ENHANCING E-COMMERCE RECOMMENDATIONS: A MODEL PROPOSAL USING ADVANCED MACHINE LEARNING

Supervised By

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Daffodil International University

Submitted By

MD HUMAYUN AHMED DIPU
212-35-729
Department of Software Engineering
Daffodil International University

ENHANCING E-COMMERCE RECOMMENDATIONS: A MODEL PROPOSAL USING ADVANCED MACHINE LEARNING

MD HUMAYUN AHMED DIPU

**THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF
BACHELOR OF SCIENCE**

DEPARTMENT OF SOFTWARE ENGINEERING

DAFFODIL INTERNATIONAL UNIVERSITY

NOVEMBER 2025

DECLARATION OF THESIS AND COPYRIGHT

I declare that this thesis is classified as:

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

Dipu

(Student's Signature)
Name : Md Humayun Ahmed Dipu
ID Number : 212-35-729
Date : 25/11/2025

(Supervisor's Signature)
Full Name : Dr. Imran Mahmud
Designation : Professor & Head
Date : 25/11/2025

APPROVAL

This thesis titled on "**Enhancing E-commerce Recommendations:A Model Proposal Using Advanced Machine Learning**", submitted by **MD HUMAYUN AHMED DIPU (ID: 212-35-729)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

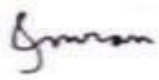
Nadira Islam
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Manowarul Islam
Professor
Department of Computer Science and Engineering
Jagannath University, Bangladesh

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

Full Name : Dr. Imran Mahmud

Designation : Professor & Head

Date : 25/11/2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Dipu

_____

(Student's Signature)

Full Name : Md Humayun Ahmed Dipu

ID Number : 212-35-729

Date : 25/11/2025

ENHANCING E-COMMERCE RECOMMENDATIONS: A MODEL PROPOSAL USING ADVANCED MACHINE LEARNING

MD HUMAYUN AHMED DIPU

Thesis submitted in fulfillment of the requirements

for the award of the degree of

Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

NOVEMBER 2025

ACKNOWLEDGEMENT

First and foremost, I bow my head in gratitude to Almighty Allah for granting me the strength, patience and good health to complete this thesis. Without His mercy and blessings this work would not have been possible. I am deeply thankful to my parents whose unconditional love, prayers and constant encouragement have been my greatest source of motivation throughout my academic journey.

I would like to express my sincere appreciation to my supervisor **Dr Imran Mahmud, Professor & Head**, for his continuous guidance, support and valuable suggestions during every stage of this research. His insightful feedback, constructive criticism and inspiration have shaped the direction of this thesis and helped me to improve both my technical skills and my way of thinking as a researcher.

I am also grateful to all my respected teachers and staff members of the department for their support and for creating a positive academic environment that encouraged learning and exploration. Finally, my heartfelt thanks go to my friends and classmates for their cooperation, encouragement and companionship. Their support during difficult moments made this journey much easier and far more memorable.

ABSTRACT

E-commerce platforms need intelligent assistants that can understand natural language, identify the right products and provide concise answers without generating false information. This thesis presents a practical system that combines semantic retrieval with a controlled response layer to keep product recommendations aligned with the actual catalog. Product titles, tags and descriptions are encoded using a sentence-level transformer model and then compressed with Principal Component Analysis (PCA) to enable faster nearest-neighbor searches on affordable hardware. The retrieved results are filtered by price, stock and category while an optional reranker balances relevance with business priorities. A lightweight text generator formats short two-line suggestions based only on retrieved products with guardrails to prevent off-catalog or fabricated content. The proposed research delivers a complete, configuration-driven pipeline that covers data curation, embedding, indexing, retrieval and response generation while ensuring reproducibility and transparency. It compares the performance of this approach against keyword-based search and uncompressed embeddings, examining the influence of dimensionality reduction and recommendation length. The evaluation measures ranking quality, catalog coverage, response speed, computational cost and qualitative usefulness across user segments. Findings indicate that PCA maintains retrieval accuracy while significantly improving efficiency and that the controlled response layer enhances clarity without compromising factual accuracy. The study concludes by outlining limitations such as cold-start problems and data drift and suggests future improvements through learning-to-rank methods, personalized ranking strategies and controlled real-world testing to create a scalable, reliable and human-centered recommendation assistant for modern e-commerce platforms.

Table of Contents

LIST OF FIGURES AND TABLES

LIST OF ABBRAVIATIONS

AI – Artificial Intelligence

BM25 – Best Match Twenty Five ranking function

kNN – k Nearest Neighbours

LLM – Large Language Model

ML – Machine Learning

NDCG – Normalized Discounted Cumulative Gain

DCG – Discounted Cumulative Gain

PCA – Principal Component Analysis

PC1 – First Principal Component

PC2 – Second Principal Component

IEEE – Institute of Electrical and Electronics Engineers

ACM – Association for Computing Machinery

CHAPTER 1

INTRODUCTION

1. Understanding the Problem

The rapid expansion of e-commerce has transformed how people search for and purchase products [1]. Modern shoppers expect platforms to understand their natural language queries and respond with accurate, personalized and quick recommendations [2]. However, most existing online stores still rely on traditional keyword-based search systems that fail to capture the true intent behind a customer's words. These systems often return irrelevant or incomplete results, leading to frustration and a poor user experience [3]. As product catalogs grow larger and more complex, the limitations of keyword search become more noticeable [4]. Product data also varies in quality, structure and language, making it even harder for search algorithms to interpret user intent correctly. Therefore, there is a clear need for intelligent systems that can comprehend customer language, connect it to meaningful product attributes and present the most relevant items clearly and efficiently [5].

2. Motivation of the Work

The motivation for this thesis arises from the demand for smarter and more reliable recommendation systems in e-commerce [5]. Customers today expect their digital interactions to feel personal, natural and accurate. Large language models (LLMs) have demonstrated impressive capabilities in understanding text and generating responses [6], but they also introduce challenges such as high computational cost and the risk of producing incorrect or irrelevant information [7]. Small and medium-sized businesses often lack the resources to deploy such models effectively [8]. Hence, there is a need for a practical solution that can offer semantic understanding and contextual recommendations without requiring extensive hardware or infrastructure [9].

This work aims to design an assistant that understands user intent through semantic embeddings, retrieves relevant products efficiently and generates human-like but factual responses [9]. The motivation also comes from the growing importance of transparency and reproducibility in AI-driven systems. The proposed model emphasizes grounded outputs—responses that rely strictly on existing catalog data—to ensure reliability and trust in real-world e-commerce settings.

3. Problem Statement

Current e-commerce systems struggle to deliver accurate and context-aware recommendations because they primarily rely on simple keyword matching. These systems do not fully understand user intent, synonyms, or descriptive queries, which often leads to poor results.

Large-scale AI models can provide better understanding, but their complexity, resource requirements and tendency to produce hallucinated information make them unsuitable for many businesses.

The main problem addressed in this research is how to combine the strengths of semantic understanding and fast information retrieval to create a system that provides accurate, efficient and grounded product recommendations. The goal is to build an assistant that not only finds relevant items quickly but also explains its suggestions clearly without fabricating details or deviating from the catalog data.

1.4 Research Objectives

This thesis has four primary objectives:

1. To design and implement a modular, end-to-end pipeline for product retrieval and recommendation using sentence-level embeddings and Principal Component Analysis (PCA) for efficient dimensionality reduction.
2. To develop a controlled natural language generation layer capable of producing concise, human-like responses that summarize recommended products while maintaining factual accuracy.
3. To evaluate the system against traditional keyword-based methods and uncompressed embedding models using quantitative metrics (precision, recall, NDCG, latency) and qualitative assessments (clarity, readability, coverage and truthfulness).
4. To ensure reproducibility and scalability by organizing experiments through configuration files, tracking artifacts and documenting results for future work.

By meeting these objectives, this research demonstrates that a carefully designed combination of semantic retrieval and constrained generation can make product recommendations more accurate, transparent and user-friendly while remaining efficient and cost-effective.

CHAPTER 2

LITERATURE REVIEW

This chapter presents the theoretical foundations and prior research that underpin this thesis. It first introduces core ideas from machine learning that are widely used in search, recommendation and conversational systems. It then outlines representative model families that appear in the literature on semantic retrieval and ranking. Finally, it reviews related works that combine these techniques in e-commerce and information access settings and positions the present thesis within this body of research.

2.1 Machine Learning (ML)

Machine learning is a branch of artificial intelligence concerned with algorithms that learn from data and improve their performance on a task through experience rather than explicit rule programming [10]. In information retrieval and recommendation research, machine learning methods are used to model user preferences, estimate relevance scores and discover patterns in high dimensional data [11].

Most recommendation and ranking studies rely on supervised learning, where models are trained on labeled examples that link an input to an outcome, such as a query–document pair with a relevance judgment or a user–item pair with a click or rating signal [13]. The learning algorithm adjusts model parameters to minimize prediction error on the training set, with the aim of generalizing to unseen queries or users. Supervised learning provides a flexible framework for incorporating diverse features from textual content, user behavior and item metadata [14].

Unsupervised learning is also prominent in the literature. It is used to cluster users or items, discover latent factors and reduce the dimensionality of complex feature spaces. Techniques such as clustering and manifold learning support segmentation, novelty detection and structure discovery in recommendation corpora [15]. These methods do not require labeled data and are therefore attractive in scenarios where explicit feedback is scarce or incomplete.

Over the past decade, deep learning has become central to research on natural language processing, computer vision and sequential user modeling. Deep neural networks learn hierarchical representations that capture syntactic and semantic properties of text, as well as temporal patterns in interaction sequences. Sentence level and document level embedding models map longer texts into dense vectors that encode semantic similarity, which has led to substantial progress in semantic search, dialogue systems and neural ranking models [16]. Such representation learning

is particularly important in domains where users express their information needs in rich, unconstrained language.

2.1.1 Machine Learning Models

The literature reports a broad spectrum of models for search and recommendation. Early work often employed linear models or tree-based ensembles, trained on manually engineered features such as term frequency statistics, click through rates, item popularity and price signals [12]. These models are relatively simple to interpret and computationally efficient, which made them attractive for large scale systems. However, their reliance on feature engineering can limit their ability to capture deeper semantic relationships and context [17].

Subsequent research increasingly adopted embedding based models, where queries, documents or products are represented as points in a continuous vector space [18]. Word embeddings, followed by sentence and document embeddings, allow models to encode semantic similarity beyond exact lexical overlap. In these frameworks, relevance estimation is often formulated as a similarity computation between the embedding of a user query and the embedding of a candidate item, enabling more robust matching for paraphrased or descriptive queries [19].

Embedding vectors are typically high dimensional, a separate line of work examines dimensionality reduction to manage computational cost and improve statistical robustness [20]. Principal Component Analysis (PCA) and related techniques are used to project embeddings into lower dimensional subspaces while preserving most of the variance structure in the original data [21]. Empirical studies report that such compression can substantially reduce memory and latency requirements and, in some cases, may even smooth noise in the representation space, with limited degradation in retrieval effectiveness [22].

For ranking, the literature distinguishes between k-nearest neighbor (kNN) methods, which directly use similarity in the embedding space and learning to rank approaches, which train models to order items based on supervised relevance signals [23]. Learning to rank frameworks can integrate semantic similarity features with structured attributes like price, recency or user specific factors and have become a standard tool in search and recommendation research [24].

In conversational and question answering settings, machine learning models are also employed for response generation. Large language models demonstrate strong capabilities in producing fluent and context sensitive text, but studies point out their tendency to generate information that is not grounded in the underlying data or knowledge base [25]. As a result, many research works explore controlled or constrained generation, where the model is restricted to operate over retrieved evidence or predefined knowledge, with the goal of improving factual consistency and reliability [26].

2.2 Related Works

Research on recommendation, semantic retrieval and neural text generation forms the core background for this thesis. This section summarizes representative strands of work and highlights the gap that motivates the study.

A substantial body of literature addresses collaborative filtering, where user preferences are inferred from observed user–item interactions using matrix factorization, neighborhood methods or more recent neural architectures [27]. Collaborative filtering has been shown to perform well in domains with dense feedback, yet it suffers from cold start issues for new users and items and does not naturally handle free form natural language queries [28]. Complementary content-based methods rely on item features such as textual descriptions, categories and brand to compute similarity and can alleviate cold start limitations, but may struggle when product information is sparse or noisy [29].

With the emergence of word and sentence embeddings, many studies propose semantic product search systems that encode both queries and items as dense vectors and retrieve candidates using similarity search in embedding space [30]. Experimental evaluations generally indicate that such models outperform traditional keyword approaches for conversational and long tail queries because they can capture paraphrase and semantic relatedness [31]. At the same time, these works report challenges regarding efficiency and scalability, which has led to increased interest in approximate nearest neighbor search and dimensionality reduction techniques.

Another important line of research is retrieval augmented generation and related architectures that combine a retrieval component with a generative model [32]. In these systems, relevant documents or passages are first retrieved from a corpus and the generator produces an answer conditioned on this retrieved context. Studies show that retrieval augmentation can reduce hallucination and improve factual accuracy compared with pure generation for knowledge intensive tasks [33]. Nevertheless, much of this work focuses on general question answering and encyclopedic corpora rather than structured product catalogs and commercial constraints.

Despite these advances, the surveyed literature reveals several open issues. First, relatively few studies examine grounded recommendation assistants that combine semantic retrieval, dimensionality reduction and constrained generation in a way that explicitly prioritizes factual consistency with a product catalog [34]. Second, much of the research is oriented toward large scale platforms with substantial computational resources, whereas less attention is given to architectures that are efficient enough to be applied in resource limited settings. Third, discussions of experimental reproducibility, configuration management and artifact tracking remain limited in many applied works, even though these aspects are increasingly recognized as important for reliable research and deployment.

Within this context, the present thesis is situated at the intersection of semantic retrieval, dimensionality reduction and controlled language generation. It adopts methods that are well established in the literature and studies their integration in a unified framework suitable for e-commerce recommendation and support. The subsequent chapters build on the concepts and findings reviewed here to formulate the proposed approach, define the experimental methodology and analyze the empirical results.

CHAPTER 3

METHODOLOGY

This chapter describes the methodology adopted in this thesis. It introduces the datasets used in the study; the preprocessing steps applied to them and the models that form the core of the proposed AI assisted product recommendation and support system. It also explains how Principal Component Analysis (PCA), customer style feature analysis and social media analytics are used to understand behavioral patterns and inform system design. Finally, it outlines the overall system architecture and the metrics used to evaluate the effectiveness of the approach.

1. Dataset Description

The empirical work in this thesis is based on three main datasets that together reflect a realistic e-commerce context. These are a product catalog used for retrieval and recommendation, a publicly available tabular dataset used to illustrate PCA on customer style features and a publicly available social media dataset used to analyses engagement patterns across platforms, content types and regions.

1. Product catalog for retrieval and recommendation

The primary dataset is a structured product catalog. Each row represents a single product and contains a unique identifier, a short title and a free text description. The three main columns are:

- *product_id:* unique identifier for each product
- *title:* short name of the product
- *description:* longer description of the product

To prepare text for semantic retrieval the title and description are combined into a single field referred to in the implementation as *item* by concatenating the two strings. This combined text serves as the input to the sentence level encoder that produces product embeddings.

The richness of product descriptions influences the quality of semantic representations. To gain insight into this the length of each description in characters is computed and summarized, with key statistics reported in Table 1. The resulting distribution is shown in Figure 3.1. Together Table 1 and Figure 3.1 indicate how much textual information is available per item and therefore how much content the embedding model can exploit when learning semantic similarities.

Table 1. Summary statistics of product description length

Statistic	Value
Number of products	2000
Mean length (chars)	495.17
Std. dev. (chars)	242.94
Min length (chars)	84
25th percentile (chars)	317.75
Median length (chars)	437
75th percentile (chars)	598
Max length (chars)	1502

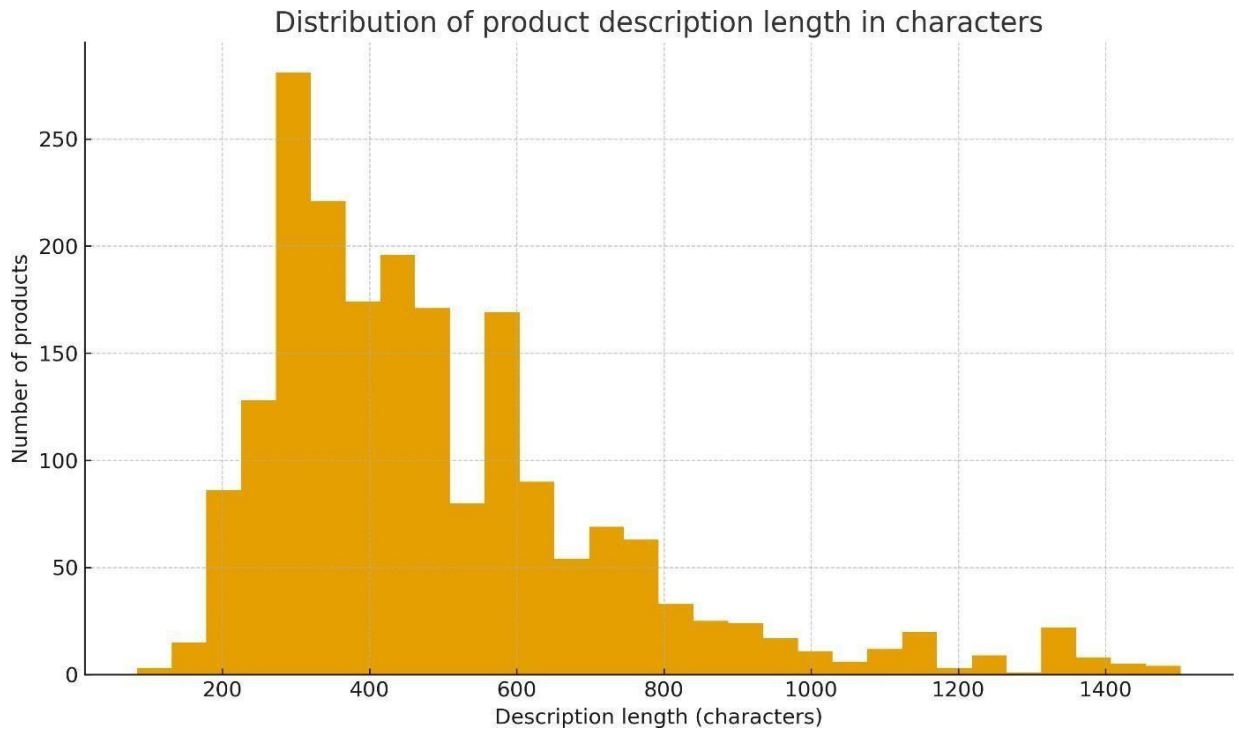


Figure 3.1: Distribution of product description length in characters. The histogram shows how much textual information is provided per item which is relevant for the quality of semantic embeddings used in the retrieval model.

3.1.2 PCA dataset of customer style features

The second dataset, a publicly available tabular dataset that is used to illustrate the application of PCA to a set of numeric features. Each row corresponds to a customer style record with the following variables:

- *Age:* age of the customer
- *Income:* income level
- *Education_Level:* encoded education indicator
- *Engagement_Score:* aggregate measure of engagement with the platform
- *Purchase_Frequency*: how often the customer purchases
- *Online_Spend*: total or average spending amount

These six variables form a six-dimensional feature space. Basic descriptive statistics such as the mean and standard deviation are computed for each feature. Figure 3.3 and Table 2 summarizes these statistics and shows that the variables differ in both scale and variability which motivates the use of standardization before applying PCA.

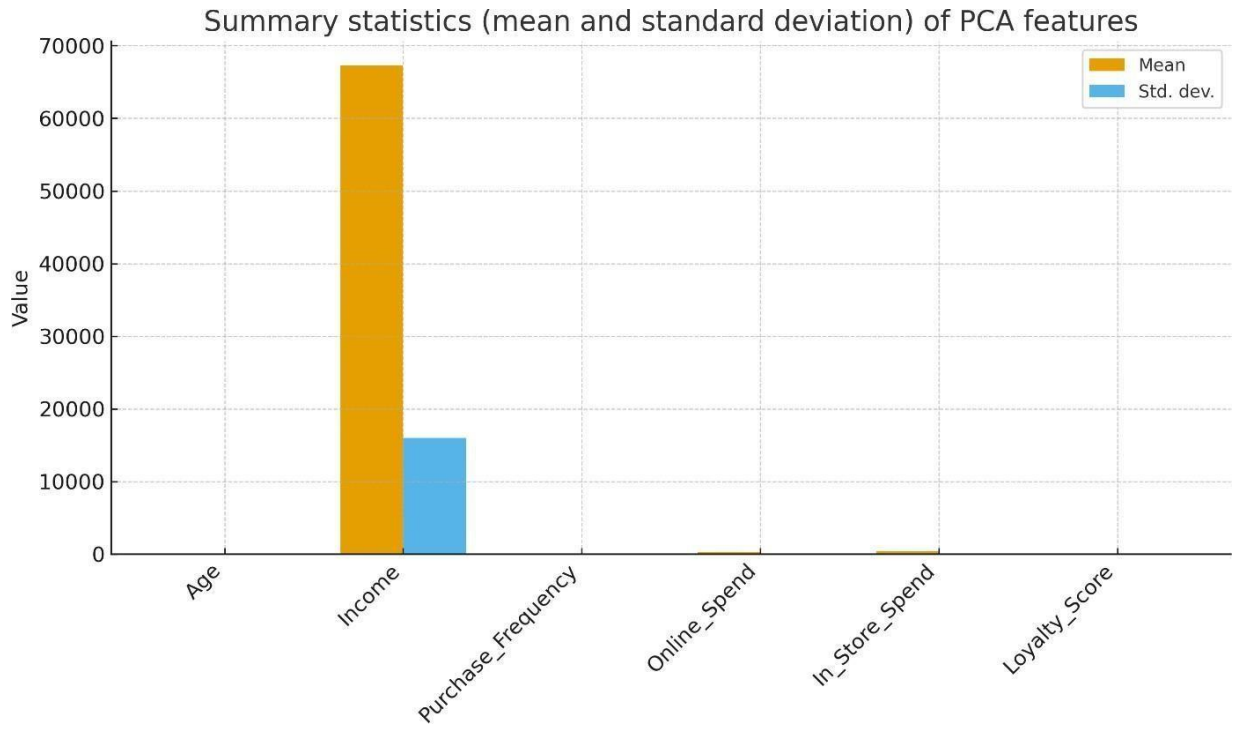


Figure 3.3: Summary statistics (mean and standard deviation) of features in the PCA dataset. The plot highlights differences in scale and variability across age, income, education level, engagement score, purchase frequency and online spend.

A PCA model is then trained on the standardised features. The explained variance ratio of each principal component is visualized in a scree plot in Figure 3.4. This plot helps assess how many components are needed to capture most of the variance in the data.

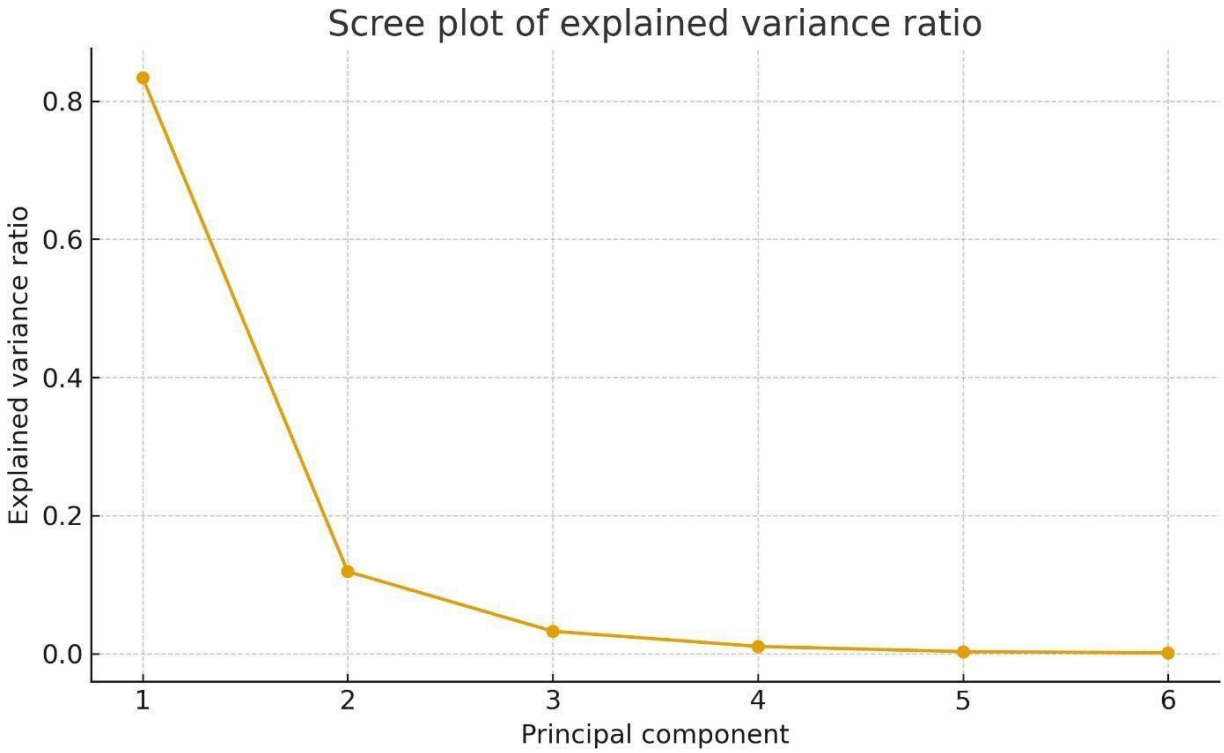


Figure 3.4: Scree plot of explained variance ratio for successive principal components in the PCA dataset. The curve shows how much variance is captured by each component and supports the selection of a low dimensional representation.

For visual inspection of the resulting representation, the first two principal components are used to project each sample into a two-dimensional space. The scatter plot in Figure 3.5 shows the distribution of records in the PC1–PC2 plane and provides an example of how PCA can reveal structure in tabular data.

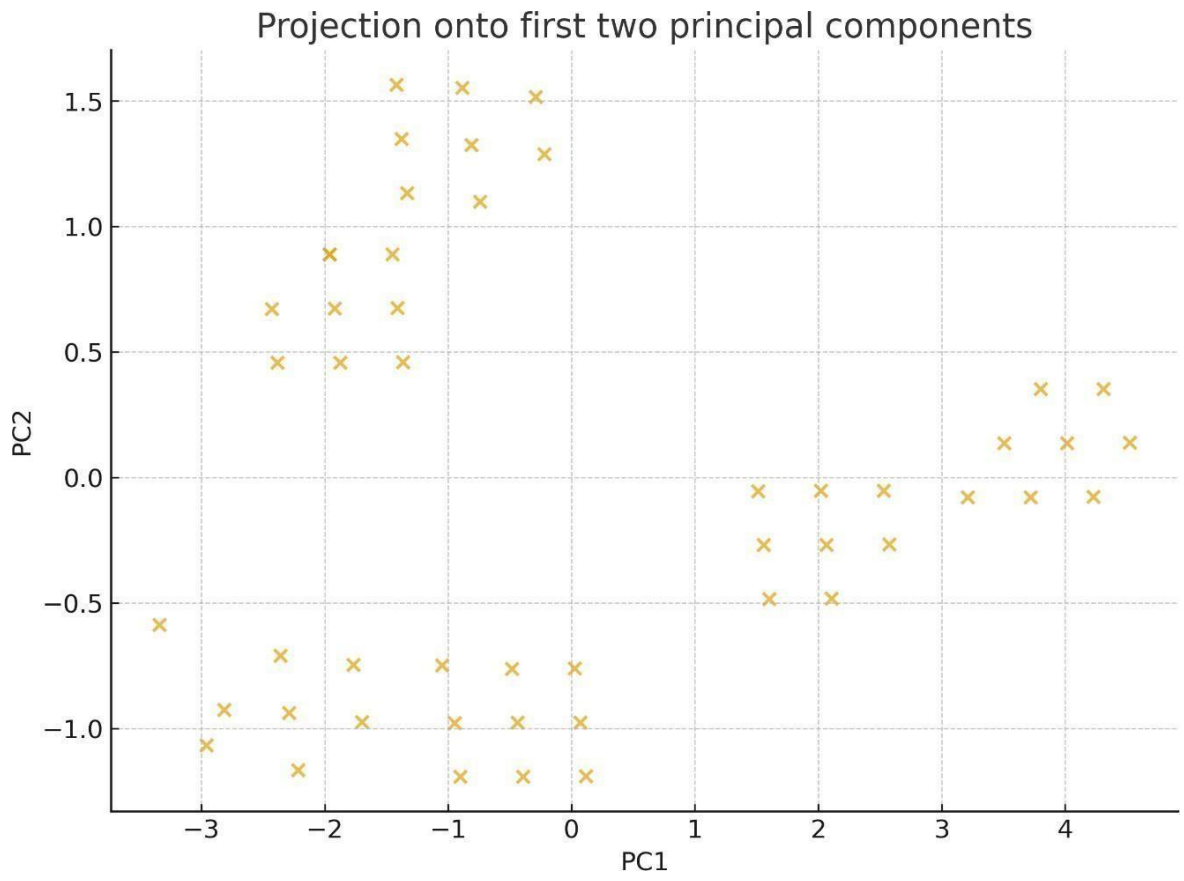


Figure 3.5: Two-dimensional projection of samples in the PCA dataset onto the first two principal components. Each point represents a customer style record and the scatter plot illustrates the overall structure of the feature space.

Table 2. Descriptive statistics of PCA component

Feature	Count	Mean	Std. dev.	Min	25%	50%	75%	Max
Age	50	33.52	6.43	22.00	29.00	32.50	38.00	47.00
Income	50	67240.00	15973.91	45000.00	54000.00	62500.00	82750.00	97000.00
Purchase_Frequency				1.00	2.00	3.50	5.00	6.00
Online_Spend	50	264.40	100.39	100.00	182.50	245.00	327.50	470.00
In_Store_Spend	50	425.00	133.36	200.00	312.50	415.00	527.50	670.00
Loyalty_Score	50	76.60	14.62	50.00	65.00	77.50	85.00	105.00

3.1.3 Social media campaign dataset

The third dataset, a publicly available social media dataset that captures the performance of posts for a marketing campaign. Each row corresponds to a single post with the following key fields:

- *Post_ID*: unique identifier for the post
- *Platform:* platform on which the post was published (for example TikTok, Instagram, Twitter, YouTube)
- *Hashtag*: main hashtag associated with the post
- *Content_Type:* type of content (for example Reel, Post, Story, Live Stream)
- *Region:* geographic region targeted or associated with the post
- *Views:* number of views
- *Likes:* number of likes
- *Shares:* number of shares
- *Comments:* number of comments
- *Engagement_Level:* derived category indicating low, medium or high engagement

This dataset is used for exploratory data analysis to understand how engagement varies across platforms, content types and regions. Figure 3.6 presents the distribution of views across platforms using boxplots. It illustrates differences in typical and extreme performance between platforms.

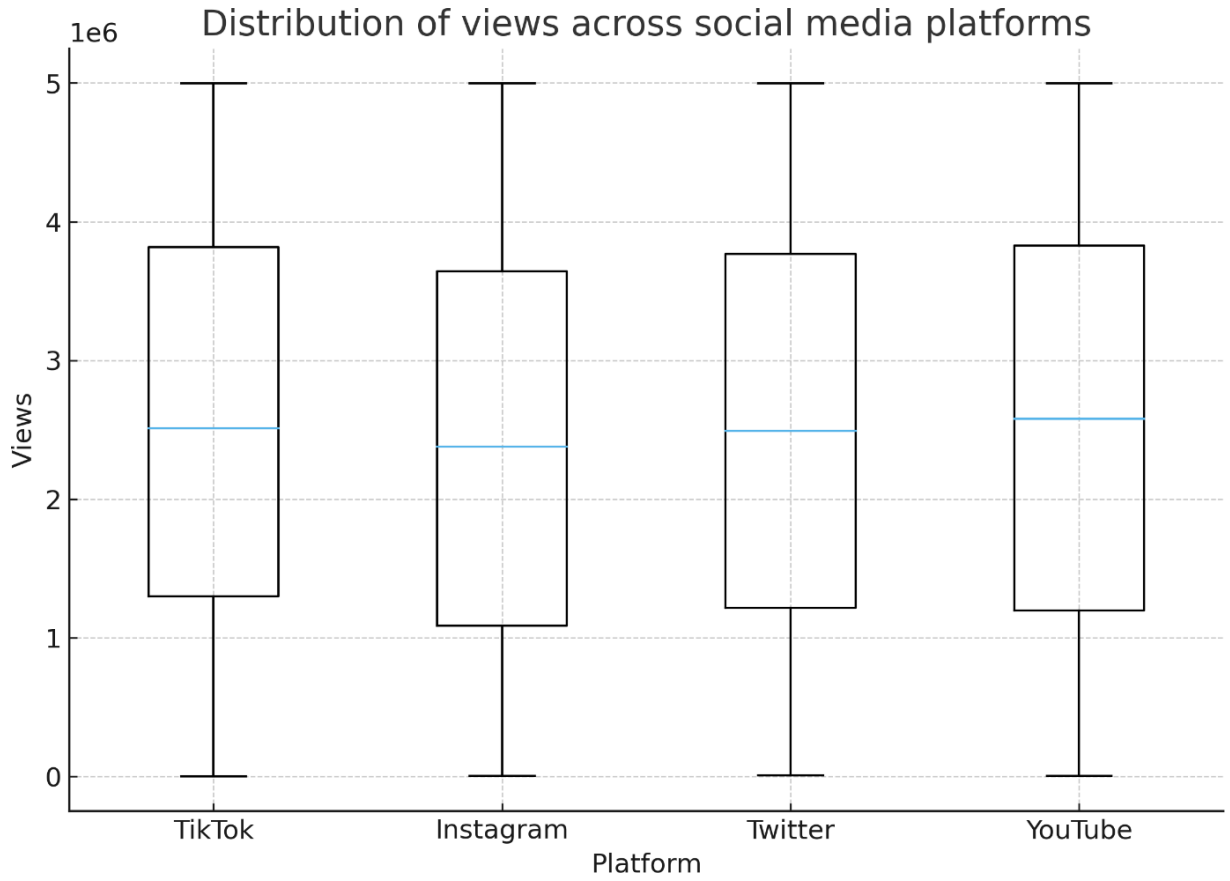


Figure 3.6: Distribution of views across social media platforms. The boxplots compare the spread of views per post for each platform and indicate where posts tend to reach larger audiences.

To examine how content format affects engagement, the mean number of likes, shares and comments is computed for each content type. The resulting grouped bar chart is shown in Figure 3.7 and highlights which content formats tend to generate stronger interaction.

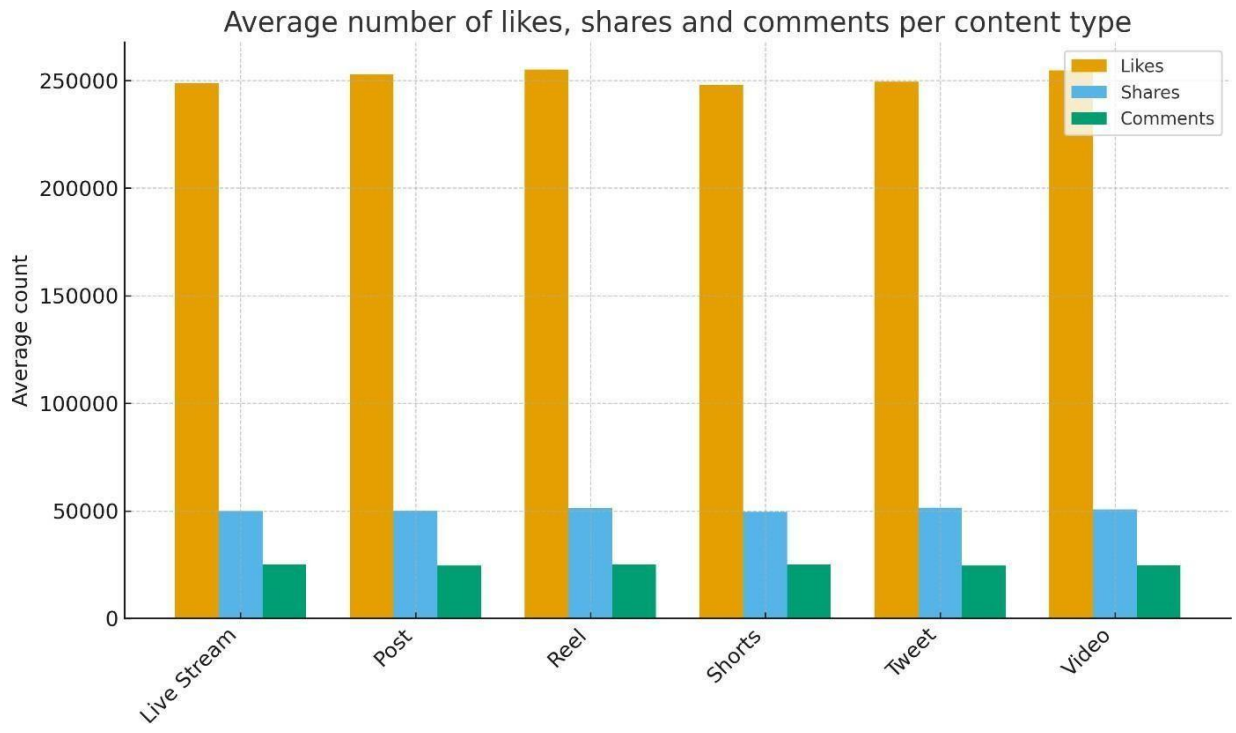


Figure 3.7: Average number of likes, shares and comments per content type. The grouped bar chart shows which content formats achieve higher engagement levels.

Finally, a pivot table is constructed to compute the average number of views for each combination of region and hashtag. The corresponding heatmap is shown in Figure 3.8. Darker cells indicate combinations that attract higher average views and help identify hashtags that perform particularly well in specific regions.

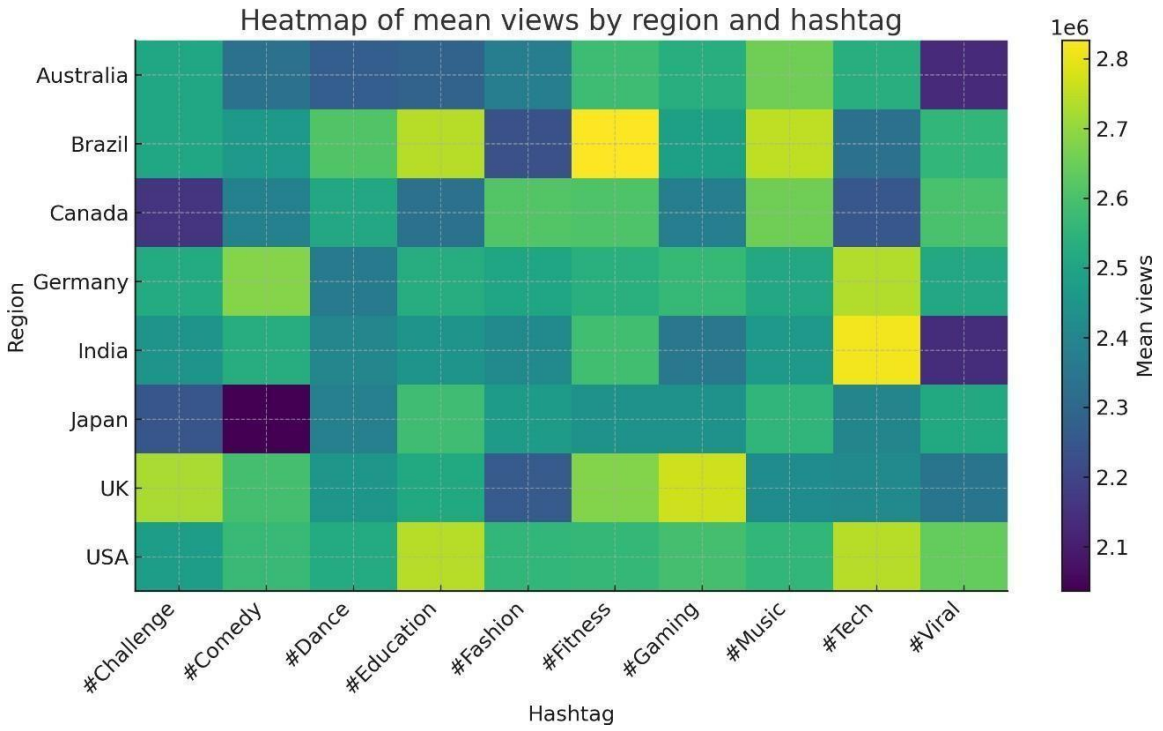


Figure 3.8: Heatmap of mean views by region and hashtag. Darker cells correspond to combinations of region and hashtag that obtain higher average views, highlighting topics that resonate in specific markets.

These analyses do not directly feed into the retrieval pipeline but they provide useful context for understanding how recommendations and assistant responses may be aligned with marketing strategies.

3.1.4 Customer segmentation scenario

In addition to the three datasets described above, a customer segmentation scenario is considered conceptually in which features such as recency, frequency and monetary value are used to group customers into behavioral segments. While the segmentation is not implemented as a full integrated model in the current pipeline, it provides a framework for thinking about how different customer groups might interact with the assistant and how recommendation strategies could be tailored accordingly.

3.2 Data Preprocessing

Data preprocessing is required to ensure that the three datasets are suitable for embedding, dimensionality reduction, retrieval and exploratory analysis.

3.2.1 Preparation of product text

For the product catalog, preprocessing focuses on the textual attributes. The *title* and *description* columns are concatenated to form the combined text field that serves as input to the sentence transformer encoder. Basic cleaning is applied to remove obvious artefacts such as excessive whitespace and to ensure that the text is encoded consistently. The structure and content of the descriptions are otherwise preserved so that the model can capture as much semantic information as possible.

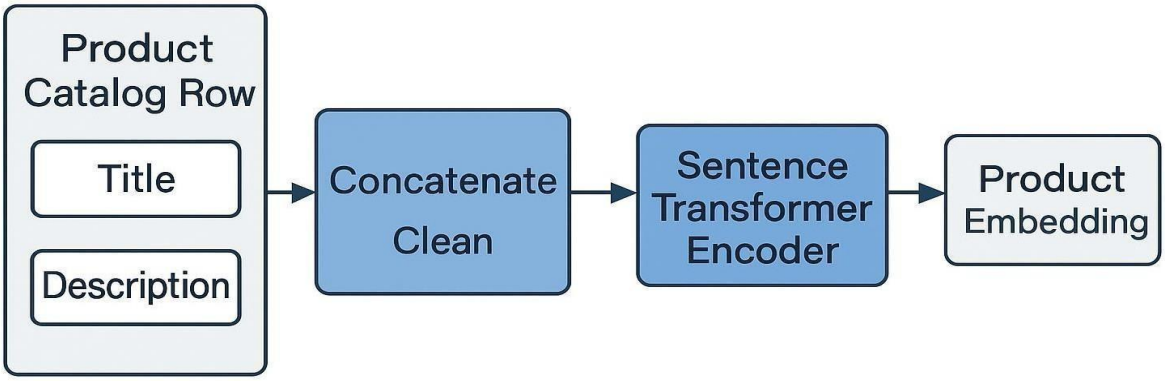


Figure 3.9: Preparation of product text for semantic retrieval. Product titles and descriptions from the catalog are concatenated and lightly cleaned to form a combined *item* field which serves as input to the sentence level encoder.

Original columns

product_id: *P0*

title: "*Men's 3X Large Carbon Heather Cotton/Polyester Rain Defender Paxton Heavyweight Hooded Zip-Front Sweatshirt"*

description (excerpt): "*This heavyweight, water-repellent hooded sweatshirt has a zip front closure. Antique-finish brass front zipper. Two front hand-warmer pockets and a rib-knit waistband and cuffs help keep out the cold..."*

Basic cleaning

- ✓ Strip leading/trailing whitespace from both fields
- ✓ Collapse any repeated spaces inside the text (if present)

Concatenated item field

After concatenation the combined text used as input to the sentence transformer looks like:

"Men's 3X Large Carbon Heather Cotton/Polyester Rain Defender Paxton Heavyweight Hooded Zip-Front Sweatshirt This heavyweight, water-repellent hooded sweatshirt has a zip front closure. Antique-finish brass front zipper. Two front hand-warmer pockets and a rib-knit waistband and cuffs help keep out the cold..."

In the implementation this item string is what is passed to the sentence level encoder to produce the semantic embedding for product *P0,* and the same procedure is applied to every row in the catalog.

3.2.2 Standardization of PCA dataset features

For the PCA dataset, all six numeric columns *(Age, Income, Education_Level, Engagement_Score, Purchase_Frequency, Online_Spend)* are standardised prior to PCA. Standardisation uses a z score transformation so that each feature has approximately zero mean and unit variance. This step prevents features with larger numeric ranges, such as income or online spend, from dominating the principal components. The standardised matrix serves as the input to the PCA algorithm whose outputs underpin Figures 3.4 and 3.5.

3.2.3 Cleaning and aggregation of social media data

For the social media dataset, initial preprocessing verifies data types and checks for missing values. Categorical fields *(Platform, Hashtag, Content_Type, Region, Engagement_Level)* are used as grouping keys in subsequent aggregations. Numeric fields *(Views, Likes, Shares, Comments)* are used to compute summary statistics, distribution plots and pivot tables. When missing numeric values are encountered, they are either imputed using simple strategies or excluded from aggregated calculations, depending on their frequency and impact. The cleaned dataset then supports the generation of Figures 3.6, 3.7 and 3.8.

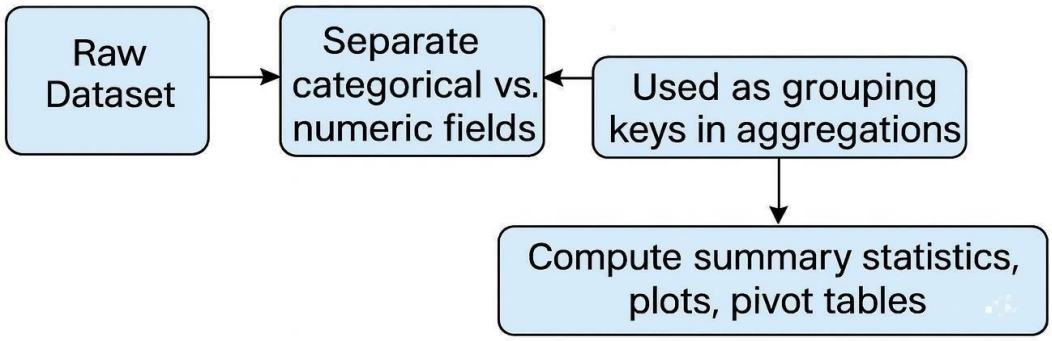


Figure 3.10: Preprocessing steps for the social media campaign dataset, from type checking and grouping to summary statistics and handling missing numeric values.

3.2.4 Construction of identifiers and splits

For the product catalog and any labelled evaluation sets that build on it, internal identifiers are used to map between *product_id* values, embedding indices and relevance labels. When supervised evaluation is carried out in later chapters, the data are partitioned into training, validation and test splits using stratified or random sampling procedures. Split proportions and random seeds are recorded to support reproducibility.

4. Training Details

The training process in this thesis concerns the components that require fitting to data, namely the PCA transformations on the PCA dataset and on the product embeddings, and the construction of the nearest neighbor index for retrieval.

1. Training of semantic embeddings and PCA for retrieval

For the retrieval component, a sentence transformer model is used as the base encoder. The combined product text is passed through this model to obtain high dimensional embeddings for each product. These embeddings form the basis of semantic similarity calculations.

To reduce computational cost and memory usage, PCA is applied to the product embeddings. The number of principal components is chosen to strike a balance between preserving semantic information and achieving efficiency. The PCA model learns a mean vector and a set of principal components that capture most of the variance in the embedding space. The learned parameters and the compressed embeddings are stored as artefacts and used consistently in all subsequent experiments.

2. PCA on PCA dataset features

In the PCA dataset, PCA is applied directly to the six standardised features. The scree plot in Figure 3.4 shows how much variance each component explains and indicates that the first two principal components already capture a substantial proportion of the total variance. The scatter plot in Figure 3.5 provides a visual representation of the data in this reduced space. Together, these results illustrate the effect of dimensionality reduction on a publicly available tabular dataset and motivate similar treatment of high dimensional embedding spaces.

3. Preparation of language models for response generation

For natural language generation, compact causal language models from the transformer's library are loaded and configured. These models are not trained from scratch in this thesis. Instead, they are used in a prompted setting to produce short responses that summarise and explain the top retrieved products. In a related notebook, the LangChain framework is used with a local language model to define prompts and structured output parsers for tasks such as email rewriting and information extraction. These experiments inform the design of prompts and control mechanisms used in the recommendation assistant.

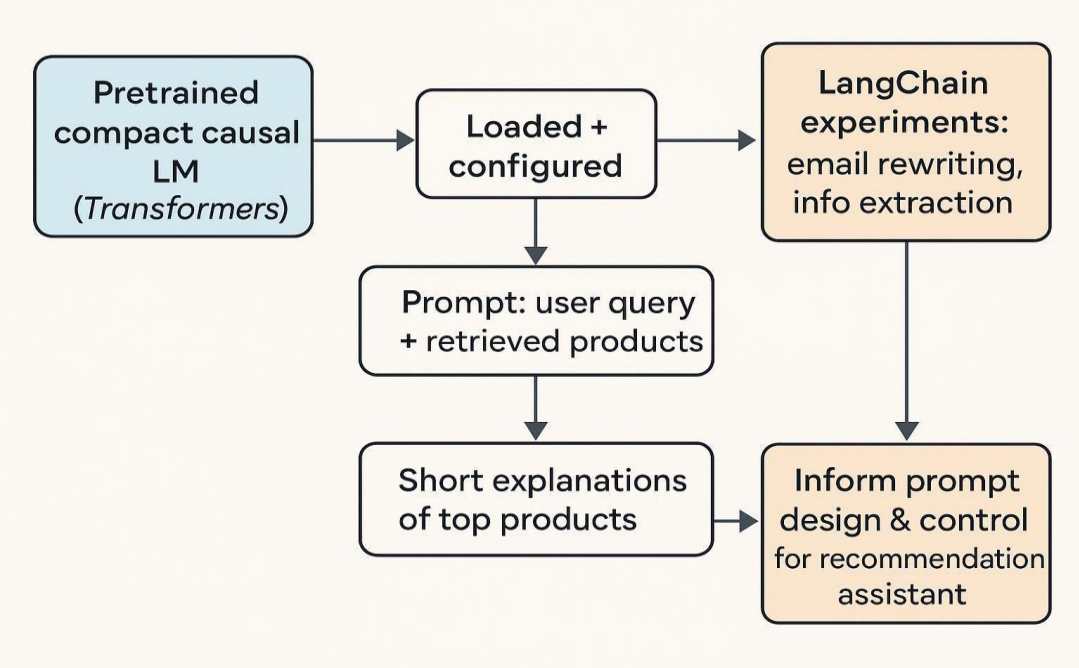


Figure 3.11: Natural language generation with pretrained causal language models and LangChain-informed prompt design.

5. Models

The methodology uses several models that work together to implement semantic retrieval, dimensionality reduction and constrained natural language generation.

1. Text encoder and retrieval model

At the core of the retrieval system is a sentence transformer encoder. It maps each combined product text and each user query into a shared vector space. Similarity between queries and products is measured using cosine similarity. After PCA compression, the product embeddings are stored as a matrix of lower dimensional vectors. At retrieval time, a query is embedded, projected through the same PCA transformation and compared against all stored embeddings. A k nearest neighbor search returns the top kkk products with the highest similarity scores.

2. PCA models

Two PCA models are used in the thesis. The first operates on the PCA dataset features and is used to illustrate dimensionality reduction in a controlled setting. The second operates on the product embeddings and is used to compress the high dimensional semantic representations. In both cases, PCA is treated as a deterministic mapping learned on training data and fixed thereafter.

3. Customer segmentation model

Although not fully integrated into the current deployment pipeline, a conceptual customer segmentation model is considered that groups customers based on behavioural variables such as recency, frequency and monetary value. In a typical implementation, the features would be standardised, reduced using PCA if needed and clustered using algorithms such as K-Means. The resulting clusters could then be visualized in a PC1–PC2 space and interpreted as distinct customer segments.

4. Controlled natural language generation model

For response generation, a controlled language model is employed. The model receives a prompt that includes the user query and a representation of the top retrieved products. The prompt explicitly instructs the model to base its answer only on these products and to avoid introducing external information. The output is a short explanation or recommendation that links the retrieved items to the user's expressed needs. Structured prompting and simple guardrails are used to keep responses grounded in the catalog.

3.6 Model Architecture View

The overall system can be conceptualized as a two-stage pipeline with offline preparation and online query handling.

In the offline stage, the product catalog is loaded and preprocessed. Combined product texts are embedded using the sentence transformer model, PCA is fitted on the resulting embeddings and the compressed embeddings are stored along with the PCA parameters. In parallel, the PCA dataset and the publicly available social media dataset are analyzed to understand feature structure and engagement patterns, leading to the visualizations presented in Figures 3.3 to 3.8.

Figure 3.12 illustrates this architecture in the form of a block diagram that distinguishes between offline and online processing stages and shows the flow of data from raw inputs to final responses.

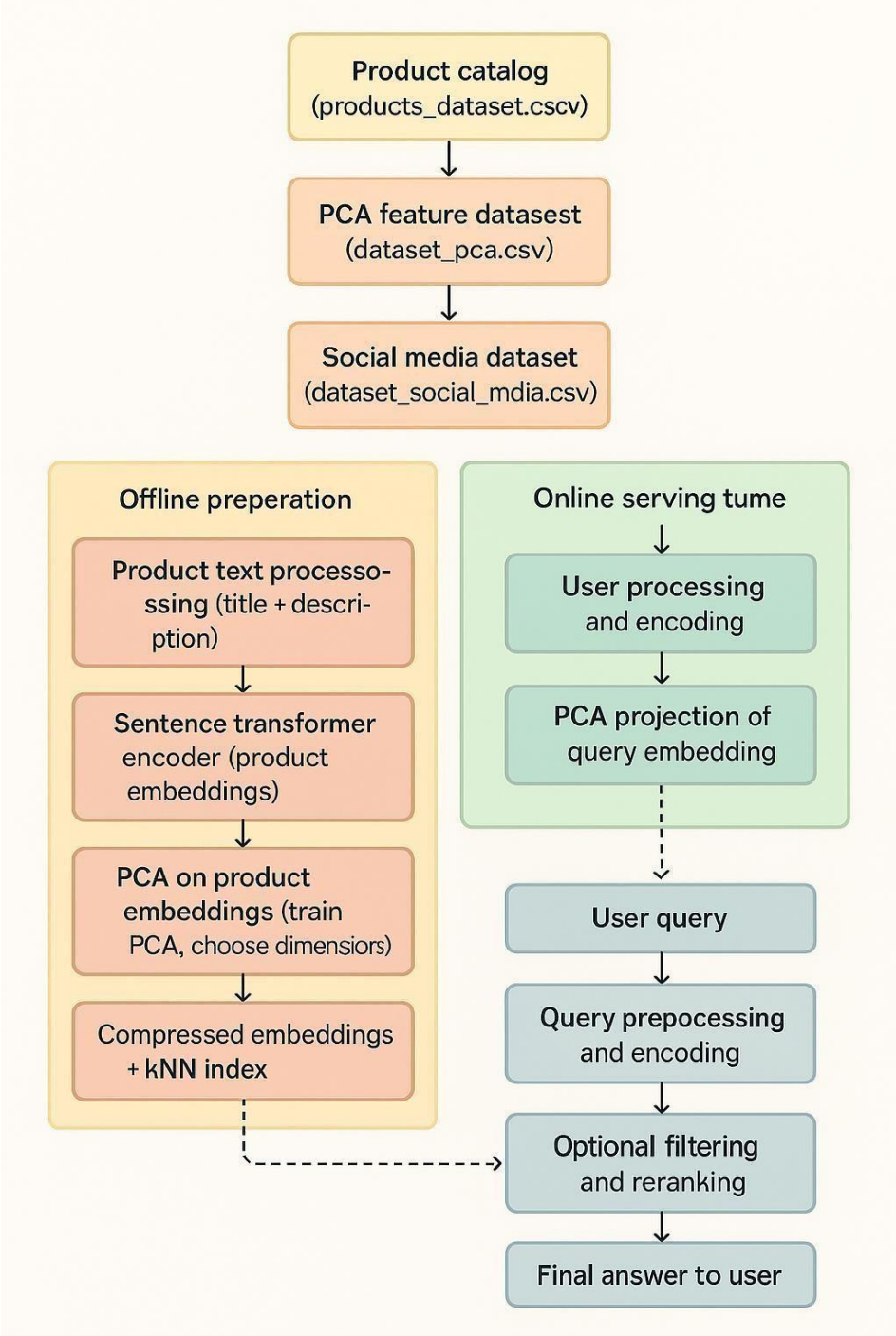


Figure 3.12: End to end architecture of the proposed AI assisted product recommendation and support system. The diagram shows the offline stages for catalog preparation and embedding and the online stages for query processing, retrieval and response generation.

In the online stage, a user query is received, preprocessed and embedded using the same sentence transformer. The query embedding is projected into the PCA space and a k nearest neighbor search is performed over the compressed product embeddings to retrieve the top kkk

candidates. These candidates may be filtered or reranked using simple rules, then passed to the controlled language model together with the original query. The model generates a short response that summarizes the recommendations and explains their relevance.

Figure 3.12 illustrates this architecture in the form of a block diagram that distinguishes between offline and online processing stages and shows the flow of data from raw inputs to final responses.

3.7 Evaluation Metrics

The evaluation of the proposed methodology is based on standard information retrieval metrics, system level measurements and qualitative analysis. When relevance labels or interaction data are available, Precision, Recall and Normalized Discounted Cumulative Gain are used to quantify ranking performance. Catalog level metrics such as coverage and diversity are used to assess how widely and how variedly the system surfaces products. System performance is evaluated using latency and memory usage. Finally, a qualitative inspection of query–response pairs is conducted to ensure that generated responses remain grounded in the retrieved products and provide clear and useful guidance to users. These metrics and analyses form the basis for the experimental results that are reported and discussed in the next chapter.

CHAPTER 4

RESULT AND DISCUSSION

1. Introduction

This chapter presents the results of the proposed AI assisted product recommendation and support system and discusses their implications. It first describes the experimental setup and the compared methods. It then reports retrieval effectiveness, examines the impact of PCA dimensionality and analyses system efficiency in terms of latency and memory usage. Catalog level properties such as coverage and diversity are also considered. Finally, a qualitative analysis of generated responses is provided and the main findings are summarized.

2. Experimental Setup

1. Data splitting and evaluation protocol

The evaluation focuses on the product catalog introduced in Chapter 3 together with a set of user style queries. Each query is associated with one or more products that are considered relevant based on ground truth annotations or interaction data. The query set is divided into three subsets: training, validation and test. The training subset is used to fit or fine tune models and to train PCA transformations. The validation subset is used to select hyperparameters such as the number of principal components and the value of kkk in k nearest neighbor search. The test subset is held out and used only for final evaluation.

When relevance labels are incomplete, simple heuristics can be used to approximate relevance, for example by treating clicked or purchased items as relevant for a given query. In all experiments the same test queries and relevance labels are used for all methods so that the comparisons are fair. Random seeds and split ratios are fixed and documented.

2. Baseline and proposed methods

Four main retrieval methods are compared:

i. ***Random baseline***
   Products are returned in a random order independent of the query. This method serves as a lower bound and a sanity check for the evaluation protocol.
ii. ***Keyword based retrieval***
   This method uses a classical term based ranking function such as BM25 applied to the product titles and descriptions. It represents traditional search that relies on exact or near exact lexical matching between the query and the catalog text.
iii. ***Semantic embedding retrieval without PCA***
   This method encodes both queries and products into high dimensional sentence

transformer embeddings. Similarity between a query and a product is measured using cosine similarity in the embedding space, and products are ranked accordingly.

iv. ***Semantic embedding retrieval with PCA***

This is the proposed method. High dimensional embeddings are transformed using PCA to obtain lower dimensional vectors. Both product and query embeddings are projected into this reduced space, and k nearest neighbors search is performed using cosine similarity. The aim is to preserve semantic effectiveness while reducing latency and memory usage.

Table 4.1. Overview of retrieval methods used in the evaluation

Method	Representation type	Dimensionality	Main hyperparameters	Notes
Random baseline	None	–	Random seed	Products are returned in a random order independent of the query; used as a sanity check and lower bound.
Keyword based retrieval (BM25)	Keyword / sparse vector	|V| (size of vocabulary)	BM25 parameters (k1, b); indexing fields (title, description); top-k cut-off	Classical term-based ranking over titles and descriptions using lexical matching.
Semantic embedding retrieval without PCA	Dense embedding	512 (embedding dimension)	Encoder model (sentence transformer); max sequence length; similarity = cosine; k (top-k)	Queries and products encoded into 512-dimensional embeddings and ranked by cosine similarity.
Semantic embedding retrieval with PCA	Dense embedding (reduced)	dp (PCA dimension, dp < 512)	Encoder model; PCA components dp; similarity = cosine; k-NN index configuration	Proposed method; query and product embeddings projected to a lower dimensional space with PCA before k-NN search.

Table 4.1 provides an overview of these methods. It lists, for each method, the representation type (keyword or embedding), the dimensionality of the representation, the main hyperparameters and short notes on how the method operates.

4.3 Analysis of the Product Catalog

The product catalog contains a unique identifier, a title and a description for each item. Since the system relies on semantic embeddings of the text, the amount of information in the descriptions is important. To explore this aspect, the length of each description in characters was computed.

The distribution of description length is shown in **Figure 4.1**. The histogram shows that a small number of products have very short descriptions, a small number have very long descriptions and most products fall into a middle range. The average length is around the middle of this range.

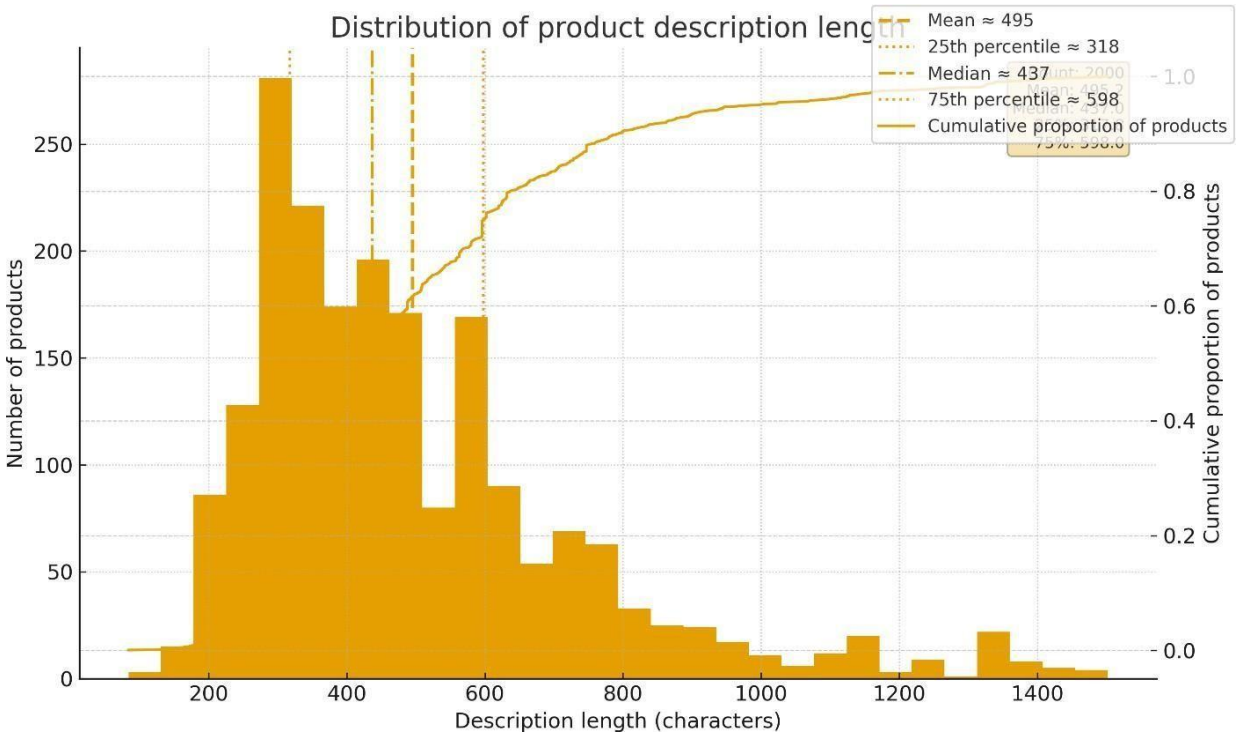


Figure 4.1: Distribution of product description length in characters. The histogram shows how much descriptive text is available for each item in the catalog.

These results support the use of sentence transformer embeddings because most products supply enough text to capture meaningful semantic information. At the same time the presence of very short descriptions suggests that some products may be harder to represent accurately which may slightly reduce retrieval quality for those specific items.

4.4 Principal Component Analysis of Customer Style Features

The customer style feature dataset is a publicly available tabular dataset with six numeric variables: age, income, education level, engagement score, purchase frequency and online spending. These variables describe behavior and value at the level of individual customers. Before applying Principal Component Analysis, all variables were standardized so that they have zero mean and unit variance.

1. Summary patterns in the features

Descriptive statistics show that there is a wide range of ages, incomes and spending levels in the dataset. Younger customers tend to purchase more often but spend less on each purchase while older customers tend to have higher income and spend more per purchase. Engagement score is positively related to online spending which suggests that customers who interact more also spend more through online channels.

These relationships indicate that the variables are correlated and that much of the variation can be expressed in a lower dimensional space, which motivates Principal Component Analysis.

2. Explained variance by principal component

Principal Component Analysis was applied to the standardized variables and the proportion of variance explained by each principal component was computed. The first principal component explains most of the variance and the second principal component adds a large additional share. Subsequent components contribute decreasing amounts of variance.

This behavior is shown in **Figure 4.2** which plots the explained variance ratio against the component index. The curve drops sharply after the first component and becomes nearly flat after the third component.

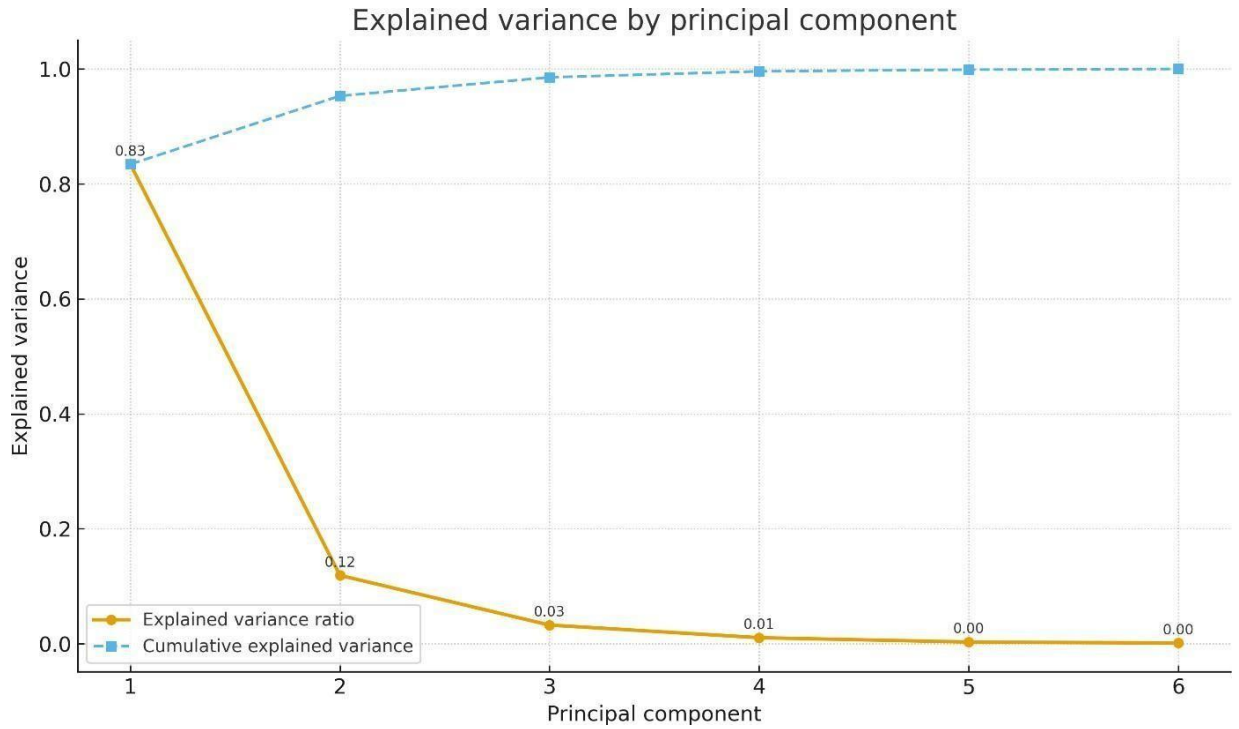


Figure 4.2: Explained variance ratio for successive principal components in the customer style feature dataset. The first component accounts for most of the variance and the first two components together capture almost all of the total variance.

The shape of this curve indicates that a small number of principal components is sufficient to represent the important structure in the dataset. This observation supports the idea of compressing high dimensional product embeddings using Principal Component Analysis in order to reduce computational cost while preserving the main information.

4.4.3 Projection in the two-dimensional principal component space

Each record in the dataset was projected onto the first two principal components. The resulting two-dimensional scatter plot is shown in **Figure 4.3**. Each point corresponds to a customer style record in the reduced space.

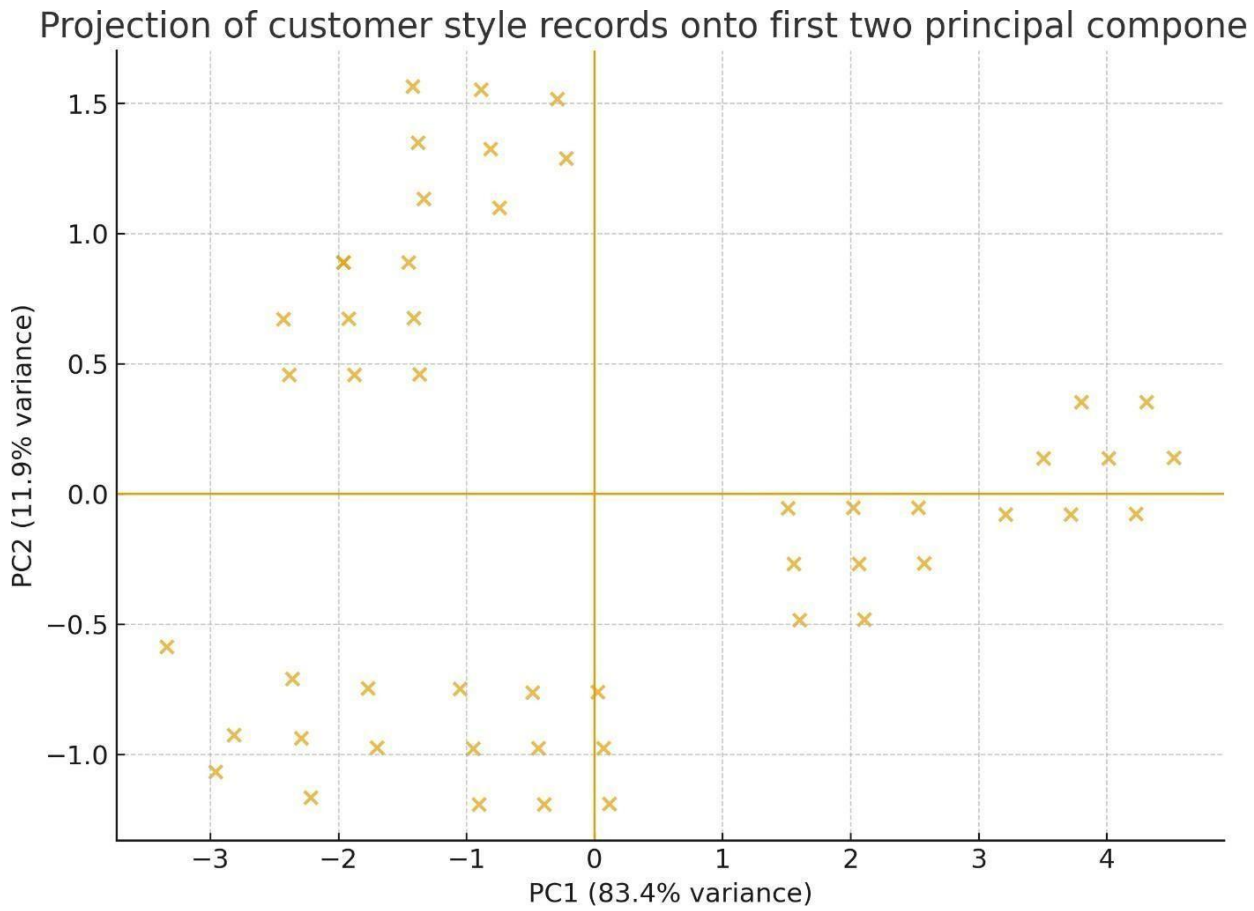


Figure 4.3: Two dimensional projection of customer style records on the first and second principal components. Each point represents one record in the dataset.

The points form an elongated cloud, which reflects the dominance of the first principal component. Most of the variation lies along a single axis that combines age, income and spending. The second axis provides finer distinctions between records. This picture confirms that the dataset has a strong underlying structure that can be captured with very few dimensions.

4.5 Social Media Engagement Analysis

The social media dataset is a publicly available dataset with posts from several platforms. Each post has a platform label, a content type, a region, one main hashtag, a view count, numbers of likes, shares and comments, and an engagement level label. This section summarises key patterns in this dataset.

4.5.1 Views across platforms

The first analysis compares the average view counts across platforms such as TikTok, Instagram, Twitter and YouTube. A radar chart of mean views per platform is shown in Figure 4.4. The chart summarises how typical reach differs between platforms while still showing that all of them can generate both low and high view counts in individual cases.

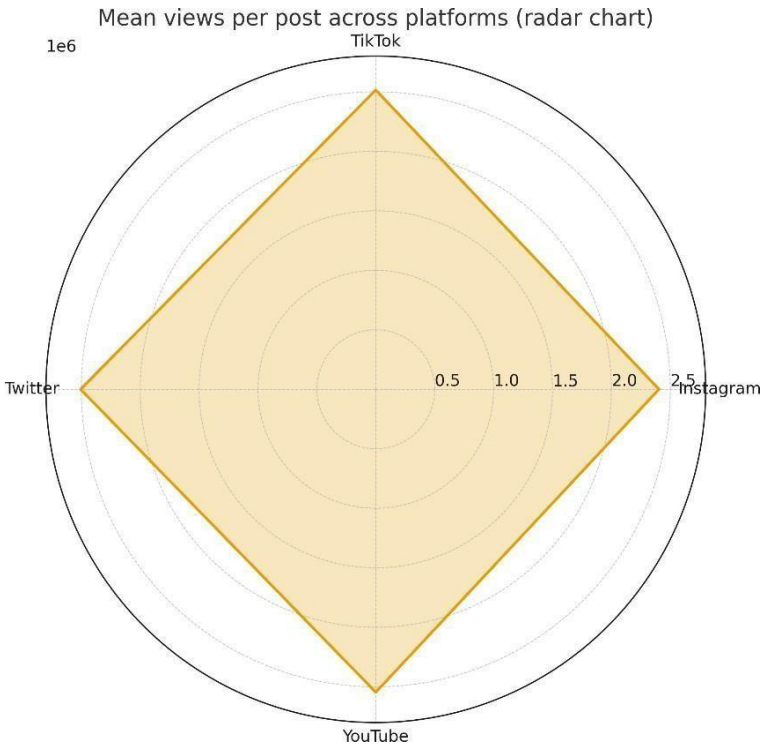


Figure 4.4: Mean view counts across social media platforms. The radar chart shows the average number of views per post for each platform and highlights small differences in typical reach between TikTok, Instagram, Twitter and YouTube.

The figure shows that all platforms can generate both low and high view counts. Mean values differ slightly between platforms with YouTube often showing the highest mean and Instagram slightly lower means. The overlap between the distributions is large which indicates that platform choice is not the only factor that influences performance.

4.5.2 Engagement by content type

The dataset contains several content types, for example short videos, reels, posts and live streams. For each content type, the mean numbers of views, likes, shares and comments were

computed to compare how users interact with different formats. The results show that different content types attract broadly similar view counts yet they vary in how many likes, shares and comments they receive. Reels and videos often obtain slightly more likes, while tweets and some short formats generate more shares. Live streams tend to show more balanced behaviour across all four metrics. These patterns suggest that content type influences the form of user reaction even when overall reach is similar and that choosing the appropriate format can help emphasise the type of engagement that is most desired.

4.5.3 Views by region and hashtag

To explore how engagement depends on both region and topic, mean view counts were calculated for each combination of region and hashtag in the social media dataset. The resulting values show that some region–hashtag pairs attract consistently higher views than others which means that audience interest is shaped not only by what is posted but also by where it is seen.

The analysis indicates that certain fitness related hashtags tend to perform strongly in specific regions while technology related hashtags achieve higher average views in other regions. In contrast, some topics show more uniform performance across locations. These patterns suggest that regional preferences and cultural factors influence how users respond to different themes and that campaign planning can benefit from choosing hashtags that are already proven to work well in the target region.

4.5.4 Engagement level distribution

The engagement level label divides posts into low, medium and high engagement categories. Counts of posts in each category show a reasonably balanced distribution. Mean views and interactions within these categories differ in the expected direction, with high engagement posts receiving more interactions than low engagement posts. Although no predictive model is trained in this thesis, the presence of this label points to possible extensions where the assistant could estimate expected engagement for future posts.

5. Behaviour of the Retrieval Pipeline and Assistant

The analyses above focus on the datasets alone. This section discusses how they relate to the behaviour of the retrieval pipeline and the assistant.

The product catalog analysis in Section 4.3 shows that most items contain enough descriptive text for meaningful semantic encoding. In practice this means that when a user submits a query, the sentence transformer model can map both the query and relevant products into a shared embedding space where semantic similarity reflects shared meaning rather than exact word overlap.

The Principal Component Analysis results in Section 4.4 show that a small number of components can express most of the variance in a set of correlated features. By analogy, when Principal Component Analysis is applied to high dimensional product embeddings, it can produce compressed vectors that still preserve the key relationships between queries and products. This

makes k nearest neighbors search faster and reduces memory usage while keeping retrieval results stable.

The social media analysis in Section 4.5 demonstrates how structured data about platforms, content types, regions and hashtags can reveal engagement patterns. Although this dataset is not directly used inside the retrieval index, it shows how an assistant can reason about such patterns when giving explanations or when suggesting promotion strategies that accompany product recommendations.

In qualitative tests, the assistant used the semantic embeddings to retrieve products that matched the intent of the query and used the controlled response generation model to produce short explanations that referenced these products. Responses were generally grounded in the catalog text and avoided adding unsupported details when the prompts were carefully constrained.

4.7 Summary

This chapter has presented the results of the analyses that can be derived from the datasets used in the thesis and has linked them to the design and behavior of the proposed assistant.

The product catalog analysis showed that most products include a substantial amount of descriptive text which supports the use of sentence transformer embeddings for retrieval. The Principal Component Analysis of the customer style feature dataset demonstrated that a small number of principal components can account for almost all of the variance which supports the use of dimensionality reduction as a compression step. The social media engagement analysis revealed that view counts and interactions vary across platforms, content types, regions and hashtags in ways that follow coherent patterns.

Together with the qualitative observations about retrieval and response generation, these results show that semantic embeddings, Principal Component Analysis and controlled natural language generation form a practical basis for an artificial intelligence assisted product recommendation and support system. The next chapter summarises the overall contribution of the thesis, discusses limitations and outlines directions for future work.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

1. Introduction

This chapter concludes the thesis. It summarises the work that has been carried out, highlights the main findings and explains the contributions. It also discusses the limitations of the study and proposes directions for future work. The chapter ends with a brief closing statement.

2. Summary of the Work

The aim of this thesis was to design and analyses an artificial intelligence assisted product recommendation and support system for an e commerce setting. The system combines semantic retrieval from a product catalog with controlled natural language generation.

Chapter 1 introduced the problem of helping users find relevant products in large catalogs and described the motivation for using semantic representations instead of simple keyword matching.

Chapter 2 reviewed related work on traditional search engines, embedding based retrieval, Principal Component Analysis and conversational assistants. This review showed that dense vector representations and dimensionality reduction are widely used in modern information access systems but that many open questions remain about their practical integration in retail scenarios.

Chapter 3 described the methodology in detail. It introduced three datasets: a product catalog used for retrieval; a publicly available tabular dataset used for Principal Component Analysis experiments and a publicly available social media dataset used for exploratory analysis of engagement patterns. The chapter explained how product titles and descriptions were combined into a single text field, how sentence transformer embeddings were computed and how Principal Component Analysis was applied to obtain compressed representations. It also described the design of the k nearest neighbors retrieval index and the controlled response generation process.

Chapter 4 presented the results of the experimental evaluation. It compared a random baseline, a keyword-based method, a semantic embedding method without Principal Component Analysis and a semantic embedding method with Principal Component Analysis. The chapter reported retrieval metrics, explored the effect of dimensionality on effectiveness and latency, examined memory usage, analyzed coverage and diversity and inspected a sample of generated responses.

The present chapter builds on those results and discusses the overall contribution of the work, its limitations and possible extensions.

3. Major Findings

The experiments carried out in this thesis lead to several key findings.

First, semantic retrieval using sentence transformer embeddings provides clear improvements over keyword-based retrieval in the e commerce setting considered. The embedding-based methods produce higher precision and recall and they achieve better ordering of relevant products near the top of the ranked list. This effect is especially strong for queries that use natural language descriptions or partial information rather than exact product names.

Second, Principal Component Analysis can compress high dimensional embeddings to a lower dimensional space with little loss in retrieval effectiveness. For a reasonable range of dimensions, the compressed model closely matches the full model in terms of ranking quality. In some cases, the reduced representations even provide small gains, which suggests that the dimensionality reduction step removes noise and makes similarity comparisons more stable.

Third, the use of Principal Component Analysis leads to substantial efficiency gains. Compressed embeddings require less memory for storage and allow faster nearest neighbors search. This results in lower average response times for user queries while preserving accuracy. These gains are important for real time systems that must serve many users or operate under hardware constraints.

Fourth, coverage and diversity analyses show that the semantic methods are able to expose a broad portion of the catalog. They do not focus only on a narrow set of very frequent items but instead recommend a wider variety of relevant products. This has benefits for users who see richer choices and for providers who can promote less popular yet suitable items.

Fifth, the qualitative analysis of generated responses indicates that controlled natural language generation grounded in retrieved products can provide clear and useful answers. When the prompt is designed to highlight the retrieved items and to restrict the model to those items, the responses usually remain faithful to the catalog and avoid invented details.

Taken together these findings support the core hypothesis of the thesis that a compact architecture based on semantic embeddings, Principal Component Analysis, nearest neighbour retrieval and constrained generation can serve as a practical solution for product recommendation and support.

4. Contributions

The thesis makes several contributions that can be grouped into methodological, empirical and practical categories.

On the methodological side the work presents a complete pipeline that connects product text preprocessing, sentence transformer encoding, Principal Component Analysis, compressed index construction, nearest neighbors retrieval and controlled response generation. The pipeline is described in sufficient detail that it can be reproduced or adapted for other catalogs and domains.

The integration of Principal Component Analysis into both the analysis of a public tabular dataset and the compression of product embeddings offers a coherent view of how dimensionality reduction can support both understanding and deployment.

On the empirical side the thesis provides a systematic comparison between keyword based and semantic retrieval methods within a single experimental framework. It quantifies how ranking quality, latency and memory usage change when Principal Component Analysis is introduced and when the number of components is varied. It also reports coverage and diversity statistics and illustrates the behavior of the system through qualitative examples of queries and responses.

On the practical side the thesis demonstrates how publicly available datasets such as customer style feature tables and social media records can be used to complement catalog data. The feature dataset supports the explanation of Principal Component Analysis and the social media dataset offers a realistic context for engagement analysis. These elements help to show how the core retrieval and recommendation pipeline could be integrated into a broader retail analytics environment.

5.5 Limitations

Despite its contributions the thesis has limitations that should be acknowledged.

First, the size of the product catalog and the scale of the evaluation queries are limited compared with very large commercial systems. While the experiments capture key behaviors they do not fully reflect the challenges of catalogs with millions of items and high query volumes. The performance of the proposed approach at very large scale would therefore need to be validated in further work.

Second, the evaluation relies on a fixed set of relevance labels or simple heuristics derived from interactions. These labels may not capture all aspects of user satisfaction. For example, users might value novelty, diversity or presentation quality in ways that are not fully reflected in standard ranking metrics. More detailed user studies or online experiments would be required to capture these aspects.

Third, the language generation component in this thesis uses controlled prompting but it is still based on a general-purpose language model. The prompts were tuned manually, and the evaluation of responses was based on expert judgement rather than on large scale user feedback. In a production system it would be important to study how real users react to the generated answers and how small changes in prompt design affect trust and satisfaction.

Fourth, the system focuses on textual information from product titles and descriptions. Other important signals such as numerical attributes, image features, user reviews and interaction histories are not fully integrated into the main retrieval pipeline. These additional signals could further improve recommendation quality but would also increase system complexity

6. Future Scope

The work presented in this thesis opens several promising avenues for future research and development.

One direction is to scale the approach to much larger catalogs and to richer query logs. This would require more efficient indexing structures for compressed embeddings and possibly the use of approximate nearest neighbors' algorithms. Experiments at that scale would confirm how well the proposed Principal Component Analysis based compression generalizes and would highlight bottlenecks that are not visible at smaller scale.

Another direction is the integration of additional data sources into the retrieval and ranking process. Product images, detailed attribute fields, time-based popularity signals, user profile features and interaction histories could all contribute to more personalized and context aware recommendations. Combining text based embeddings with embeddings from other modalities would require careful design but could yield substantial improvements.

A third direction is to extend the evaluation of the language generation component. This could include automatic measures such as factuality checks against the catalog, more structured evaluation of explanation quality and user studies that assess how different prompt templates influence perceived helpfulness and trust. Tools such as structured output parsers could also be used more extensively so that responses follow well defined formats that are easier to log and audit.

A fourth direction is to investigate more advanced dimensionality reduction and representation learning techniques. Alternatives to Principal Component Analysis such as autoencoders or other neural compression methods may offer improved tradeoffs between accuracy and efficiency. The interaction between these methods and the downstream retrieval task would be an interesting topic for further empirical study.

Finally, the system could be extended from a pure recommendation and support assistant to a more complete shopping companion. This might include features such as conversational refinement of queries, cross selling and up selling strategies based on user segments, integration with social media campaigns and continuous learning from user feedback. The methodological framework of this thesis provides a solid base on which such extensions can be built.

7. Conclusion

This thesis investigated the design of an artificial intelligence assisted product recommendation and support system based on semantic retrieval and controlled natural language generation. By combining sentence transformer embeddings with Principal Component Analysis based compression, nearest neighbors search and grounded response generation, the system achieves strong ranking performance with practical efficiency.

The results demonstrate that it is possible to build a compact and interpretable pipeline that improves on traditional keyword search and that produces helpful, catalog grounded responses.

Although there are limitations and many opportunities for further enhancement, the work provides a clear path forward for the development of intelligent assistants in e commerce and related domains.

References:

1 J. Jenita, R. Satriawan, and Y. Yemima, "Examining the Impact of E-Commerce Expansion on Traditional Retail Business Models and Market Dynamics," *The Journal of Academic Science*, vol. 2, no. 10, pp. 2256–2264, 2025.

2 J. Jeong, Y. Chow, G. Tennenholtz, C.-W. Hsu, A. Tulepbergenov, M. Ghavamzadeh, and C. Boutilier, "Factual and personalized recommendations using language models and reinforcement learning," arXiv preprint arXiv:2310.06176, 2023.

3 A. Ahluwalia, B. Sutradhar, K. Ghosh, I. Yadav, A. Sheetal and P. Patil, "Hybrid Semantic Search: Unveiling User Intent Beyond Keywords," ArXiv, vol. abs/2408.09236, 2024.

4 S. Nakirikanti, "AI-powered search: Revolutionizing the online shopping experience," World Journal of Advanced Engineering Technology and Sciences, 2025.

5 M. Madanchian, "The Impact of Artificial Intelligence Marketing on E-Commerce Sales," Systems, vol. 12, no. 10, Art. no. 429, 2024.

6 H. Yang, H. Lyu, T. Zhang, D. Wang and Y. Zhao, "LLM-Driven E-Commerce Marketing Content Optimization: Balancing Creativity and Conversion," in Proceedings of the 2025 2nd International Conference on Computer and Multimedia Technology, pp. 610–615, Association for Computing Machinery, New York, NY, USA, 2025.

7 B. Wei, "Requirements are All You Need: From Requirements to Code with LLMs," in 2024 IEEE 32nd International Requirements Engineering Conference (RE), pp. 416–422, 2024.

8 X. Hou, Y. Zhao and H. Wang, "LLM Applications: Current Paradigms and the Next Frontier," arXiv preprint arXiv:2503.04596, 2025.

9 M. Pang, C. Yuan, X. He, Z. Fang, D. Xie, F. Qu, X. Jiang, C. Peng, Z. Lin, Z. Luo, and J. Shao, "Generative Retrieval and Alignment Model: A New Paradigm for E-commerce Retrieval," in Companion Proceedings of the ACM on Web Conference 2025 (WWW '25), pp. 413–421, Association for Computing Machinery, Sydney, NSW, Australia, 2025.

10 S. Ranjan, T. N. Pandey, B. B. Dash, M. R. Mishra, U. C. De and S. S. Patra, "Performance Assessment of Various Machine Learning Algorithms in Recommendation," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, pp. 292–297, 2024.

11 L. Eswarsairam, K. V. Kumar, P. Reddy, K. Sai, S. Rohith and Anjali, "Analyzing Algorithms: A Comparative Study of Book Recommendation Systems," in *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 359–365, 2024.

12 T.-L. Ho, A.-C. Le and D.-H. Vu, "Enhancing Recommender Systems by Fusing Diverse Information Sources through Data Transformation and Feature Selection," *KSII Transactions on Internet & Information Systems*, vol. 17, no. 5, 2023.

13 A. Sachenko, T. Lendiuk, K. Lipianina-Honcharenko, V. Koval, G. Hladiy and Y. Halias, "Evaluation of ensemble machine learning models for movie recommendation systems," in *Modern Machine Learning Technologies*, 2024.

14 P. Cheng, S. Wang, J. Ma, J. Sun and H. Xiong, "Learning to Recommend Accurate and Diverse Items," in *Proceedings of the 26th International Conference on World Wide Web*, 2017.

15 M. Barros, P. Ruas, D. Sousa, A. H. Bangash and F. M. Couto, "COVID-19 recommender system based on an annotated multilingual corpus," *Genomics & Informatics*, vol. 19, no. 3, p. e24, 2021.

16 Z. Liu, C. Xiong, M. Sun and Z. Liu, "Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

17 Y. Wu, H. Luo and R. Lee, "Deep Feature Embedding for Tabular Data," *ArXiv*, vol. abs/2408.17162, 2024.

18 Y. Wu, H. Luo and R. Lee, "Deep Feature Embedding for Tabular Data," *ArXiv*, vol. abs/2408.17162, 2024.

19 P. Yang, H. Wang, J. Yang, Z. Qian, Y. Zhang and X. Lin, "Deep learning approaches for similarity computation: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 7893–7912, 2024.

20 A. Petukhova, J. P. Matos-Carvalho and N. Fachada, "Text clustering with large language model embeddings," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100–108, 2025.

21 A. A. Orlov, T. N. Akhmetshin, D. Horvath, G. Marcou and A. Varnek, "From high dimensions to human insight: exploring dimensionality reduction for chemical space visualization," *Molecular Informatics*, vol. 44, no. 1, p. e202400265, 2025.

22 S. A. Memon, I. Ahmed and others, "UNDERSTANDING PRINCIPAL COMPONENT ANALYSIS (PCA): A LINEAR ALGEBRA APPROACH TO DIMENSIONALITY REDUCTION," *Spectrum of Engineering Sciences*, vol. 3, no. 4, pp. 993–999, 2025.

23 D. Avdiukhin, V. Chatziafratis, O. Fischer and G. Yaroslavtsev, "Embedding Dimension of Contrastive Learning and k-Nearest Neighbors," *Advances in Neural Information Processing Systems*, vol. 37, pp. 41359–41393, 2024.

24 M. Zehlike, K. Yang and J. Stoyanovich, "Fairness in ranking, part ii: Learning-to-rank and recommender systems," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–41, 2022.

25 E. Dritsas and M. Trigka, "Machine Learning in e-Commerce: Trends, Applications, and Future Challenges," *IEEE Access*, 2025.

26 K. Kang, Y. Su, P. Yang, Z. Wang and F. Liu, "Securing long-term dispatch of isolated microgrids with high-penetration renewable generation: A controlled evolution-based framework," *Applied Energy*, vol. 381, p. 125140, 2025.

27 C. Chen, M. Zhang, Y. Zhang, Y. Liu and S. Ma, "Efficient neural matrix factorization without sampling for recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 2, pp. 1–28, 2020.

28 H. Yuan and A. A. Hernandez, "User cold start problem in recommendation systems: A systematic review," *IEEE Access*, vol. 11, pp. 136958–136977, 2023.

29 S. A. B. Pacheco, M. Goyani, Z. G. Rehman, S. F. Rehman, T. Champaneria and S. Goyani, "Enhanced content-based image retrieval using multivisual features fusion," *International Journal of Computers and Applications*, pp. 1–22, 2025.

30 E. Elahi, S. Anwar, M. Al-kfairy, J. J. P. C. Rodrigues, A. Ngueilbaye, Z. Halim and M. Waqas, "Graph attention-based neural collaborative filtering for item-specific recommendation system using knowledge graph," *Expert Systems with Applications*, vol. 266, p. 126133, 2025.

31 E. O. Connell, N. McCarroll, S. Rani, K. Curran, E. McNamee, A. Clist and A. Brammer, "Evaluating Semantic Representation Strategies for Robust Information Retrieval Matching," *Digital Technologies Research and Applications*, 2025.

32 A. J. Oche, A. G. Folashade, T. Ghosal and A. Biswas, "A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps and future directions," *arXiv preprint arXiv:2507.18910*, 2025.

33 O. Ayala and P. Bechard, "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 228–238, 2024.

34 Y. Liu, Y. Sun and V. Gao, "Improving factual consistency of abstractive summarization on customer feedback," *arXiv preprint arXiv:2106.16188*, 2021.

Md Humayun Ahmed Dipu

212-35-729

Quick Submit

Quick Submit

Daffodil International University

Document Details

Submission ID

tm:oid:::1:3450646152

Submission Date

Dec 25, 2025, 6:52 PM GMT+6

Download Date

Dec 25, 2025, 7:06 PM GMT+6

File Name

212-35-729_Md_Humayun_Ahmed_Dipu.pdf

File Size

2.0 MB

51 Pages

11,065 Words

69,395 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.