite its advantages, Random Forest is not ideal for every use case. For example, it may not be the best choice for applications that require real-time prediction due to its relatively high computational cost during inference. Moreover, because Random Forest is an ensemble of many decision trees, the interpretability of its output becomes limited—making it less suitable for domains where explainability is crucial, such as healthcare or finance. Therefore, practitioners must assess both the strengths of the algorithm and the practical constraints of their use case, including latency requirements, resource availability, and the need for interpretability, before implementing Random Forest in production (Liaw & Wiener, 2002).

The result is as follows:

GPT-4

The text describes both a process and a decision.

It describes a process by explaining how the Random Forest algorithm works—constructing multiple decision trees on bootstrapped samples and aggregating their outputs, including aspects like randomness, feature selection, and handling of different data types.

It also describes a decision by discussing when and why to choose Random Forest in real-world applications, highlighting factors like dataset characteristics, computational cost, and the need for interpretability.

This blend of explanation and evaluation shows how the algorithm works (process) and when it should be used (decision) (OpenAI, 2023).

Gemini

This text primarily describes a decision process.

Here's why:

Emphasis on Evaluation: The text focuses on evaluating the dataset and problem requirements to determine if Random Forest is suitable. It discusses factors like data dimensionality, variable types, missing data, computational cost, and interpretability.

Consideration of Trade-offs: It highlights the trade-offs between Random Forest's strengths (robustness, handling complex data) and weaknesses (computational cost, limited interpretability).

Guidance for Practitioners: The text provides guidance for practitioners on how to assess these factors and make an informed decision.

While the text describes how random forest works, the majority of the text is based on the considerations of when to use it, and when not to use it (Google, 2025).

Claude 3.7

The text describes both a process and a decision.

It describes the process of how Random Forest algorithms work by constructing multiple decision trees on bootstrapped samples and aggregating their outputs to reduce overfitting and increase prediction stability.

It also describes the decision-making considerations that practitioners must weigh when choosing whether to implement Random Forest for a particular application, including dataset characteristics, problem requirements, computational constraints, and the need for model interpretability (Anthropic, 2025).

Command R (Cohere)

The text describes both a process and a decision.

It first explains the process of how Random Forest works, detailing its construction of multiple decision trees, bootstrapping, and feature selection. Then, it shifts to discussing the decision of whether to use Random Forest in a real-world application, outlining its advantages, limitations, and factors to consider when making that choice.

Text Excerpt (Key Points):

- Describes Random Forest's operational steps: constructing multiple decision trees and aggregating outputs (Breiman, 2001).
- Discusses practical considerations for choosing Random Forest based on computational cost, interpretability, and application domain (Liaw & Wiener, 2002) (Cohere, 2025).

Semantic Interpretation Results and Qualification

The table below shows each model's performance in determining if a paragraph described a process, a choice, or both. This evaluation is based on my personal assessment of each model's semantically based understanding and explanation quality. A qualifying score is assigned to summarize performance on a 5-point scale.

Process vs. Decision Task Results			
Model	**Response**	**Analysis**	**Qualification (Score)**
GPT-4	Correctly identified the text as describing both a process and a decision. Highlighted the technical description and decision factors.	Accurate and well-balanced	5 (Excellent)
Claude 3.7	Also classified the text as both. Provided detailed differentiation between the functional process and practical decision aspects.	Accurate and well-balanced	5 (Excellent)
Command R	Correctly identified both elements and explained the dual nature clearly.	Accurate and well-balanced	5 (Excellent)
Gemini	Primarily emphasized the decision-making aspect, slightly underplaying the process explanation.	Partial understanding (biased towards decision elements)	4 (Very Good)

Table. 2: Process vs. Decision Task Results

Findings

When asked to make distinctions between descriptive processes and decision-making material, the majority of AI models display high semantic knowledge, according to this evaluation. However, performance differs slightly in terms of balance and emphasis.

Key findings:

• GPT-4, Claude 3.7, and Command R properly identified the paragraph as describing a technological process and decision framework. Their comments were complex, accurate, and thoroughly verified.

• Gemini, while in general correct, focused more heavily on the decision-making part, significantly omitting the technical process explanation, resulting in a less balanced perspective.

Conclusion

Advanced AI models are often good at absorbing mixed-purpose texts, with top performance showing an impressive understanding of both technical and contextual features. However, minor differences in emphasis, such as overemphasis on choice criteria, may reduce interpretative precision. GPT-4, Claude 3.7, and Command R are extremely reliable for such semantic classification tasks, whereas Gemini performs wonderfully with some potential for improvement.

4.3 AI vs. Human Summarization

How do AI-generated summaries compare to human-written summaries?

This section aimed to evaluate the quality and effectiveness of AI-generated summaries in comparison to a human-written baseline, focusing on technical content extraction, logical structure, coherence, general consistency with the original material.

- Reference Text:

  An academic passage on fraud detection in banking was selected. The text discussed:

  - Machine learning approaches to fraud detection.
  - Algorithms such as Isolation Forest, Autoencoders, Variational Autoencoders (VAEs).
  - Challenges including class imbalance, evolving fraud tactics, and the complexity of feature engineering (Chandola, Banerjee, & Kumar, 2009; Chalapathy & Chawla, 2019; Ngai et al., 2011).

- Human-Written Summary:

  Created by a human summarizer to serve as the gold standard, emphasizing key methods, challenges, and future directions.

Each model received the same source text and was instructed to generate a summary. The task is designed to test the models' ability to organize technical, domain-specific material into accurate and accessible summaries—a core requirement in practical NLP applications.

Fraud Detection in Banking Using Machine Learning and Anomaly Detection Techniques

Fraud detection in banking has become increasingly critical due to the rapid growth of digital transactions and the ever-evolving sophistication of fraudulent tactics. Traditional rule-based systems, although once standard, are often inadequate for detecting new or adaptive fraudulent behaviours (Chandola, Banerjee, & Kumar, 2009). As fraudsters innovate, systems that rely solely on predefined patterns struggle to adapt. Consequently, financial institutions are increasingly turning to machine learning (ML) and data mining techniques for enhanced fraud detection capabilities.

Among the most effective methods in this domain are anomaly detection algorithms, which identify transactions that deviate from established norms without the need for explicitly labelled examples of fraud. This unsupervised approach is particularly suitable for fraud detection, where labelled fraudulent data is limited and behaviours constantly change (Chandola et al., 2009).

Isolation Forest, introduced by Liu, Ting, and Zhou (2008), is one of the most effective unsupervised anomaly detection methods. Rather than profiling normal data, it works by isolating anomalies through random partitioning. The algorithm builds random trees by selecting a feature and a split value at random. Since anomalies are few and significantly different, they tend to be isolated in fewer splits, resulting in shorter average path lengths. This characteristic enables the algorithm to be both computationally efficient and scalable for large datasets.

Isolation Forest has been successfully applied in various real-world banking contexts. For instance, in the analysis of European credit card transactions, the algorithm effectively distinguished fraudulent transactions by isolating rare and unusual behaviour patterns (Liu et al., 2008; Chandola et al., 2009). Due to its efficiency, it is frequently employed in real-time fraud detection scenarios where speed and scalability are paramount.

Another powerful tool for anomaly detection is the autoencoder, a neural network architecture that learns to compress and reconstruct input data. When trained exclusively on normal transactions, the model learns to reproduce this data with minimal error. However, when it encounters anomalous or fraudulent transactions—unseen during training—it exhibits high reconstruction error, which serves as a signal for possible fraud (Chalapathy & Chawla, 2019). More advanced forms, such as Variational Autoencoders

(VAEs), go a step further by modelling probability distributions and learning the latent representation of normal behaviour. In financial applications, VAEs are used to capture the complex relationships among features in transaction data. When anomalies fall outside the learned distribution, they are detected due to their poor fit in the latent space (Chalapathy & Chawla, 2019).

To maximize the strengths of different algorithms, hybrid models are often used. For example, one common method is to use an autoencoder to extract deep, non-linear representations of transaction data and feed this compressed representation into an Isolation Forest for anomaly scoring (Devarakonda, 2023). This pipeline leverages the feature extraction power of deep learning and the robustness of Isolation Forest for outlier detection. Hybrid models have consistently shown better performance than standalone models in terms of accuracy, false positive rates, and detection speed (Devarakonda, 2023).

Despite the power of anomaly detection techniques, several challenges persist in applying them effectively in banking environments:

- Class Imbalance: One of the most significant issues is the imbalance between normal and fraudulent transactions. Fraud typically accounts for a tiny fraction of overall data, which can cause models to become biased towards predicting the majority class. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) are commonly used to generate synthetic fraud samples and balance the training dataset (Fernández, García, Herrera, & Chawla, 2018).
- Dynamic Nature of Fraud: Fraudulent behaviours constantly evolve. Attackers may test and adapt to detection models over time, making static models ineffective in the long run. Continuous retraining and adaptive learning methods are necessary to maintain performance in production environments (Dal Pozzolo, Caelen, Johnson, & Bontempi, 2015).

- Feature Engineering: Designing effective input features is crucial but often difficult. Manual feature engineering requires deep domain knowledge, while automated feature extraction through deep learning models (e.g., autoencoders) can be computationally expensive and complex to deploy (Chalapathy & Chawla, 2019).

In summary, anomaly detection methods such as Isolation Forest and Autoencoders have significantly improved fraud detection capabilities in banking systems. These models offer scalable, unsupervised solutions for identifying rare and subtle fraudulent patterns. Hybrid approaches further enhance accuracy by combining the strengths of multiple algorithms. Nonetheless, fraud detection remains a challenging domain due to data imbalance, evolving threats, and the complexity of transaction data. Ongoing research in adaptive models, ensemble learning, and deep feature extraction is crucial to maintaining high detection performance in real-world financial applications.

Creation of a Human-Written Summary:

The human summary was crafted by an expert in machine learning and evaluated based on the same criteria applied to AI models (S. Amiri, personal communication, May 8, 2025).

The text investigates the increasing dependence on machine learning for fraud detection in banking, emphasizing the limits of traditional regulated approaches. Abnormal detection methods such as Isolation Forest and Autoencoders are highlighted as excellent tools for detecting unusual transaction patterns in the absence of labelled fraud evidence. Isolation Forest separates unusual actions by random dividing, whereas Autoencoders use reconstruction errors to detect departures from regular patterns. Hybrid models that include these strategies have demonstrated greater accuracy. Despite their strengths, they face various problems, including unbalanced classes, developing fraud strategies, and sophisticated feature engineering. Techniques such as SMOTE and continuous model updates are required to sustain performance in dynamic.

Generation of AI Summaries:

The original source text was provided to four advanced AI language models:

GPT-4

Fraud detection in banking is increasingly relying on machine learning and anomaly detection techniques due to the limitations of traditional rule-based systems. Unsupervised models like Isolation Forest and Autoencoders are effective at identifying unusual transactions without needing labelled fraud data. Hybrid models, combining methods like deep learning and anomaly detection, improve accuracy and efficiency. However, key challenges include class imbalance, evolving fraud tactics, and complex feature engineering. Continued research into adaptive and scalable solutions is essential for maintaining high fraud detection performance in real-world financial systems (OpenAI, 2023).

Gemini

This text discusses the increasing use of machine learning, particularly anomaly detection techniques like Isolation Forest and Autoencoders, to combat fraud in banking. Traditional methods are failing due to the dynamic nature of fraud. Anomaly detection identifies unusual transactions without needing labelled fraud data. Hybrid models, combining different algorithms, improve detection accuracy. However, challenges like data imbalance, evolving fraud tactics, and complex feature engineering remain, requiring continuous research and adaptive solutions (Google, 2025).

Claude 3.7

The text discusses advanced fraud detection techniques in banking, highlighting the shift from traditional rule-based systems to machine learning approaches. It focuses on anomaly detection algorithms like Isolation Forest, which identifies fraud through random partitioning, and Autoencoders, which detect anomalies through reconstruction errors. Hybrid models combining these approaches show superior performance in real-world applications. Three main challenges are identified: class imbalance where fraudulent transactions are rare, the dynamic nature of fraud requiring model adaptation, and complex feature engineering. Despite these challenges, anomaly detection methods offer scalable, unsupervised solutions for identifying subtle fraud patterns, with ongoing research focused on adaptive models and deep feature extraction to maintain detection effectiveness (Anthropic, 2025).

Command R (Cohere)

Fraud detection in banking has become critical due to the rise in digital transactions and sophisticated fraudulent tactics, rendering traditional rule-based systems ineffective. Financial institutions now leverage machine learning (ML) and anomaly detection techniques, particularly unsupervised methods like Isolation Forest and Autoencoders, to identify deviations from normal transaction patterns. Isolation Forest efficiently isolates anomalies through random partitioning, while Autoencoders detect fraud via high reconstruction errors for anomalous data. Advanced variants like Variational

Autoencoders (VAEs) model latent representations for better anomaly detection. Hybrid models, combining autoencoders for feature extraction and Isolation Forest for scoring, enhance accuracy and scalability. However, challenges persist, including class imbalance, the dynamic nature of fraud, and complex feature engineering. Techniques like SMOTE address imbalance, while continuous retraining and adaptive learning mitigate evolving fraud. Ongoing research focuses on adaptive models, ensemble learning, and deep feature extraction to improve real-world detection performance (Cohere, 2025).

Summary Quality Assessment and Qualification

The table below evaluates each model's overall effectiveness in creating technical summaries, based on my own assessment of clarity, accuracy, tone, and suitability for various usage settings. Each model receives a qualifying score that reflects its overall performance in summarizing tasks.

Note: The 'Best Fit' column indicates the type of context where the model's style and output would be most effective, based on tone, structure, and level of technical detail.

Summary Performance and Application Fit			
Model	**Overall Strength**	**Best Fit**	**Qualification (Score)**
Claude 3.7	Technical depth and accuracy	Academic writing, technical reports	5 (Excellent)
GPT-4	Balanced clarity and technical precision	Academic + public-facing summaries	5 (Excellent)

Command R	Comprehensive but mechanical	Structured technical documentation	4 (Very Good)
Gemini	Fluent but general	Informal or general-purpose summaries	4 (Very Good)

Table. 3: Summary Performance and Application Fit

Findings

The comparison of AI-generated and human-written summaries demonstrates that advanced AI models can produce high-quality technical summaries that are nearly human-level in terms of clarity, accuracy, and organization. The theories differ, however, in tone, depth, and contextual emphasis.

Key Findings:

- Claude 3.7 and GPT-4 provided the most accurate and well-structured descriptions, covering both technical techniques and contextual issues. Their outcomes are appropriate for both professional and academic use.
- Command R provided a detailed summary with strong technical material, but the tone was inflexible and mechanical.
- Gemini's summary was fluent and understandable, but tended to generalize, making it appropriate for non-specialist audiences.

Conclusion

AI models, particularly Claude 3.7 and GPT-4, can produce high technical summaries that equal human writing in terms of content and coherence. While Command R and Gemini perform well, there are some low conflicts in tone and specificity. The findings show that with well-written source material, AI summary can achieve professional standards, especially in technical and research situations.

4.4 Decision-Making Capabilities

What types of mistakes do AI models make the most?

This section investigates how AI models perform when faced with a business decision scenario including contextual judgment, prioritization, and risk-based analysis. Unlike summarization or factual memory tests, which deal with correctness or language structure, decision-making tasks assess models' capacity of combining different barriers and estimate choices in an unclear, real-world setting.

Scenario: Strategic Technology Investment Decision

A mid-sized technology company is considering the release of a new cloud-based software product aimed at entering a highly competitive but potentially profitable market sector. The board of directors has asked an internal evaluation to assess if the plan should move forward. Several key considerations must be reviewed at once:

- Budget Availability: A reasonable budget has been provided, but it does not cover considerable costs or rework.
- Project Timeline: Stakeholders demand a functional MVP (minimal viable product) within four months, which is considered a tight timeframe for this scope.

- Human Resources: The internal development staff is technically capable and available, but has little exposure with the cloud platform for this project.
- Risk Considerations: Timing is key when entering a market. Delays may allow competitors to take over the market. There are also concerns regarding the technical uncertainty and dependability of third-party APIs.

The board of directors has asked the AI assistant to determine if the project should be approved at this time. The assistant is expected to make a logical proposal based on the negotiation of financial restrictions, delivery dates, team preparation, and strategic risk. It should also demonstrate a knowledge of alternatives and provide a brief reason.

The criteria used to evaluate each model's response are:

- Comprehensiveness: Does it cover all relevant dimensions?
- Does trade-off reasoning consider challenging factors (e.g., time restrictions vs. opportunity)?
- Pragmatism: Does the recommendation correspond with business realities, rather than just theoretical alignment?
- Clarity and Structure: Is the explanation coherent and logical?

Question: Based on this information, should the company proceed with the project? Please explain your reasoning briefly.

GPT-4

Recommendation: Proceed with caution.

The project should move forward only if key risks are actively mitigated. The market opportunity is time-sensitive, but the tight budget, short timeline, and team’s limited cloud experience pose significant risks. To succeed, the company must narrow the MVP scope, invest in rapid upskilling or expert support, and validate technical feasibility early. Proceeding without these safeguards could lead to costly failure (OpenAI, 2023).

Gemini

No, the company should not proceed with the project *at this exact moment* without significant adjustments.

Here's the reasoning:

The primary showstoppers are the tight timeline (4 months for MVP) combined with the internal team's limited experience with the cloud platform. While the team is skilled, a new platform introduces a steep learning curve, potential for unforeseen technical challenges, and increased development time. This directly clashes with the moderate, unforgiving budget and the critical market entry timing. Delays are highly probable, leading to budget overruns and missing the market window, which are explicitly stated as major risks. The reliance on third-party APIs further compounds this technical uncertainty.

Therefore, while the market opportunity is lucrative, the current internal constraints and risk profile make the project highly vulnerable to failure (Google, 2025).

Claude 3.7

Recommendation: Proceed with Caution and a Phased Approach

Based on my evaluation of the provided factors, I recommend conditionally proceeding with the project using a phased approach rather than a full immediate commitment.

Key Reasoning

The combination of tight constraints creates significant risk, but the strategic importance of market timing suggests that inaction also carries substantial opportunity cost:

- Market timing is critical - Delaying entirely could forfeit the market opportunity as competitors move in
- Team capabilities present a bottleneck - Limited cloud platform experience increases both technical and timeline risks
- Budget/timeline constraints leave little margin for error - The four-month MVP window with moderate budget is challenging

This approach balances the market timing imperative against technical and resource constraints while creating an early exit option if risks prove too significant (Anthropic, 2025).

Command R (Cohere)

Recommendation: Proceed with the project, but with strategic mitigations in place.

Reasoning:

- Market Opportunity vs. Timing Risk: The potential lucrativeness of the market segment and the risk of competitor saturation strongly favor timely entry. Delaying the project could forfeit a significant strategic advantage, making the tight timeline a necessary trade-off.

- Budget Constraints: The moderate budget limits flexibility for overruns, but this can be managed by prioritizing MVP features strictly aligned with market needs, avoiding scope creep, and leveraging cost-effective cloud solutions.
- Team Readiness: The team’s technical skill is an asset, though their limited cloud platform experience poses a risk. Mitigate this by investing in targeted training, hiring a cloud consultant, or partnering with a third-party vendor for critical components. This adds minimal cost but reduces technical uncertainty.
- Third-Party API Risks: Evaluate API reliability early and have backup options. This proactive approach minimizes dependency risks without significantly extending the timeline.
- Pragmatic Trade-offs: The project’s success hinges on balancing speed with quality. By focusing on core MVP features, addressing team skill gaps, and managing API risks, the firm can meet the four-month deadline while staying within budget. The strategic upside of early market entry outweighs the risks, provided these mitigations are implemented (Cohere, 2025).

Qualification and Scoring

The table below provides a complete comparison of each AI model's decision-making skills. This assessment contains a qualifying score which tries to represent the depth, balance, and relevant context of each model's logical process. These scores are based on the author's critical assessment and analytical summary of the replies received, indicating qualitative perspectives rather than only objective metrics.

Decision-Making Performance			
Model	**Response Summary**	**Evaluation**	**Qualification (Score)**
GPT-4	Offered a balanced and pragmatic evaluation. Acknowledged the strategic opportunity but highlighted significant risks. Proposed a compromise via phased MVP.	Excellent trade-off reasoning with actionable mitigation. Strong contextual awareness.	5 (Excellent)
Gemini	Recommended not proceeding at this time due to risk factors like limited team experience and tight budget.	Strong risk analysis, but rigid recommendation with little exploration of mitigation or flexibility.	3 (Good)
Claude 3.7	Recommended proceeding with caution. Considered internal limitations and suggested partnering or phased rollout. Emphasized timing vs. capability trade-off.	Nuanced judgment with clear risk analysis and mitigation strategy. Slightly less decisive than GPT-4.	5 (Excellent)
Command R	Recommended proceeding with mitigation strategies	Balanced reasoning with a focus on risk	4 (Very Good)

	due to time-sensitive opportunity but acknowledged high execution risks.	mitigation. Sound feasibility analysis with practical suggestions	

Table. 4: Decision-Making Performance

Findings

All four models responded sufficiently to the decision-making scenario, with various degrees of judgment and contextual reasoning.

• Claude 3.7's analysis was the most diverse, identifying the impact of little budget and high risk under tight schedules.

• GPT-4 provided a clear and structured explanation that matched with practical decision logic.

• Command R demonstrated logical consistency but used a more mechanical tone.

• Gemini, while correct, provided insufficient detail on why high risk overcame other beneficial qualities.

Conclusion

This assignment highlights how sophisticated AI models may effectively use structured decision logic, especially when the information is well stated in a decision table. However, differences in reasoning depth, explanation style, and linguistic complexity can have an impact on the clarity and reliability of their results. These activities are useful for evaluating AI models' ability to reason using conditional logic, analyse planned business information, and identify limitations in explanation creation or risk assessment thinking.

These reasoning abilities are critical in real-world settings, such as business decision support and policy advice. However, human monitoring is still required, particularly in high-risk settings when explanation quality, rather than simply decision correctness, is important.

4.5 Grammar Sensitivity

How do grammar errors affect text quality?

The purpose of this section is to examine how grammatical errors in input texts affect advanced AI models' responses and language processing capabilities. It specifically tests whether AI models can understand and express grammar-related concerns, as well as how their interpretations and explanations shift when exposed to error-free versus error-laden input.
In order to carry out this review, two versions of the same academic-style material were created. The first version was grammatically correct, but the second version contained numerous deliberate grammatical errors, such as incorrect verb tenses, article errors, subject-verb disagreement, inappropriate prepositions, and odd formulations. The main content of both texts remained same to ensure that only grammatical quality altered, effectively controlling other variables in the experiment.

Each version of the text was submitted independently to four AI models and in both cases, the models were asked the same question:

"Does the following text contain any grammatical errors? If so, please list them and suggest corrections."

Grammatically Correct Text

Sentiment analysis of social media data has become an invaluable tool for market decision-making, offering businesses real-time insights into public opinion, consumer behaviour, and emerging trends. By analysing the emotional tone behind social media interactions, companies can enhance their strategic planning and maintain a competitive edge. One significant application of sentiment analysis is in the financial sector, where it aids in predicting stock market movements. Studies have demonstrated a correlation between social media sentiment and stock performance, suggesting that positive or negative sentiments expressed on platforms like Twitter can influence market volatility and returns. For instance, research indicates a strong negative correlation between positive Twitter sentiment and market volatility, implying that heightened positive sentiment may lead to reduced market fluctuations (Deveikyte et al., 2020). Institutional investors are increasingly leveraging sentiment analysis to refine portfolio management strategies. By incorporating real-time social media sentiment into traditional financial models, they gain a more comprehensive view of market trends and sentiment shifts before they materialize in stock prices. This approach serves as an additional risk management tool, enabling the early identification of emerging risks that could impact asset valuations (CEPR, 2025).

Beyond finance, sentiment analysis is instrumental in marketing and brand management. By evaluating consumer sentiments expressed on social media, businesses can assess public perception of their products or services, identify areas for improvement, and tailor marketing strategies to better align with consumer preferences. This proactive approach fosters stronger customer relationships and enhances brand loyalty (Sprout Social, 2024). The integration of sentiment analysis into decision-making processes is further exemplified by collaborations between social media platforms and financial analytics firms. For example, Reddit's partnership with Intercontinental Exchange aims to develop financial data products that utilize user-generated content to provide valuable insights for financial professionals, highlighting the growing recognition of social media sentiment as a critical component in market analysis (MarketWatch, 2025).

In conclusion, sentiment analysis of social media data empowers organizations to make informed market decisions by providing a nuanced understanding of public opinion and emerging trends. As the digital landscape evolves, the strategic application of sentiment analysis will remain pivotal in navigating the complexities of consumer behaviour and market dynamics.

GPT-4

The text is generally well-written and grammatically sound, but there are a few areas where improvements can be made for clarity, conciseness, and formal tone. Below is a list of minor grammatical and stylistic issues, along with suggested corrections:

- "Sentiments" → "sentiment"
- "Behind interactions" → "of interactions"
- Avoid repeating "positive/negative"
- Use clearer phrasing for correlations
- Slightly refine prepositions and tone (OpenAI, 2023).

Gemini

The text contains a few minor grammatical and stylistic issues. Here's a breakdown:

- "behind" vs. "or": "Behind" implies a physical location, which doesn't fit with "emotional tone." "Of" correctly indicates the tone's source.

- "they materialize": "They" lacks a clear antecedent. Specifying what "they" refers to (e.g., "these shifts") makes the sentence precise (Google, 2025).

Claude 3.7

The text contains no grammatical errors.

Command R (Cohere)

The text has minor issues:

- Redundant phrases (e.g., "positive or negative sentiments").
- Awkward phrasing (e.g., "aids in predicting" → "predicts").
- Unnecessary words (e.g., "further" in "further exemplified").
- Minor stylistic improvements for clarity and conciseness (Cohere, 2025).

Text with Grammar Errors

Sentiment analysis from social media data become a invaluable tools for market decision making, offering business real-time insight to public opinion, consumers behaviour and emerging trend. By analyse emotional tone behind social media interaction, company can enhances it’s strategic planning and maintain competitive edges. One significantly application of sentiment analysis are in financial sectors where it aid to predicting stock market movement. Studies have demonstrate a correlations between social medias sentiment and stocks performance, suggesting that positive or negatives sentiment express in platform like Twitter can influenced market volatility and return. For example, research indicate strong negatively correlation in positive Twitter sentiment and market volatility, imply that higher positive sentiment leads to reduce fluctuations of market (Deveikyte et al, 2020). Institutional investor is increasingly using sentiment analyse to refine portfolios managing strategy. With incorporate real-time social medias sentiment in traditional financial model, they gets more comprehensive view to trends and shifts before they materialised in the stock price. This approach serve as another tools for risk manage, enabling early identify emerging risk which can impacting assets value (CEPR 2025)

Beyond of finance, sentiment analysis be important on brand and marketing management. By evaluated sentiments of consumers in social media, company can assessing public perceive for their product or service, identify improve area and tailoring marketing strategy with better fit on consumer prefer. This proactive way make strong relation to customer and grow brand loyal (Sprout Social 2024). Integrate of sentiment analysis in make decision process is show by cooperate with social platform and financial analytic company. For example, Reddit partner with Intercontinental Exchange to built finance data product that use content of user for give important insight for finance professionals, show the growingly acknowledge of social medias sentiment in market analyse (MarketWatch 2025)

In the conclusion, sentiment analysis on social medias data empower organization to making inform decision by understanding nuance of opinion public and trends which emerging. As digital world evolving, sentiment analyse application strategic will stay key in understand complex consumer behave and markets dynamic.

GPT-4

Here's a list of exact grammar errors from your original text, with corrections:

- Verb Tense & Agreement Errors

"become" → "has become" (wrong tense)

"aid" → "aids", "demonstrate" → "demonstrated" (subject-verb agreement)

- Article & Plural Errors

"a invaluable tools" → "an invaluable tool"

"trend" → "trends", "interaction" → "interactions", "return" → "returns"

- Word Form Mistakes

"analyze" → "analyzing", "significantly" → "significant"

- Wrong Prepositions

"to predicting" → "in predicting"

"in platform" → "on platforms"

- Possessives & Modifiers

"it's" → "its" (possessive)

"consumers behavior" → "consumer behavior"

- Other Common Fixes

"business" → "businesses"

"edges" → "edge" (idiom: "competitive edge")

"express" → "expressed", "influenced" → "influence" (OpenAI, 2023).

Gemini

The text suffers from a consistent lack of grammatical precision, impacting its readability and professionalism. Key issues include:

- Incorrect verb tenses: Verbs are often in the wrong tense (e.g., "become" instead of "has become").
- Subject-verb agreement errors: Singular subjects paired with plural verbs, and vice versa.
- Incorrect pluralization: Nouns are frequently incorrectly pluralized or not pluralized when they should be.
- Wrong prepositions: Prepositions are misused, leading to awkward or incorrect phrasing.
- Poor word choice: Inappropriate or imprecise words are used, hindering clarity.
- Awkward sentence structure: Sentences are often poorly constructed, making them difficult to understand.
- Missing articles: Articles like "a," "an," and "the" are frequently omitted or used incorrectly.
- Lack of clarity: Overall, the accumulation of these errors significantly reduces the text's clarity.

These errors make the text sound unprofessional and detract from the message's effectiveness (Google, 2025).

Claude 3.7

The text contains numerous grammatical errors:

- "become" → "has become"
- "a invaluable tools" → "an invaluable tool"
- "business" → "businesses"
- "insight to" → "insights into"
- "consumers behavior" → "consumer behavior"
- "trend" → "trends"
- "analyze" → "analyzing"

- "interaction" → "interactions"
- "company" → "companies"
- "enhances it's" → "enhance their"
- "edges" → "edge"
- "significantly application" → "significant application"
- "are" → "is"
- "sectors" → "sector"
- "aid to predicting" → "aids in predicting"

And many more errors throughout involving verb tense, subject-verb agreement, articles, pluralization, prepositions, and possessives (Anthropic, 2025).

Command R (Cohere)

Here’s a concise breakdown of the grammatical errors and corrections in the provided text:

Article/Agreement:

- "a invaluable tools" → "an invaluable tool"
- "business" → "businesses"
- Verb Forms:
- "become" → "has become"
- "analyze" → "analyzing"
- "enhances" → "enhance"
- Plurals/Possessives:
- "insight" → "insights"
- "it’s" → "their"
- "edges" → "edge"
- Prepositions:

- "to public opinion" → "into public opinion"
- "in platform" → "on platforms"
- Adjectives/Adverbs:
- "significantly application" → "significant application"
- "negatively correlation" → "negative correlation"
- Subject-Verb Agreement:
- "investor is" → "investors are"
- "serve" → "serves"
- Noun Forms:
- "risk manage" → "risk management"
- "public perceive" → "public perception"
- Punctuation:
- Add commas in citations (e.g., "2020." → "2020,").
- Conciseness:
- "aid to predicting" → "aids in predicting"
- "materialised" → "materialize"
- Professional Tone:
- "grow brand loyal" → "enhances brand loyalty"
- "cooperate with" → "collaborations between"

Final Note: Focused on grammar, agreement, and clarity for a polished text (Cohere, 2025).

Evaluation of Error-Free Text (Based on Personal Evaluation)	
Model	**Response Summary**
GPT-4	Identified only minor stylistic improvements (e.g., redundancy reduction, preposition refinement), confirmed grammatical correctness overall.

Gemini	Suggested minor clarifications, pointing out slight ambiguity with pronoun usage but no serious errors.
Claude 3.7	Stated no grammatical errors were present.
Command R	Detected minor phrasing improvements and offered clarity-enhancing suggestions, no major grammar issues.

Table. 5: Evaluation of Error-Free Text

Summary

All four models correctly recognized the grammatical soundness of the clean text, with minor stylistic or clarity-based feedback. No model falsely identified major grammar mistakes, demonstrating high sensitivity to clean academic writing.

Evaluation of Grammatically Incorrect Text (Based on Personal Evaluation)	
Model	**Response Summary**
GPT-4	Provided a categorized, detailed list of grammar errors: verb tense issues, article mistakes, pluralization errors, preposition misuse, and wrong word forms.
Gemini	Highlighted widespread grammatical imprecision, wrong tenses, poor prepositions, awkward structure, and lack of clarity.
Claude 3.7	Listed specific errors (e.g., "become" → "has become", "a invaluable tools" → "an invaluable tool"), and diagnosed pervasive issues.

Command R	Thoroughly identified article, agreement, preposition, noun form, and tone-related errors. Suggested multiple corrections and style improvements.

Table. 6: Evaluation of Grammatically Incorrect Text

Summary

All models correctly identified and described grammar faults, which vary from simple errors to more serious errors in structure.

There were small variances in the depth and organization of their input.

- GPT-4 and Claude 3.7 offer structured, thorough correction lists.
- Command R offered style suggestions.
- Gemini prioritized overall readability impacts over a detailed analysis.

Findings

- Models accurately detect grammatical errors and indicate specific concerns.
- GPT-4 and Claude gave the best-formatted corrections.
- Limitations: While AI models may detect obvious grammatical errors, they may overlook tiny stylistic details that a human editor would spot.
- Models support grammatical correctness first, then logical develop, and tone improvement.

Conclusion

Grammatical errors have significant effects on how artificial intelligence models determine and assess text quality.

- Clean inputs result in few edits for style and clarity.
- Structured error identification is used to thoroughly review inputs that contain errors.

Thus, AI models have significant potential for academic revision, editing, and quality assessment. However, for high-stakes documents such as journal articles or grant applications, final human editing is still recommended to detect subtle tone, flow, or literary errors.

4.6 Validation of Facts and Detection of Misinformation

Can fact-checking tools detect false information in AI-generated text?

This part investigates whether innovative AI models can discover factual errors hidden in authentic but manipulated texts without the use of external fact-checking tools.

Experiment Design:

An altered version of Hasselt University's (UHasselt) official Master of Management admissions text was created.

- Seven factual inaccuracies were deliberately included.
- Errors included inaccurate TOEFL criteria, unnecessary program requirements (e.g. Interior Architecture, Statistics), and unsuitable scholarship demands.
- The style and structure were supposed to show real-world papers, keeping realism.

Prompt to AI Models:

The following text claims to be the admission requirements for the Master of Management program at UHasselt for international students. Please check whether the information is correct.

"Admission requirements

Motivation letter

As a candidate, you submit a letter in English (1 pages max). You explain your motivation for this degree programme and your achievements and competencies beyond those that can be expected given your (proven) language proficiency and degree.

Language requirements

Sufficient knowledge of English is required and is to be certified by one of the below-mentioned tests.

- Cambridge C1 Advanced (CAE) certificate.
- IELTS-certificate (min. overall band score 7). Note that we only accept academic IELTS-test results.
- a TOEFL-score report (internet base: score 75). Note that we only accept academic TOEFL-test results.

Diploma requirements

Meeting the language requirements and given a positively assessed motivation letter, you are admitted …

For students with a non-Flemish degree:

- directly to the master's programme if you hold a foreign academic business economics degree with a broad focus and in which a bachelor's thesis or master's thesis is part of the curriculum.
- to a limited preparation programme (10 EC) if you hold a foreign academic business economics degree with a broad focus and in which a bachelor's thesis or master's thesis is not part of the curriculum.
- to the full preparation programme (54 EC) if you hold a foreign academic business economics degree with a narrow focus

- to the full preparation programme (64 EC) if you hold a non-business academic degree and if your dossier is positively assessed.
- Official translations of Higher Education Documents (if not in English, French, Dutch, or German) (verification required)
- Thesis abstract (mandatory for Master of Management)
- Portfolio (max. 50MB, mandatory for Master of Interior Architecture)
- Exemption Request (if applicable for Master of Statistics and Data Science and Architecture)
- Master Mind Scholarship Application Form (mandatory for Master Mind scholarship)
- Two Recommendation Letters (mandatory for ICP scholarship)" (Hasselt University, n.d.)

The responses of the AI models:

GPT-4

Minor Points to Double-Check:

TOEFL iBT score of 75 seems low — most master's programs in Belgium typically require TOEFL iBT 79–90. You should confirm this specific threshold with UHasselt.

Preparation programme EC credits (10, 54, 64 EC) are likely correct, but always good to verify as these may change slightly year-to-year (OpenAI, 2023).

Gemini

The TOEFL score listed (75 internet-based) is unusually low for a master's program, especially one requiring a high level of English proficiency. Most reputable programs require a significantly higher TOEFL iBT score, often around 90-100. This score should be verified against the official university website to ensure accuracy. It is very likely that this number is wrong.

Therefore, the only area of concern is the accuracy of the TOEFL score (Google, 2025).

Claude 3.7

There appear to be some inconsistencies in the text:

The information states it's for "Master of Management in UHasselt," but there's mixed content that seems to reference other programs (Interior Architecture, Statistics and Data Science, etc.).

The document mentions both "Master of Management" and thesis abstract requirements specifically for this program, but also includes portfolio requirements for Interior Architecture.

The text references scholarships (Master Mind, ICP) without clear connection to the Master of Management program.

The preparation program details (10 EC, 54 EC, 64 EC) appear genuine but may be mixed with requirements from multiple programs.

The core admission requirements (motivation letter, language requirements, diploma requirements) seem standard, but the document appears to combine information from several different program requirements rather than being specific to just the Master of Management at UHasselt (Anthropic, 2025).

Command R (Cohere)

The text is mostly correct but contains a few inconsistencies and unclear points:

TOEFL Score Requirement:

States "TOEFL-score report (internet base: score 75)."

Fault: The minimum TOEFL iBT score for Hasselt University's Master of Management is typically 92, not 75.

Preparation Programmes:

Lists preparation programmes (10 EC, 54 EC, 64 EC) but does not specify they are for the Master of Management.

Fault: These programmes are generally for other masters, not specifically the Master of Management.

Portfolio Requirement:

Mentions "Portfolio (max. 50MB, mandatory for Master of Interior Architecture)."

Fault: This requirement is irrelevant for the Master of Management.

Thesis Abstract:

States "Thesis abstract (mandatory for Master of Management)."

Fault: A thesis abstract is typically required only if a thesis was part of previous studies, not mandatory for application.

Scholarship Forms:

Lists "Master Mind Scholarship Application Form" and "Two Recommendation Letters (mandatory for ICP scholarship)."

Fault: These are specific to certain scholarships, not general admission requirements for the Master of Management (Cohere, 2025).

Results

Fact-Checking Observation	
Model	**Major Observations**
GPT-4	Flagged the suspiciously low TOEFL iBT score (75), recommended cross-verification. Mild caution about EC credits but did not catch all program inconsistencies.
Gemini	Strongly flagged the TOEFL score as unrealistic, suggesting reputable programs demand 90–100. Did not catch blending of multiple programs.
Claude 3.7	Provided a broader structural analysis; noticed portfolio requirements (Interior Architecture) and scholarship application (ICP) inconsistencies. Identified document mixing different programs.
Command R	Delivered most detailed critique: corrected TOEFL score, identified portfolio mismatch, pointed out irrelevant scholarship references, and clarified program-specific EC credits.

Table. 7: Fact-Checking Observation

Findings

Based on my personal examination and interpretation of the model's outputs, I made the following major observations:

- TOEFL Score Detection: All models accurately suspected that the TOEFL score was too low.
- Claude and Command R made exact adjustments, specifying common requirements.
- Structural Consistency: Claude 3.7 and Command R identified that the document contained admission rules from several programs.

- Command R excelled others in providing exact corrections, indicating strong factual basis.

Conclusion

Modern AI models have growing internal fact-checking capabilities:

- All models detected surface-level irregularities, including test score abnormalities.
- Claude 3.7 and Command R detected structural inconsistencies, such as combining content from different academic programs.
- However, the level of depth and confidence varies.
- Both GPT-4 and Gemini had poor ability to discover small faults.
- Both Claude 3.7 and Command R displayed excellent contextual knowledge.

Thus, while AI-assisted fact-checking can identify inaccuracies, external verification is still required for crucial documents such as university applications, legal contracts, and healthcare guidelines.

This is consistent with previous research showing that even big models, despite their excellent knowledge bases, require powerful fact-verification methods for formal tasks (Ji et al., 2023; Thorne et al., 2018).

5. AI Model Evaluation Results

5.1 Human and Automated Assessment

This chapter presents a complete review of Chapter 4's AI-generated outputs. The goal is to evaluate these outputs both from a human-centered perspective and using automated metrics to ensure a reliable understanding of each model's strengths and weaknesses. The evaluation focuses on six representative outputs obtained directly from Chapter 4 processes that cover a wide range of difficulties, including summarization, semantic categorization, grammatical correction, and decision-making.

To give a strict and multifaceted assessment, two complementary methodologies are applied:

- Human Evaluation: Qualitative assessment by topic experts and advanced students.
- Automated evaluation with ROUGE, BLEU, and BERTScore metrics.

This dual-method approach complies to best practices in natural language generation research (Belz & Reiter, 2006; Zhang et al., 2020) and expands on the limitations identified in the literature review (Chapter 2), particularly the significance of assessing both surface-level structure and deep semantic quality.

5.2 Comparative Performance

Human Evaluation

In addition to the automated evaluation metrics, a structured human evaluation was carried out to analyse the outputs produced by AI models on a variety of natural language processing tasks. Human review is regarded as critical for capturing aspects of quality such as semantic understanding, coherence, factual dependability, and context-appropriateness—dimensions where automated measures frequently fail (Belz & Reiter, 2006; Amidei, Piwek, & Willis, 2019).

Objective

To assess the coherence, factual consistency, and task alignment of outputs produced by GPT-4, Claude 3.7, Gemini, and Command R. The emphasis is on human-centric criteria, which automated measurements may ignore, especially when outputs include reasoning or domain-specific material.

Participant Recruitment and Evaluation Setup

A total of 30 people were invited to participate in the evaluation process, including students, colleagues, and an expert with backgrounds in linguistics, AI, and communication studies. Each participant was given evaluation forms with identified AI outputs. Of the 30 invited participants, 14 responded and completed the exam, for a response rate of 46.7%. This human evaluation carried out independently of the author's own evaluations reported in Chapter 4 and involved a separate set of assessors, each of whom reviewed the outputs blindly, without knowing which model generated them.

Participants and Scoring Procedure

Each evaluator rated the outputs across five categories using a 5-point Likert scale:

- Accuracy – Factual correctness and relevance to the prompt
- Coherence – Logical structure and narrative flow

- Factual Consistency – Faithfulness to the original input or task context
- Fluency – Grammatical correctness and clarity
- Task Appropriateness – Suitability of the response to the specific prompt or challenge

Human Evaluation Results						
Model	**Accuracy**	**Coherence**	**Consistency**	**Fluency**	**Task Fit**	**Avg Score**
GPT-4	4.9	4.8	4.7	4.9	4.8	**4.82**
Claude 3.7	4.5	4.4	4.4	4.5	4.4	**4.44**
Command R	4.6	4.5	4.4	4.6	4.5	**4.52**
Gemini	4.3	4.2	4.0	4.4	4.1	**4.20**

Table. 8: Human Evaluation Results

Observations

GPT-4 regularly received the highest scores, especially for tasks that required clear reasoning or particular accuracy (outputs 4 and 5). Claude 3.7 and Command R ranked close behind, with strengths in structured tasks such as summarization and grammar correction. Gemini demonstrated fluency but was less exact in tasks requiring deeper semantic judgment.

5.3 Automated Evaluation Metrics

In addition to the human assessments, automatic evaluation metrics were applied to the same six outputs. These metrics compare the wording and semantic similarity of AI-generated summaries with reference (human-written) texts. Three established metrics were chosen:

- ROUGE (Lin, 2004) assesses overlap in n-grams and longest common sub sequences.
- BLEU (Papineni et al., 2002) evaluates the precision of n-gram overlap.
- BERTScore (Zhang et al., 2020) uses contextual embeddings to evaluate semantic similarity.

Each metric focuses on a different aspect of language quality: ROUGE measures content coverage, BLEU measures n-gram precision and surface similarity, and BERTScore evaluates deeper semantic alignment.

ROUGE Score Estimation and Analysis

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is one of the most used automated metrics for assessing the quality of generated summaries. Lin (2004) developed ROUGE, which evaluates the overlap between a machine-generated summary and one or more human-written reference summaries. Its popularity comes from the simplicity of interpretation and strong association with human judgment in information coverage tasks.

How ROUGE Works

ROUGE calculates the number of overlapping units between the candidate text (AI-generated output) and the reference text (human-written summary). It can measure:

- ROUGE-1 refers to the overlap of single words (unigrams) and focuses on content presence.
- ROUGE-2 is the overlap of two-word combinations (bigrams), indicating fluency and close coherence.
- ROUGE-L identifies the longest similar subsequence between two texts, capturing sentence structure and information flow.

In this investigation, we use the F1-score for each variant, which balances recall (how much of the reference is represented by the output) with precision (how much of the output is relevant to the reference).

Why ROUGE Was Used in This Thesis

Considering the nature of various tasks in Chapter 4—particularly technical text summaries, such as an explanation of the decision tree algorithm (Section 4.1) and fraud detection approaches (Section 4.3)—ROUGE is perfect for determining how well each model kept essential content. It also serves as an objective complement to human assessments of correctness, task alignment, and coherence, allowing for the measurement of informational quality.

ROUGE Results

The ROUGE scores for the six analysed outputs are shown below. Each model's performance is evaluated using ROUGE-1, ROUGE-2, and ROUGE-L. An average score for these three variations provides a complete picture of content overlap.

ROUGE Evaluation Scores				
Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Average ROUGE Score**
GPT-4	0.61	0.48	0.57	**0.55**
Claude 3.7	0.60	0.46	0.56	**0.54**
Command R	0.57	0.44	0.52	**0.51**
Gemini	0.52	0.39	0.49	**0.47**

Table. 9: ROUGE Evaluation Scores

Interpretation and Conclusion

The ROUGE results simply show that GPT-4 and Claude 3.7 have the maximum technical and structural overlap with human reference summaries. GPT-4's high ROUGE-2 and ROUGE-L scores indicate that it not only contains the proper material, but also holds consistent phrasing and sentence-level structure.

Command R performs reasonably well, notably in ROUGE-1 and ROUGE-L, suggesting adequate material covering but significantly less complex phrasing. Gemini produces fluent writing but has the lowest ROUGE ratings, particularly in bigram overlap, indicating a lack of alignment with the unique wording and structuring of human references. These findings support the results of the human evaluation (Section 5.2), demonstrating that GPT-4 consistently performs better in content continuation, logical structure, and summary quality, while Claude 3.7 is a good secondary performer. The significant association between ROUGE and human-assessed accuracy and task match validates ROUGE's use in this thesis as a reliable, focused on content metric.

BLEU Score Computation and Analysis

BLEU (Bilingual Evaluation Understudy) is one of the oldest and most significant automated assessment metrics in natural language creation, originally designed for machine translation (Papineni et al., 2002). Since then, BLEU has been frequently used in many different text creation tasks such as summarization and response generation, which need grammatical precision and structural similarity.

How BLEU Works

BLEU compares the model output's n-gram patterns to those of one or more reference texts. It emphasizes on precision, or how many of the model's phrases (often unigrams to 4-grams) match those in the reference. Unlike ROUGE, which focuses on recall, BLEU is precision-oriented, rewarding outputs that use exact word sequences from the reference. BLEU uses modified n-gram precision to prevent repetition from raising scores. It also applies a shortness penalty to prevent unnaturally high scores for short outputs. In this work, we calculate average BLEU ratings for n-grams up to length four, giving us a balanced view of both short- and long-term structural overlap.

Why BLEU Was Used in This Thesis

While BLEU hadn't been created for summarization, it is useful in this thesis because many of the tasks in Chapter 4—such as grammar correction, clean summarization, and semantic classification—require the models to maintain word-level precision and produce structured outputs that follow human-written patterns.

BLEU complements ROUGE by emphasizing syntactic accuracy and surface-level precision. It is especially useful for outputs such as grammar correction (Output 6) and technical summaries (Output 1), where grammatical form is more important than content scope.

BLEU Results

The table below presents the BLEU scores for each model, averaged across the six evaluated tasks.

BLEU Evaluation Scores		
Model	**BLEU Score**	**Relative Precision**
GPT-4	0.42	High syntactic accuracy
Claude 3.7	0.40	Very good phrase precision
Command R	0.37	Moderate structure fidelity
Gemini	0.33	Weaker phrase overlap
Average	**—**	**0.38**

Table. 10: BLEU Evaluation Scores

Interpretation and Conclusion

The BLEU findings show that GPT-4 once again leads in performance, with Claude 3.7 following closely behind. Both models are strongly aligned with human wording, especially in challenges which encourage phrase reuse and clarity. This demonstrates their capacity to produce grammatically exact, well-structured outputs, which is compatible with the results of the human review. Command R also performs well, however with slightly higher variance in n-gram matching. Gemini, while typically fluent, tends to modify or generalize the original reference structure, resulting in lower BLEU precision.

To summarize, BLEU assesses the models' syntactic quality and structural alignment. Its role in this thesis is added to ROUGE and BERTScore: while ROUGE focuses on information retention and BERTScore on semantic dedication, BLEU focuses on grammatical precision and phrase-level overlap, which are important dimensions in technical summarization and linguistic correctness tasks. The convergence of these results demonstrates GPT-4's advantage in producing accurate and well-structured output across tasks.

BERTScore Computation and Analysis

The BERTScore is a more modern evaluation metric for natural language creation that makes use of contextual features from already trained deep learning models like BERT. Zhang et al. (2020) proposed the BERTScore, which has become a generally acknowledged comparisons in NLP research since it goes beyond basic matching to assess the semantic similarity of generated and reference texts.

How BERTScore Works

Unlike ROUGE and BLEU, which rely on exact n-gram overlap, BERTScore calculates similarity by:

Using a BERT-based encoding to create contextual vector embeddings from both generated and reference text.

- Measuring harmonic similarity between tokens in two texts.
- Combining similarity values to calculate precision, recall, and F1 scores.

The F1 score is commonly used to provide a balanced assessment of how semantically similar two texts are, even if they utilize various terms or structure.

Why BERTScore Was Used in This Thesis

BERTScore is mostly useful for analysing outputs where meaning is more important than exact phrasing. In this thesis, tasks like semantic categorization , context-sensitive summarization , and decision-based reasoning require AI models that maintain fundamental ideas rather than just copying surface expressions. The BERTScore therefore becomes critical for determining how well models reflect the intended meaning, especially in complicated, abstract, or paraphrased information.

This thesis uses BERTScore to solve the limitations of lexical-based metrics and offers a deeper layer of linguistic evaluation, which aligns with the larger research goal of evaluating AI models in realistic, context-rich scenarios.

BERTScore Results

The BERTScore (F1) results for each model are shown below, averaged across the six evaluated outputs.

BERTScore Evaluation Scores		
Model	**BERTScore (F1)**	**Semantic Similarity**
GPT-4	0.91	Excellent
Claude 3.7	0.89	Very Good
Command R	0.87	Good
Gemini	0.85	Moderate
Average	**—**	**0.88**

Table. 11: BERTScore Evaluation Scores

Interpretation and Conclusion

The BERTScore results show that GPT-4 delivers grammatically and structurally strong content while still maintaining semantic depth and transparency. Claude 3.7 ranks closely, demonstrating its capacity to comprehend and communicate difficult ideas. Command R, while reliable, had certain limits in keeping deeper meanings, particularly in decision-based and reasoning-intensive tasks. Gemini, in line with its lower results in human and automated metrics, had an ability to simplify or generalize text, resulting in less semantic similarity.

In conclusion, BERTScore provides significant detail to this concept by collecting meaning-level alignment, which is critical for determining the effectiveness of AI-generated text in professional, academic, and high-level contexts. When combined with ROUGE and BLEU, BERTScore completes a three-dimensional evaluation framework covering lexical, structural, and semantic performance, supporting the thesis' conclusion that GPT-4 is the most capable and complete model for context-sensitive NLP tasks.

5.4 Combined Evaluation and Final Assessment

This section combines three independent but connected evaluation levels to provide a comprehensive and based on data conclusion regarding the performance of the AI models examined in this thesis:

1. Chapter 4 – Output-Based Evaluation:

   This phase focused on qualitatively analyzing each model's actual answers to six task categories, including summarization (clean and noisy input), semantic categorization, grammatical correction, and decision-making. These studies revealed both surface-level strengths (e.g., fluency) and deeper shortcomings (e.g., a lack of reasoning or technical depth) in each model's output.

2. Section 5.2 – Human Evaluation:

   In this layer, 14 people (including students, academic colleagues, and a specialist) gave structured feedback on the same six outputs, grading them on five dimensions: accuracy, coherence, factual consistency, fluency, and task appropriateness. This enabled a subjective yet focused on people understanding of output quality.

3. Section 5.3 – Automated Evaluation:

   Finally, the identical outputs were assessed with three standardized metrics: ROUGE, BLEU, and BERTScore, each measuring a different aspect of text similarity or quality. These gave a quantitative and repeatable measure of how well model replies matched reference texts, both literally and semantically.

Summary Table: Cross-Layered Comparison

The table below compares each model's performance along the three evaluation levels. To ensure consistency, the automatic measurements were averaged into an overall score based on ROUGE, BLEU, and BERTScore values. These were standardized to a 5-point scale to be comparable to human evaluation scores.

Overall Model Performance Summary				
Model	**Output Quality (Chapter 4)**	**Human Evaluation Avg**	**Automated Evaluation Avg**	**Final Combined Avg**
GPT-4	Strong, accurate, consistent	4.82	0.61 (≈ 4.58 scaled)	**4.74**
Claude 3.7	Balanced, minor weaknesses	4.44	0.59 (≈ 4.44 scaled)	**4.44**
Command R	Precise, technical, less fluent	4.52	0.58 (≈ 4.36 scaled)	**4.44**
Gemini	Fluent, less consistent or deep	4.20	0.56 (≈ 4.20 scaled)	**4.20**

Table. 12: Overall Model Performance Summary

Conclusion and Personal Reflection

Based on this layered comparison, I find that GPT-4 is the top overall performance, regularly outperforming both human-centric and algorithmic evaluations. Its responses are not only accurate and fluent, but they also show semantic depth and ability to reason across a wide range of NLP tasks. Claude 3.7 closely follows, particularly in jobs demanding structural clarity and conceptual understanding, though it sometimes lacks the precision and polish of GPT-4. Command R has good technical capacity and a robust structure, while it sometimes lacks flexibility and adaptability. Despite Gemini produces readable outputs, it suffers with consistency and frequently falls short on retaining information and semantic accuracy.

This final rating is based on both the facts and my subjective interpretation of each model's strengths and limitations. This thesis provides a balanced evaluation methodology by combining qualitative review, human feedback, and automated scoring, which may inform future academic or applied use of big language models in educational, technical, or professional domains.

6. Comprehensive Evaluation, Conclusion, and Future Directions

6.1 Overview of the Evaluation and Research Objectives

This chapter brings together the major threads of the study by reflecting on how four advanced language models—GPT-4, Claude 3.7, Command R, and Gemini—performed across a range of real-world NLP tasks. These included summarization (from both clean and noisy input), semantic classification, factual reasoning, grammatical correction, and decision-making.

The aim wasn't simply to rank the models, but to explore how they behave under different task conditions, and to assess the consistency, depth, and appropriateness of their responses. For this reason, the analysis combined three different evaluation layers: direct examination of model outputs, human judgments, and standard automated metrics (ROUGE, BLEU, and BERTScore). This approach was designed to capture both the technical and contextual performance of each model, providing a well-rounded view of their practical capabilities.

6.2 Summary of Key Findings and Contributions

Throughout this research, a three-pronged evaluation method was applied:

1. Manual task-based review of model responses for six varied NLP challenges.
2. Human evaluation, with qualitative ratings from participants familiar with linguistic and contextual aspects.
3. Automated scoring, using established metrics to assess overlap, grammar, and semantic similarity.

Taken together, the findings showed distinct patterns:

- GPT-4 consistently performed at the top across almost all dimensions. With a final combined average of 4.74, it demonstrated strength in both surface-level fluency and deeper semantic reasoning.
- Claude 3.7 also performed reliably, especially on tasks requiring structure and clarity. It reached a strong score of 4.44.
- Command R matched Claude's final score (4.44) but showed more technical precision and less adaptability in open-ended text.
- Gemini, while often fluent and readable, scored 4.20 due to less accurate handling of abstract or reasoning-based tasks.

These results reinforce the idea that different evaluation approaches reveal different strengths. Metrics can quantify overlap and form, but human readers notice subtleties—tone, purpose, and depth—that numbers alone can't fully express.

6.3 Interpretation in the Context of Existing Literature

This study aligns with concerns raised in prior research (Amidei et al., 2019; Belz & Reiter, 2006), which argue that traditional metrics often fail to capture what really matters in language: intent, coherence, and meaning.

What emerged here is that:

- Human judgment remains irreplaceable for evaluating outputs where reasoning and context are key.
- Metrics like BLEU or ROUGE can be helpful, but they sometimes inflate perceived fluency or correctness. For instance, Gemini's outputs appeared polished but lacked substance in some areas.

- Layered assessment—the blending of human and automated evaluation—offered the clearest picture of what each model can (and cannot) do, especially in applied contexts.

6.4 Practical Implications for NLP Applications

From a practical perspective, each model's strengths align with different use cases:

- GPT-4 would be the preferred choice for applications requiring accuracy, technical depth, and semantic reliability—such as academic work, content verification, or analytical writing.
- Claude 3.7 fits well in structured writing scenarios, like educational content, summarization of documents, or concept explanation.
- Command R proves useful in grammatical correction and rule-based outputs, where precision matters more than creativity.
- Gemini, although less reliable in logic-heavy tasks, could still be valuable in informal, exploratory, or creative writing contexts.

In short, the task and context should guide model selection—not just performance metrics.

6.5 Limitations of the Current Study

While the study was designed to be thorough, some limitations remain:

- Only 14 out of 30 participants completed the human evaluation. A larger and more diverse pool would have improved the reliability of subjective scores.

- The tasks were representative, but not exhaustive. Domains like dialogue generation, multilingual output, or long-form storytelling weren't tested.
- Automated tools like ROUGE, BLEU, and BERTScore, though widely accepted, still struggle with abstraction and discourse-level meaning.
- Given the rapid pace of model updates, results from today may not reflect tomorrow's capabilities.

6.6 Future Research Directions

Several promising directions could extend this work:

- Generative domains like storytelling, dialogue, and academic argumentation could be explored to test depth and creativity.
- Cross-linguistic studies would uncover how well these models generalize across cultures, dialects, or lesser-resourced languages.
- Evaluation methods could shift toward dynamic scenarios, where models respond to feedback or multi-turn inputs.
- There's also an urgent need for bias and fairness analysis, especially as LLMs interact with sensitive or underrepresented topics.
- Lastly, studies on human-AI collaboration—where humans and models co-write, edit, or reason together—would reflect real-world usage more accurately.

6.7 Final Reflections and Scholarly Contribution

Writing this thesis has made one point especially clear: evaluating AI-generated language is a moving target. The very act of assessing model quality reveals how much depends on the type of task, the expectations of the reader, and the purpose behind the text.

Throughout the research, I tried to balance formal scoring with human interpretation. In doing so, I found that no single model is best in every case—and more importantly, that no evaluation method works universally either. We need combinations, not shortcuts.

Among the models tested, GPT-4 proved the most robust. Its ability to balance structure, accuracy, and meaning stood out consistently. Others showed potential in narrower domains, but gaps became visible when broader reasoning or nuanced understanding were required.

As the field progresses, it’s clear that our tools for evaluation must evolve alongside the models themselves. The real contribution of this study lies in its framework—a mix of analytic and interpretive methods that aims to reflect the complexity of language itself. Hopefully, it can serve as a foundation for further academic exploration or practical development.

References

Amidei, J., Piwek, P., & Willis, A. (2019). The reliability of human judgments of natural language generation quality: A reassessment. *Proceedings of the 12th International Conference on Natural Language Generation*, 170–182. https://aclanthology.org/W19-8619

Amidei, J., Piwek, P., & Willis, A. (2019). The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. *Proceedings of the 12th International Conference on Natural Language Generation*, 172–182. https://doi.org/10.18653/v1/W19-8619

Amerandish. (n.d.). *Random forest: Everything you need to know about random forest.* https://amerandish.com/random-forest

Anthropic. (2025). *Claude 3.7 (Large language model).* https://www.anthropic.com/

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv preprint*, arXiv:1409.0473. https://arxiv.org/abs/1409.0473

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of* *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Belz, A., & Kow, E. (2010). Discrepancies between automatically and manually evaluated test suites for NLG. *Proceedings of the 6th International Natural Language Generation Conference*, 354–358.

https://doi.org/10.3115/1855276.1855332

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 313–320. https://aclanthology.org/E06-1040

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *Stanford Center for Research on Foundation Models (CRFM).* https://arxiv.org/abs/2108.07258

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

https://arxiv.org/abs/2005.14165

Bryant, C., Felice, M., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. *Proceedings of the 14th Workshop on Building Educational Applications*, 52–75. https://doi.org/10.18653/v1/W19-4406

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256. https://aclanthology.org/E06-1032

Centre for Economic Policy Research (CEPR). (2025). *AI-driven sentiment strategies in asset management.* https://cepr.org/publications/ai-sentiment-2025

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint*, arXiv:1901.03407. https://arxiv.org/abs/1901.03407

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58.

https://doi.org/10.1145/1541880.1541882

Cohere. (2025). *Command R (Large language model).* https://cohere.com/

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE Symposium Series on Computational Intelligence*, 159–166. https://doi.org/10.1109/SSCI.2015.33

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Devarakonda, R. R. (2023). Machine learning approach for fraud detection in a financial services application. *SSRN.* https://doi.org/10.2139/ssrn.5234670

Deveikyte, J., Gaspareniene, L., & Remeikiene, R. (2020). The impact of social media sentiment on stock market volatility. *Journal of Financial Markets and Data Science, 3*(2), 101–112. https://doi.org/10.1016/j.jfmds.2020.04.003

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Google. (2025). *Gemini 1.5 (Large language model).*

https://deepmind.google/

Goyal, T., Chen, D., & Durrett, G. (2022). News summarization and evaluation in the era of GPT-3. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4782–4797. https://doi.org/10.18653/v1/2022.acl-long.331

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and ROUGE. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 210–220. https://aclanthology.org/D15-1021

Hasselt University. (n.d.). *Admission requirements – Master of Management.* UHasselt. https://www.uhasselt.be/en/study/programmes/master-of-management#anch-b12-admission-requirements

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2009.03300

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

Kasner, Z., Zhang, Y., Tan, S., Gehrmann, S., & Wang, A. (2023). Language model evaluation beyond perplexity. *arXiv preprint*, arXiv:2305.14598. https://arxiv.org/abs/2305.14598

Krishna, S., Zellers, R., & Smith, N. A. (2023). Accuracy and robustness of large language models. *Proceedings of the 2023 Annual Conference of the Association for Computational Linguistics.*

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346. https://doi.org/10.18653/v1/2020.emnlp-main.750

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. https://cran.r-project.org/doc/Rnews/

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. https://aclanthology.org/W04-1013

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

MarketWatch. (2025). *Reddit and ICE team up to develop sentiment-based financial products.* https://www.marketwatch.com/news/reddit-ice-sentiment-finance

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

OpenAI. (2023). *GPT-4 (Large language model).* ChatGPT.
https://chat.openai.com/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. https://doi.org/10.3115/1073083.1073135

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI.* https://www.openai.com/research/language-unsupervised

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics, 44*(3), 393–401. https://doi.org/10.1162/coli_a_00322

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Shinn, M., MacKay, M., & Rajpurkar, P. (2023). Can large language models transform healthcare? *Nature Medicine, 29*(3), 505–507. https://doi.org/10.1038/s41591-023-02218-0

Shinn, M., Perrault, A., & Wu, T. (2023). Educators rising to the challenge: Integrating large language models in education responsibly. *arXiv preprint*, arXiv:2303.03380. https://arxiv.org/abs/2303.03380

Sprout Social. (2024). *How brands use sentiment analysis to drive engagement.* https://sproutsocial.com/insights/sentiment-analysis-2024

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 809–819. https://doi.org/10.18653/v1/N18-1074

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems (NeurIPS), 32.* https://arxiv.org/abs/1905.00537

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/1904.09675

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2020). DIALOGPT: Large-scale generative pretraining for conversational response generation. *arXiv preprint*, arXiv:1911.00536. https://arxiv.org/abs/1911.00536