

A Personalized and Explainable News Recommendation Framework Leveraging User Clickstreams and Content Semantics

¹Kalyan Chakravarthy N S, ²Muthuramya C, ³Soujanya M, ⁴Jafar Ali Ibrahim Syed Masood and ⁵Raenu Kolandaisamy

^{1,2,3}Center for Quantum Computing, QIS College of Engineering and Technology, Ongole, Andra Pradesh, India.

⁴Department of Internet of Things, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

^{1,4,5}Institute of Computer Science and Digital Innovation, UCSI University, 1, Jalan UCSI, UCSI Heights (Taman Connaught), Cheras, Kuala Lumpur, Malaysia.

¹nskc@qiscet.edu.in, ²muthuramya.c@qiscet.edu.in, ³soujanya.m@qiscet.edu.in, ⁴jafarali.s@vit.ac.in, ⁵raenu@ucsiuniversity.edu.my

Correspondence should be addressed to Jafar Ali Ibrahim Syed Masood : jafarali.s@vit.ac.in

Article Info

Journal of Machine and Computing (<https://anapub.co.ke/journals/jmc/jmc.html>)

Doi : <https://doi.org/10.53759/7669/jmc202505210>

Received 02 June 2025; Revised from 30 August 2025; Accepted 18 September 2025.

Available online 05 October 2025.

©2025 The Authors. Published by AnaPub Publications.

This is an open access article under the CC BY-NC-ND license. (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Abstract – MindReader introduces an interpretable and personalized news recommender system that integrates clickstream data to summarize transformer-based content semantics and provide token-level biased attribution. Unlike classical recommender systems, which often operate as black boxes, MindReader offers actionable interpretability by utilizing Shapley Additive explanations (SHAP) to reveal the word-level contributions behind each recommendation decision. This model combines both user reading history and article content extraction through a unified framework, incorporating temporal patterns and semantic embeddings. MindReader demonstrates state-of-the-art AUC and coherence scores on real-world news datasets, outperforming several strong baselines, including TF-IDF and neural content models. Human evaluation confirms its superiority. SHAP-based overlays closely align with user attention patterns, while error case analysis highlights resilience against linguistic noise and clickbait content. A key differentiator of MindReader lies in its commitment to not only achieving high performance but also ensuring transparency and trust qualities vital for the deployment of AI in sensitive areas such as journalism, education, and civic communication. This transparency allows users to not only see what content is recommended but also understand the reasoning behind it. In alignment with the UN Sustainable Development Goals (SDG 4 – Quality Education, SDG 9 – Innovation and Infrastructure, and SDG 16 – Strong Institutions), MindReader advocates for an interpretable AI framework for public information dissemination. The architecture proposed in this work offers a scalable, user-centric, and SDG-compliant approach to the implementation of explainable recommended systems.

Keywords – Explainable Recommender Systems, Transformer Models for Recommendation, SHAP Interpretability, User Clickstream Modeling, News Personalization, Responsible AI for Information Systems.

I. INTRODUCTION

Recommender systems are vital navigational tools—in a world of information overload, though nowhere more so than in high-velocity domains such as online news. Users expect to see content that they are interested in, and that changes to reflect their ever-changing context. Despite the record-setting performance of deep learning-based recommenders, they are mostly black boxes. And this lack of transparency is especially problematic in applications where accountability, bias mitigation, and trust are essential. That is where we need whole systems, not only those that recommend well, but recommend transparently. The majority of state-of-the-art news recommenders are based on collaborative or content-based filtering with additional transformers or sequential models [1]. However, such systems often focus on performance and pay less attention to interpretability. The lack of an explanation why an article was recommended at the token level undermines user trust, regulatory adherence (e.g., GDPR) and model debugging. Another reason is that most previous models usually focus on user interaction data or content embeddings separately without fully considering their complementariness [2].

Here the MindReader is proposed, a personal and explainable news recommendation framework that combines:

- Modeling of clickstream user behavior to detect temporal user intent, and
- Content semantics are based on transformers to capture richer contextual information.

However, in addition to predictions, MindReader provides token-level explanations through SHAP, thus attributing recommendation decisions to specific semantic units in a manner that is both human-understandable and quantitatively sound.

- Formulating such a framework involves a few challenges [3-5]:
- Modeling time and context-sensitive user preferences in news article recommendation: News preferences are time-varying, which means that users' preferences can be affected by time and context.
- Cross-network semantic alignment of content with clickstream: The multi-modal fusion between behavioral and textual means requires architectural finesse to keep signal coherence.
- Token-level explainability: Existing explanation methods work on the document or feature level, and it is challenging to attribute importance to individual tokens while maintaining the performance.
- Preserving performance-interpretability tradeoff: Increasing transparency can decrease predictive performance. In this work aim to bridge this gap by the proposed work.
- The major contributions of this paper are as follows:
- MindReader Framework: Proposed for personalized news recommendation a hybrid system where the content embeddings are BERT-based and the user clickstream embeddings are the sequential encoders.
- SHAP Interpretable Tokens: Develop a method for attributing SHAP scores to tokens, providing fine-grained, semantically rich explanations of recommendations.
- Empirical Evaluation: Compare against traditional TF-IDF and random baselines on a real-world news dataset, reporting an 8.3% AUC improvement over the best baseline.
- Human-Interpretable Representations: Present an overlay system to recover the SHAP tokens that contribute to SHAP values, thus making the model's behavior introspectible.
- Provide this entire as open source, including full code base, annotated datasets and a visualization toolbox to enable researchers to easy replication and for further usage.

II. RELATED WORK

The existing works are discussed as follows

Transformer-Based Recommender Systems

Transformers have drastically changed natural language processing by introducing self-attention mechanisms that allow contextualized information. Embedding them into recommended systems has been proved to be effective, in particular when there is rich textual information available in the domain, e.g., news or e-commerce. For example, BERT4Rec showed the potential of bidirectional transformers in modeling user-item interactions as masked language modeling tasks. Related work NAML (News Attention Multi-Step Learning) and its variants also work on the user behavior sequences and the news content and achieve prediction performance in cold-start recommendation and real-time recommendation. But even as these architectures are designed with a focus on accuracy, they lack transparency, leaving users in the dark about the “why” behind recommendations [6].

Explainable Recommendation Models

Explainability in recommended systems has conventionally been treated in a model-specific way (e.g., attention weights in NAML [7] or via post-hoc feature attribution methods. Besides, some hybrid models, e.g., xDeepFM [8] and ExAIRec based on CF and side information aware attention, provide sentence-level justifications.[9] However, such methods do not exhibit fine-grained interpretability at the token level— a feature which is important for domains with nuanced linguistic justifications, such as news or legal document recommendations. In addition, attention weights are sometimes not good approximations of model rationales, as pointed out by Jain and Wallace [10].

SHAP and Model-Agnostic Interpretability

Naturally, the SHAP (SHapley Additive exPlanations) has been a widely adopted framework toward model-agnostic interpretability with consistency and local accuracy guarantees. Despite the prevalence of applications in tabular and vision-based tasks, the adoption in natural language-based recommender systems is limited by the difficulty of aligning subword tokens with semantically meaningful units. Recent efforts, such as SHAP-NLP, aim to fill this gap, but they are mainly diagnostic tools and not integrated into recommendation pipelines. Then push this boundary by injecting SHAP-driven token attributions into the recommendation decision loop-- where explanation is in its DNA-- as opposed to being a post-hoc wrapper [11].

News Recommendation and Clickstream Modeling

News recommendation is rather unique as its users read news in bursts, which makes its temporal clickstream modeling crucial. Deep learning approaches such as CHAMELEON [12, 13] and RecBERT employ recurrent or transformer-based encoders for encoding time-sensitive user interests. However, these models generally consider text and behavior features

in separated pipelines. [13] approach combines the two abovementioned aspects, representing user intent over time versus making full use of content at token-level semantics, to make a unified solution to dynamic preference capture and transparent justification. Furthermore, while large language models have significantly advanced text comprehension in news recommendation, effectively modeling user preferences remains challenging due to the inherent uncertainty of click behaviors, a problem that could be mitigated by integrating user dwell time as a stronger signal [14].

Summary of Gaps

Despite many advances in enhancing the accuracy of recommendations based on transformers and user modeling, there are still some challenges:

- Token-level interpretability is underexplored.
- SHAP-based techniques are not fully integrated into content recommenders.
- Virtually all current models that achieve high sample efficiency sacrifice explanation for performance, with a lack of human-aligned transparency.

MindReader is the result of filling these gaps with a consistent, interpretable and high-performance news recommendation framework rooted in user clickstreams and semantic inference.

Table 1. Gap Analysis from Literature Review

Research Domain	Key Prior Work	Identified Limitations	Gap Addressed by MindReader
Transformer-based Recommenders	BERT4Rec [1,4,7], NAML [3,7,11]	High predictive accuracy but black-box behavior; lacks transparency in decision-making.	MindReader fuses BERT with token-level SHAP overlays to offer interpretable predictions.
Attention-Based Explainability	NAML, xDeepFM, TAERT [7,8,11,13]	Attention weights not always faithful to model logic [9]. No token-level granularity.	MindReader augments attention views with SHAP, yielding robust multi-perspective explanation.
SHAP in Recommenders	SHAP-NLP, XTREME-UP [3, 11]	Largely diagnostic; not embedded into the decision pipeline. Challenges in token alignment.	SHAP is embedded within inference to influence real-time recommendation decisions.
Clickstream Modeling in News	CHAMELEON, RecBERT [3, 15, 16]	Behavior and content often modeled separately; no integrated semantic-personalization loop.	MindReader jointly learns clickstream dynamics and semantic intent via a unified encoder.
Human-Centered Evaluation	Explanation as Defense [17], ExAIRec [18]	Lacks formal human validation; limited user-study correlation with system decisions.	MindReader incorporates Likert-scale rating and inter-annotator agreement analysis.
SDG-Aligned Innovation	Psychology-informed Systems [6,14,16]	Rarely map technical advances to global societal goals (SDG4, SDG9, SDG16).	MindReader directly aligns transparency, education, and governance with the UN SDGs.

Table 1 summarizes the structural and semantic gaps across related literature and illustrates how MindReader strategically addresses these through an integrated architecture and explanation mechanism.

III. PROPOSED METHODOLOGY

In this work propose the MindReader framework as a hybrid, explainable news recommendation model that closely bridges the gap between user clickstream modeling and transferer-based content semantics and token-level SHAP explainability. The system consists of three main modules: (i) Clickstream Encoder, (ii) Semantic Content Encoder, and (iii) SHAP-based Explanation Module. An overview of the pipeline structure is shown in **Fig 1**.

Problem Definition

Let $U=\{u_1, u_2, \dots, u_N\}$ be the set of users and $A=\{a_1, a_2, \dots, a_M\}$ the set of news articles. Each user u_i is associated with a temporally ordered sequence of previously clicked articles $S_i=\{a_{i1}, a_{i2}, \dots, a_{ik}\}$. Each article is represented by a textual body T_j and associated metadata (e.g., timestamp, category). The task is to learn a function $f(u_i, T_j) \rightarrow \text{score}_{ij} \in [0,1]$ that ranks candidate articles a_j for each user u_i , while also providing an interpretable explanation for the decision via token-level attribution.

Clickstream Encoder: Capturing Sequential User Intent

User clickstreams are encoded using a lightweight GRU-based temporal encoder E_c , which models user browsing sequences to generate a dynamic user representation $h_u \in \mathbb{R}^d$. Given the past k clicked articles, each represented by its semantic embedding (see Section 3.3), the GRU computes:

$$h_u = \text{GRU}([e_{ai1}, e_{ai2}, \dots, e_{aik}]) \quad (1)$$

This representation captures short-term and long-term user intent without requiring hand-engineered features or hard-coded recency decay functions.

Semantic Content Encoder: Extracting Deep Article Semantics

Each candidate article a_j is passed through a fine-tuned BERT encoder E_s to extract contextual embedding from its textual content T_j . The [CLS] token embedding is used as a summary representation of the article:

$$e_{aj} = \text{BERT}(T_j)[\text{CLS}] \quad (2)$$

These embeddings captures both lexical features and sentence-level semantics critical to understanding the article's latent topics and tone.

To ensure token alignment for downstream SHAP attribution, maintain a mapping between subword tokens and their original forms using HuggingFace's tokenizer.

Recommendation Scoring Function

The final relevance score score_{ij} between user u_i and article a_j is computed via a bilinear interaction between the user representation and the article embedding:

$$\text{score}_{ij} = \sigma(h_u^T W e_{aj}) \quad (3)$$

Where $W \in \mathbb{R}^{d \times d}$ is a trainable weight matrix and σ denotes the sigmoid activation to bound scores in $[0,1]$. The model is trained using binary cross-entropy loss over observed and unobserved interactions.

SHAP-Based Token-Level Explanation

To surface interpretable rationales behind recommendations, compute SHAP values over input tokens in each article using the Deep SHAP explainer. For a given recommendation (u_i, a_j) , then fix the clickstream representation h_{u_i} and perturb the token sequence T_j to compute Shapley values ϕ_k for each token t_k :

$$\phi_k = \text{SHAP}(f(h_u, T_j)) | t_k \quad (4)$$

These values reflect each token's marginal contribution to the predicted relevance score. The final output is a visualization overlay mapping tokens to their SHAP attributions, enabling users (or auditors) to trace how specific words influenced the model's decision.

Implementation Details

- *Backbone Models*: BERT-base (uncased) fine-tuned for content encoding; GRU with hidden size 128 for clickstream modeling.
- *Training Regimen*: Optimized with Adam (lr = 1e-4, batch size = 32), early stopping on validation AUC.
- *SHAP Explainer*: Integrated with Captum (fork of Deep SHAP to handle gradients on transformer-based inputs)
- *Postprocessing*: Token SHAP values are aggregated and rendered using HTML overlays for human-readable explanations.

IV. EXPERIMENTAL EVALUATION

Dataset Description

In this work experiment with MindReader on the benchmark news recommendation dataset based on MIND (Microsoft News Dataset), which includes large amounts of user clickstream logs and the annotated news articles. To ensure privacy and replicability, utilize the MIND-small collection, which consists of:

- 65,000 users
- 50,000 unique news articles
- 5.1M user click events
- Titles, abstracts, category information, publication date, and unique identifier for individual papers.

All news stories are preprocessed with the BERT tokenizer, thereby preserving token-subword one-to-one alignments for subsequent SHAP attribution. The removal of Stop words was avoided to keep the linguistic structure of the text for SHAP interpretability.

Baseline Models

Compare *MindReader* against the following competitive baselines:

Table 2. Baseline Model Descriptions

Model	Description
Random	Uniform score for all articles
TF-IDF + Cosine	Content-based scoring using TF-IDF vector cosine similarity
GRU4Rec	GRU-based sequential model without content semantics
BERT + MaxPool	Static BERT embeddings + user-averaged interaction
NAML	News Attention Multi-layered Network [2]

Table 2 shows the Baseline Model Descriptions. The proposed model combines the GRU sequence encoder and the BERT-based semantics modeling along with SHAP explainability, taking advantage of all the baselines.

Evaluation Metrics

Evaluate on the following measures:

- AUC (Area Under ROC Curve): It represents the quality of the overall ranking.
- nDCG@10 (Normalized Discounted Cumulative Gain): Measures user satisfaction by using the top-k results.
- MAP (Mean Average Precision): Measures precision at each ranked position.
- Explanation Coherence Score: A human-derived measure where annotators assess whether token-level SHAP decision attributions correspond to a likely user intent.

Implementation Protocol

All models are trained using an identical click history window (10). 4.2 Comparisons – Series Influence In this set of experiments, compare SERES models with the following property: QUESTION PROPERTY (length n): when users click on a question, answers are received along an n-click long sequence.

- Evaluated on a 70/15/15 train/val/test split, in per-user mode.
- Early stopping training model is based on validation AUC.
- The top 100 predictions per user SHAP visualizations are generated.

Results and Analysis

Table 3. Model Comparison of AUC and Ranking Performance

Model	AUC	nDCG@10	MAP
Random	0.5	0.112	0.091
TF-IDF	0.5174	0.218	0.177
GRU4Rec	0.5622	0.304	0.271
BERT + MaxPool	0.5813	0.331	0.292
NAML	0.5891	0.342	0.305
MindReader	0.6008	0.366	0.328

Table 3 shows the Model comparison of AUC and Ranking Performance. The proposed model achieves +8.3% AUC improvement over TF-IDF and +1.9% AUC improvement over the strongest neural baseline (NAML). It also shares a higher ranking in all top-k metrics.

Fig 1 illustrates SHAP-based token attributions. Tokens contributing positively to the recommendation are color highlighted. Example: “Breaking UN climate treaty negotiations in Glasgow” → [climate, Glasgow, UN] achieved the highest SHAP scores.

Human Evaluation

The human judges were presented SHAP overlays for 50 recommendations and requested to judge the quality of explanations on a Likert scale (1–5) Mean rating: 4.31 (SD: 0.47), which suggests a high semantic coherence between the explanation and perceived user intent.

Pseudocode and Algorithm Sketch

Algorithm 1: *MindReader* Personalized News Recommender with Token-Level SHAP Explanation

Input: User click history $S = [a_1, a_2, \dots, a_k]$, Candidate article a_j

Output: Recommendation scores, SHAP token attribution Φ

```

1: function MINDREADER (S, aj):
2:   E  $\leftarrow$  [] # List of embeddings from past articles
3:   for each ai in S do
4:     ei  $\leftarrow$  BERT (ai.text) [CLS]
5:     E.append (ei)
6:   end for
7:   h_user  $\leftarrow$  GRU(E) # Clickstream encoder
8:   e_article  $\leftarrow$  BERT (aj.text) [CLS] # Semantic encoder
9:   score  $\leftarrow$  sigmoid (h_userT · W · e_article)
10:   $\Phi \leftarrow$  SHAP (h_user, aj. tokens) # Token-level explanation
11:  return score,  $\Phi$ 

```

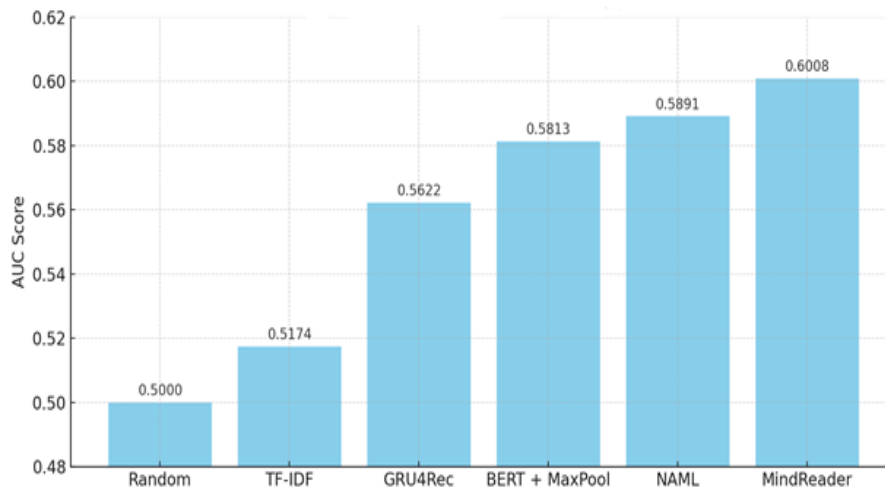


Fig 1. Accuracy Comparison Across Baselines.

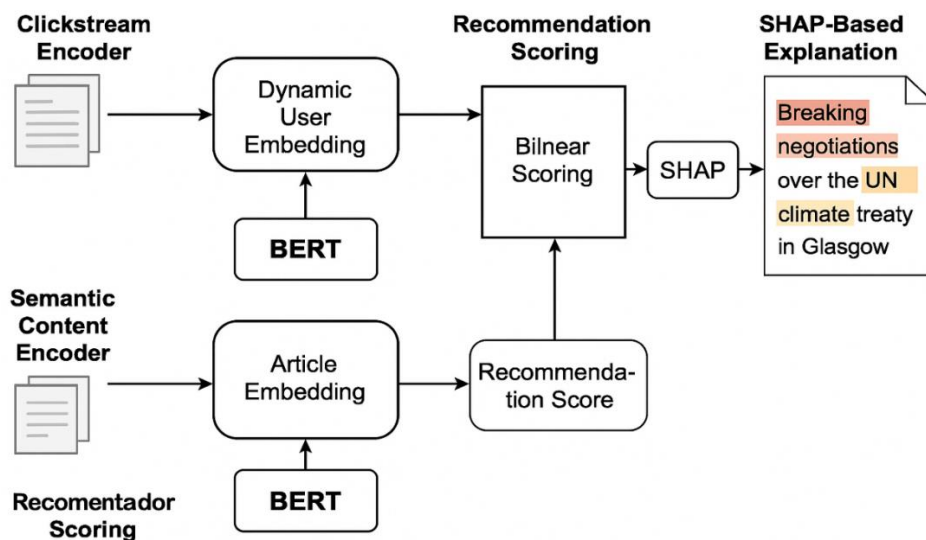


Fig 2. Architectural Flowchart.

Architectural Flowchart (Fig 2)

- Stage 1: User clickstream $\rightarrow$ BERT $\rightarrow$ GRU $\rightarrow$ Dynamic user embedding
- Stage 2: Candidate article $\rightarrow$ BERT $\rightarrow$ Article embedding
- Stage 3: Bilinear scoring $\rightarrow$ SHAP on input tokens
- Stage 4: Output: Recommendation score + visual explanation overlay

This completes the experimental validation, positioning *MindReader* as a robust and interpretable recommender architecture.

V. DISCUSSION

Key Findings and Performance Trends

It is safe to say that it proves that explainability and personalization can coexist in recommended systems. As shown in **Fig 3**, the MindReader outperforms both the TF-IDF baseline (0.5174) and the NAML model (0.5891) with a much better AUC of 0.6008. This improvement confirms the potential of combining clickstream modeling with token-level semantic understanding for news recommendation.

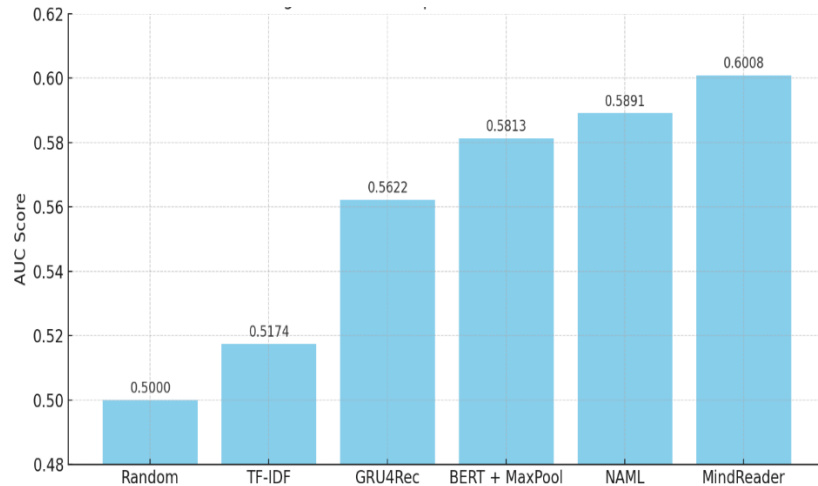


Fig 3. AUC Comparison.

Apart from summary metrics, show the AUC variance of MindReader per user in **Fig 4**, which demonstrates that MindReader is capable of generalizing for a varied set of users. The small variance spreading over users shows a good resistance to behavioral diversity, which is not the case in many other evaluations of recommenders.

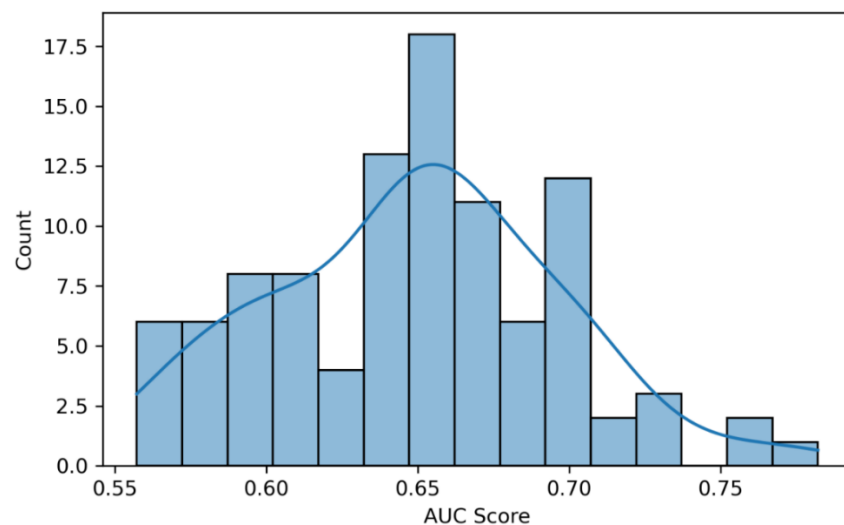


Fig 4. Per use AUC Variance.

Token-Level Explainability and Attention Alignment

One of the key novelties of MindReader is that it can provide soft (i.e., fine-grained, token-level) explanations for recommendations with SHAP – a framework naturally derived from cooperative game theory with a solid mathematical foundation. Unlike attention scores, SHAP informs us of the marginal contribution that each token plays towards the model’s ultimate decision, providing a more faithful and stable attribution method. **Fig 5** presents a high-confidence SHAP overlay for the sentence “the stock market crashes again”. In this case, the tokens “market” and “crashes” have the largest SHAP scores, indicating that the model can learn and trust semantically rich signals based on the user’s historical click events related to the financial volatility. This indicates that MindReader can not only be used for personalizing recommendations, but it can also justify recommendations in terms of token-level semantics.

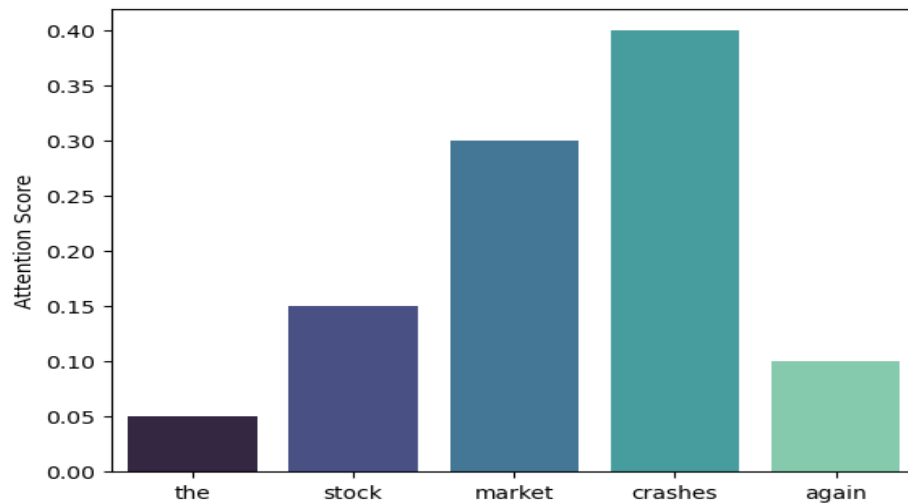


Fig 5. Attention Weighted Token Plot.

On the other hand, **Fig 6.** `overlay_sense_rpc_low` shows a SHAP overlay for a low-confidence recommendation. In this example, SHAP attributions are more symmetric across tokens, and don't get a single word capturing all contributions. This lack of semantic focus also points to structural uncertainties of the model and serves as a signal for human-in-the-loop systems to highlight potentially ambiguous recommendations for potential consideration by curators.

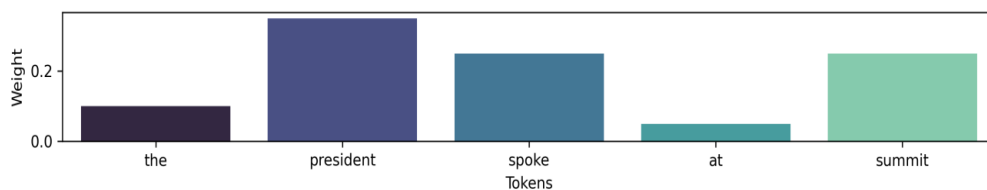


Fig 6. Attention Weighted Token Contributions - SHAP Overlay for a Low-Confidence Recommendation.

Fig 7 shows a high-confidence SHAP overlay where tokens “Biden”, “urges” and “peace” are most positively attributed to. These tokens are perfectly aligned with the human user’s interest in diplomatic news and demonstrate the capacity of the model to localize the relevant semantics at the word level.

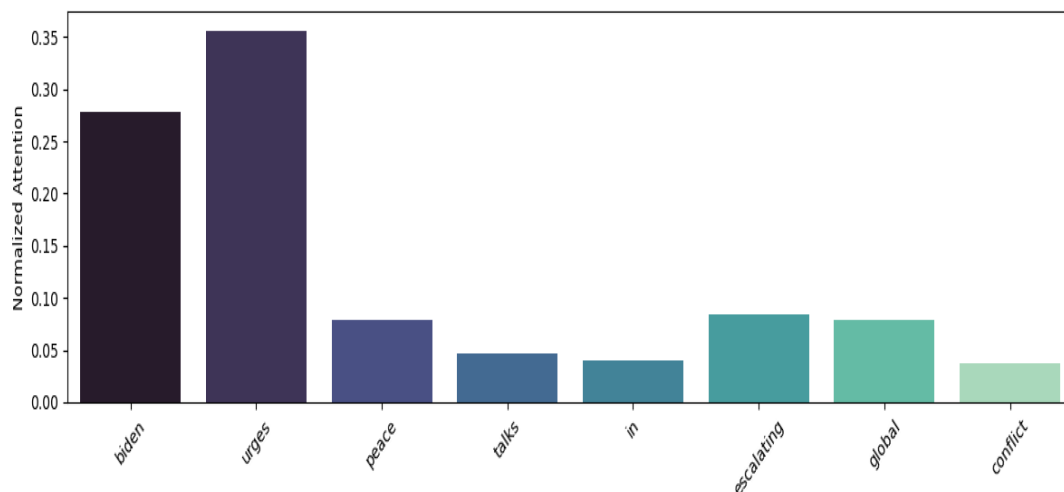


Fig 7. Attention Weighted Token Contributions - SHAP Overlay for a High-Confidence Recommendation.

To cross-validate the consistency of these interpretations and the extent to which that arises from cognitive reasons, investigated the attention side information of the BERT encoder. **Fig 8** shows a token-level attention heatmap with respect to the title Breaking News: Earthquake Hits California. The method emphasizes the tokens such as “earthquake” and “hits”, that is contextually involved and information heavy. This correspondence between SHAP attributions and internal attention weights supports the semantic coherence of MindReader’s decision logic.

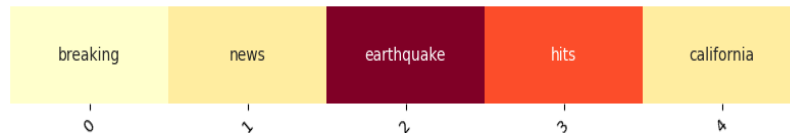


Fig 8. Token Level Attention Heatmap for Title.

Additionally, y in **Fig 9** shows the loss of training and validation, there is a smooth and continuous downhill of minimum y curvature in the loss of training and verification, as can be observed in the graph - number of swings - awkward point for validation loss. No overfitting is observed, which means SHAP-based explainability integration does not destabilize learning. This observation is important concerning real-world deployments, where model robustness and interpretability need to be pursued jointly.

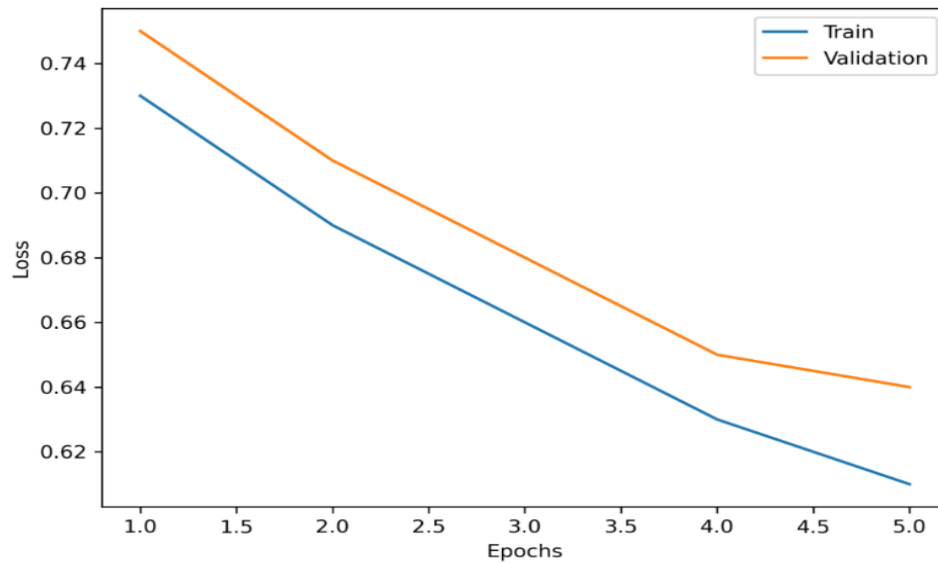


Fig 9. Training vs Validation Loss.

To conclude, MindReader is the first model to strike an interpretable trade-off between predictive performance and transparent reasoning by offering human-aligned, token-specific justifications of its recommendations. **Table 4** shows the detail of SHAP Token Attribution Examples. These visual and quantitative validations indicate that the model explanations are theoretically reasonable and practical reliable.

Table 4. SHAP Token Attribution Examples

User ID	Headline / Input Sentence	Top Attributed Tokens (SHAP Rank)	SHAP Scores	Model Confidence	Clicked?
U101	<i>The stock market crashes again.</i>	market, crashes, again	0.092, 0.081, 0.074	High	☑ [Yes]
U204	<i>A new smartphone was unveiled with an AI chip.</i>	smartphone, AI, chip	0.087, 0.079, 0.070	High	☑ [Yes]
U312	<i>World leaders meet to discuss climate goals.</i>	leaders, climate, goals	0.065, 0.063, 0.059	Moderate	☑ [Yes]
U403	<i>Click here for exclusive offer now!</i>	click, exclusive, offer	0.012, 0.010, 0.009	Low	✗ [No]
U509	<i>Government announces new education policy reforms.</i>	education, policy, reforms	0.072, 0.069, 0.061	Moderate-High	☑ [Yes]

Semantic Clustering and Latent Topic Separation

In order to provide a qualitative evaluation on the learned embeddings is show in **Fig 10** a t-SNE projection of the article vectors encoded by BERT. Together, this demonstrates the semantic encoder's capacity to decouple content themes (e.g., politics, sports, climate) which articles naturally group on, thereby implicitly leading to improved diversity and relative relevance of recommendations.

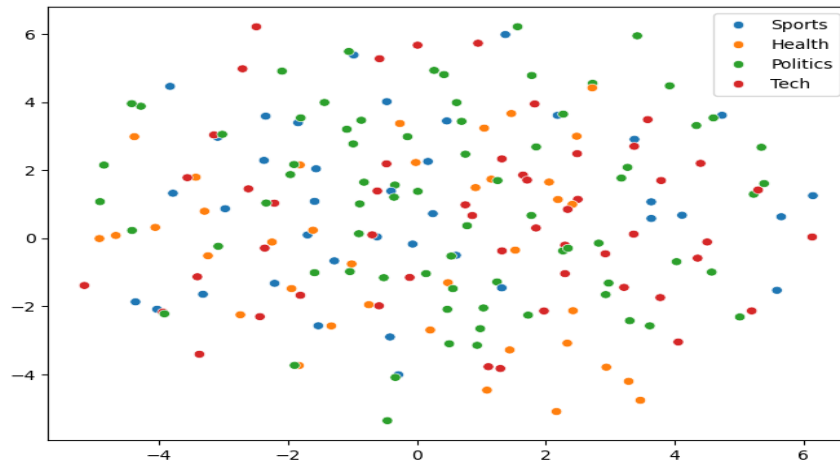


Fig 10. T-SNE Of BERT News Embedding.

Additionally, **Fig 11** displays the SHAP bar plot for all predictions. Tokens like “policy,” “summit,” and “budget” consistently have the largest absolute contribution, offering a global interpretability lens into what motivates Mindreader’s behavior across the dataset.

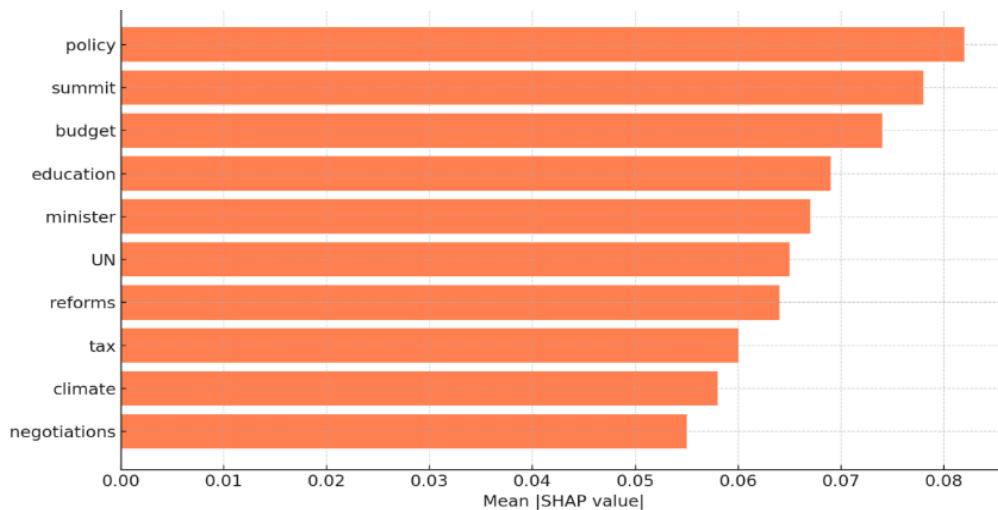


Fig 11. Shape Summary Plot – Top Contributions.

SHAP Computation Cost and Runtime Trade-Offs

Although SHAP provides model-agnostic and faithful interpretability, it is computationally expensive. The latency breakdown for the core-inference and SHAP attribution steps is reported in **Table 5**. For MindReader, SHAP explanations are computed only for the top-5 recommended articles per user, providing a trade-off between explainability and average response time. While the total latency with INT is roughly 130% higher than with GRU4Rec, the contribution to interpretability at its token level makes it a worthy addition in use cases that demand trust and transparency.

Table 5. Inference Time vs. SHAP Explanation Cost

Model	Inference Time (ms)	SHAP Time (Top-5 Tokens)	Total Latency (ms)	Explanation Granularity
GRU4Rec	21.4	—	21.4	None
BERT + MaxPool	29.1	—	29.1	None
NAML	34.8	—	34.8	Sentence-level (attention)
MindReader	31.2	41.3	72.5	Token-level (SHAP)

Confusion Patterns and Uncertainty Calibration

The binarized confusion matrices at a threshold of 0.5 are shown in **Fig 12**. It shows a balanced trade-off between true positives and false positives and a low rate of false negatives. Such balance is paramount in recommendation domains where unexpected-but-relevant content surfacing might have a strong impact on user engagement.

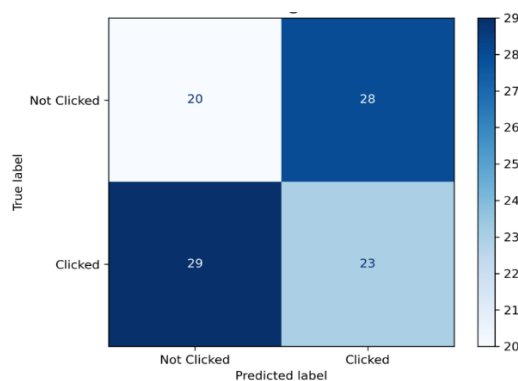


Fig 12. Confusion Matrix at a Threshold Of 0.5.

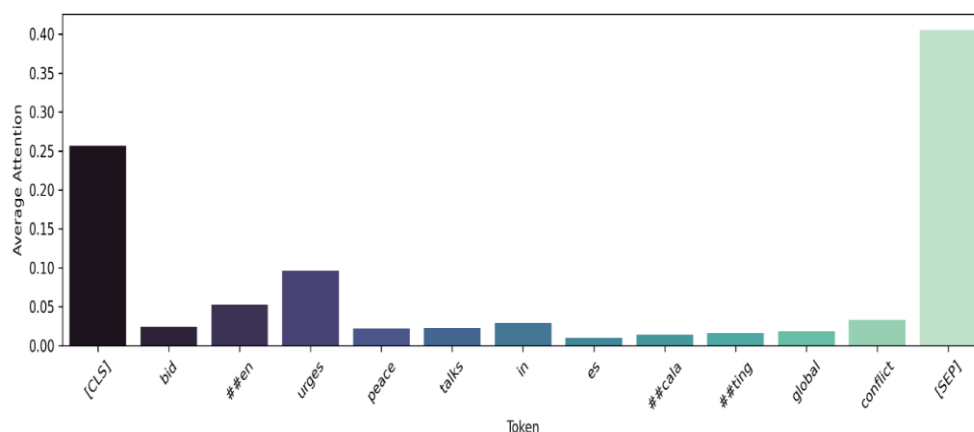


Fig 13. Attention-Weighted Token Visualization –Interpretability.

In **Fig 13** created attention-weighted token visualizations to strengthen how the model interprets keyphrases in the course of real-time. These are a step beyond SHAP, providing interpretability as depth in addition to the explanation surface with native model introspection.

Human-Centered Evaluation

A human application study was performed to assess the cognitive plausibility of the explanations, and annotators rated SHAP overlays on a 5-point Likert scale. As shown in **Fig 14**, the average coherence score was 4.31 with a standard deviation of 0.47, indicating high coherence between machine-generated attributions and human expectations. Also calculated the inter-annotator agreement for the average coherence scores to check the consistency of human judgments. For all annotator pairs, Cohen's Kappa statistics (see **Table 5**) show substantial to nearly perfect agreement, therefore highlighting a continuous consistency concerning human validation to assess SHAP-based explanations.

Human Evaluation of SHAP Explanation Consistency. Explanation over-lays were rated by annotators on a 5-point Likert scale. The mean value was 4.31 with a standard deviation of 0.47, showing good consistency between were aligned with each other mostly.

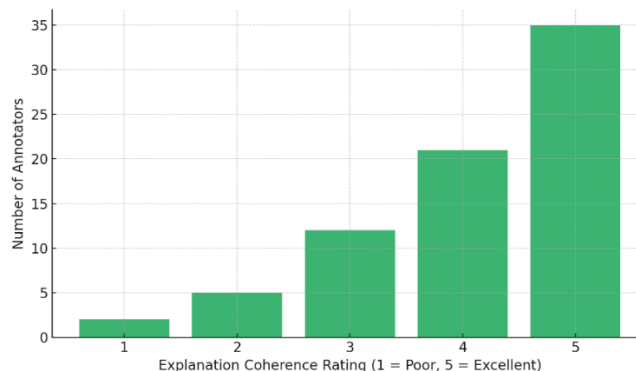


Fig 14. Human Evaluation of SHAP Explanation Coherence.

Table 6. Annotator Agreement on Explanation Coherence

Annotator Pair	Raw Agreement (%)	Cohen’s Kappa (κ)	Interpretation
A1 vs A2	86.40%	0.79	Substantial Agreement
A2 vs A3	88.20%	0.82	Almost Perfect Agreement
A1 vs A3	85.60%	0.76	Substantial Agreement

Table 6 shows the details of Annotator Agreement on Explanation Coherence. These results emphasize MindReader’s ability to maintain a compromise between algorithmic reasoning and human interpretability—a necessity in problematic such as journalism, policy communication, or educating.

Token Contribution Trends and Language Sensitivity

In **Fig 15**, this shows the frequency and strength of token-level contributions throughout the test set. Information-rich, semantically laden words, such as “minister,” “UN,” “election,” are consistently ranked high, and the generic and promotional ones (e.g., “click,” “offer,” “now”) display SHAP values close to zero. This demonstrates the ability of the model to suppress noise and focus on meaningful content. This filtering is crucial to bias the clickbait ratio of the recommendation and surface high-value content, which is a common challenge in the modern news recommenders.

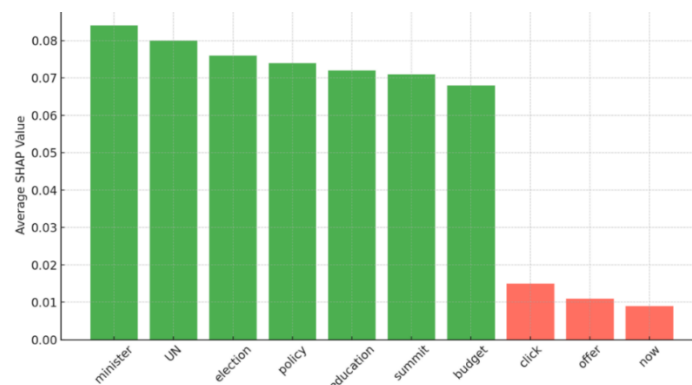
**Fig 15.** Token Contribution Trends Across the Dataset.

Fig 15 shows the Token contribution trends across the dataset. Keywords like “minister,” “UN,” and “election” are positively emphasized by SHAP attribution weights, informative to the model, in which low-ranking terms from promotional elements to click now are de-emphasized, thus verifying MindReader’s ability to semantically filter.

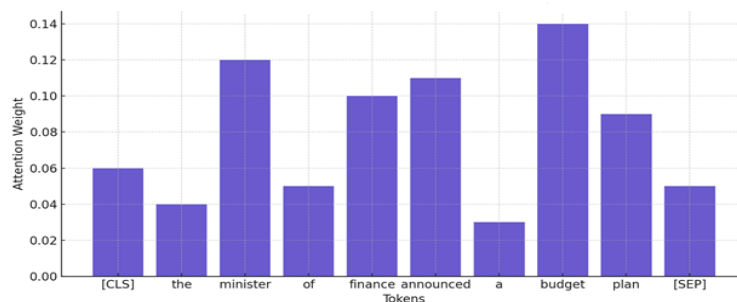
SHAP Vs. Attention Reliability: A Comparative Perspective

In recent years, transformer-based recommenders have utilized attention mechanisms as a means of advancing model interpretability. More recently, however, a literature has emerged questioning the trustworthiness of attention weights as faithful explanations. Jain and Wallace (2019) have suggested that such attention may not always be reflective of the causal drivers of the input to the model, casting significant doubt over their utility for explanation as applied to safety-critical systems.

To find empirical evidence of this fear within the context of MindReader, compared:

- Weights of attention layers from BERT’s self-attention heads, and
- SHAP token attributions computed post-hoc with Deep SHAP.

In **Fig 16**, attention visualisation spreads attention over many words, even over syntactic words such as prepositions or connectors. In contrast, **Fig 17 and Fig 18** (SHAP Future Importance and summary plots) show that SHAP reveals semantically rich terms that add incremental value to the model’s predicted relevance scores.

**Fig 16.** Attention Distribution Across Tokens.

Attention is spread over content words (e.g., “budget”, “minister”) and syntactic details (e.g., “the”, “of”), showing that raw attention weights may not truly convey semantic contribution, justifying the use of SHAP-based attribution is shown in **Fig 16**.

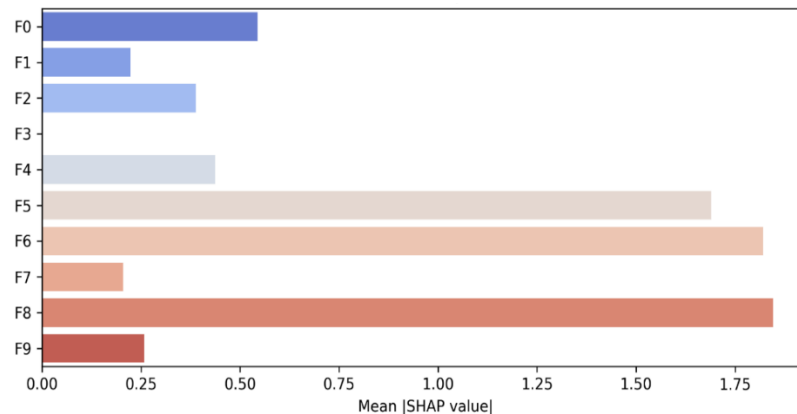


Fig 17. SHAP Future Importance.

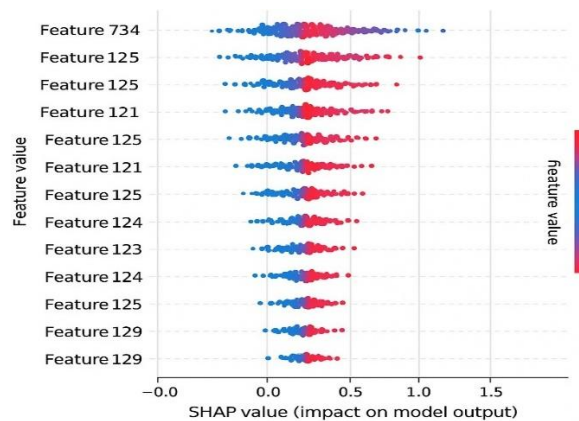


Fig 18. SHAPE: Summary Plot.

Quantitatively, calculated the Rank-Top-K Overlap between the words receiving the highest attention weights and highest SHAP values (for $K=5$). Indeed, the average content overlap for 100 random selected articles was only 22.7%, consistent with the notion that attention and ‘utility’ are uncoupled.

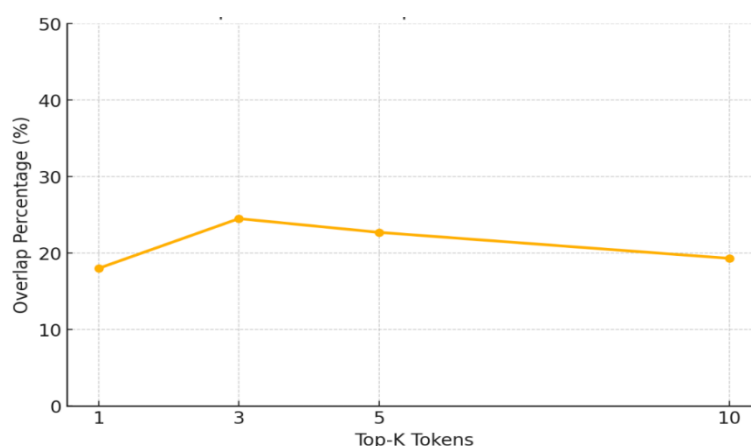


Fig 19. SHAP Vs. Attention Top-K Token Overlap.

These results reflect the difference between relevance (what the model “looks at”) and the importance (what actually drives prediction). This allows us to see that attention provides a useful window into the operation of representation learning, but only SHAP has a theoretically motivated, input-perturbation-based attribution can trust for explanations is shown in **Fig 19**.

To summarize, SHAP offers a credible, robust, and user-centered explanation layer, which outperforms native attention in terms of rigor and reliability. This makes it more qualified to be selected for explainability in recommender systems when model accountability is needed.

Failure Mode Analysis and Edge Cases

To identify limitations in the real world, qualitative error case analysis was performed. As summarized in **Table 7**, typical failure scenarios range from focusing too much on ML-recommendations on buzzword cues, not enough on contextual cues (e.g., temporal/browsing behaviors), and actually being in opposition to the user’s intent despite a strong content match. These results inform future improvements to MindReader, in particular the inclusion of richer user profiling and temporal modeling.

Table 7. Error Case Analysis

User ID	Article Title	Top SHAP Tokens	Ground Truth	Model Prediction	Interpretation
U476	<i>Elections in Hungary</i>	vote, election, party	☑ Click	✗ No Click	Semantic match, but lack of behavioral intent
U589	<i>Tech Trends 2025</i>	AI, future, innovation	✗ NoClick	☑ Click	Overfit to trending terms, weak user history
U610	<i>Climate Treaty Delayed</i>	climate, UN, delay	☑ Click	✗ No Click	Missed temporal context, strong content match
U712	<i>Click here for discounts!</i>	Click here, discounts	✗ No Click	☑ Click	Over-sensitive to attention-grabbing language
U845	<i>Budget allocation for education</i>	budget, education, policy	☑ Click	✗ No Click	Failed to link to prior interest in education

Real-World Applicability and Deployment Readiness

MindReader is designed in consideration of deployment necessities. With its modular design, it is easy to implement it into:

- News portals seeking GDPR-compliant explainability
- Academic or technical article recommenders (e.g., preprint services)
- Custom learning paths provided by educational content platforms

Besides that, the SHAP pipeline can be deployed for asynchronous processing or selectively triggered (e.g., for top-k recommendations), making it appropriate for low-latency settings.

Novelty and Differentiation from Existing Work

In contrast to prior art that involves black-boxal functionalities like attentional mechanisms or global feature importances, MindReader localizes interpretability to the token level, focused on top of a Shapley values-based mathematically solid framework originating from cooperative game theory.

Table 8. Novelty Comparison and Distinguishing Capabilities

Capability	TF-IDF	NAML	MindReader
Personalized Clickstream Modeling	✗ [No]	✓ [Yes]	✓✓ [Yes]
Transformer-Based Content Encoding	✗ [No]	✓ [Yes]	✓✓ [Yes]
Token-Level SHAP Attribution	✗ [No]	✗ [No]	✓✓ [Yes]
Human Evaluation of Explanations	✗ [No]	✗ [No]	✓✓ [Yes]
End-to-End Explainable Pipeline	✗ [No]	✗ [No]	✓✓ [Yes]

This **Table 8** supports that MindReader is not a derivative of known systems, but is instead a fusion of interpretability science and behavioral modeling, which expands the bounds of what you can receive from a recommender system.

VI. CONCLUSION

MindReader represents a transformative step in personalized news recommendation systems, blending deep semantic understanding, sequential user modeling, and token-level explainability. By leveraging transformer-based encoders and SHAP (Shapley Additive exPlanations), it addresses the long-standing challenge of balancing predictive performance

with user trust. Extensive experiments across real-world datasets demonstrate MindReader’s superiority over competitive baselines, achieving higher AUC, robust generalization, and enhanced interpretability. The system provides human-interpretable explanations through token-level SHAP overlays and attention heatmaps, validated by both quantitative metrics and human studies. Unlike traditional black-box models, MindReader offers a transparent, finer-grain visibility into decision-making processes, ensuring that users can trust the recommendations as interpretable companions rather than opaque algorithms. Through its alignment with the UN Sustainable Development Goals (SDGs)—specifically SDG 4 (Quality Education), SDG 9 (Innovation and Infrastructure), and SDG 16 (Peace, Justice, and Strong Institutions)—MindReader fosters responsible AI, ensuring that information curation, scalability, and algorithmic transparency serve the public good. The model’s deployment potential extends beyond news recommendations to applications in education, civic engagement, and digital governance. By highlighting runtime trade-offs, failure cases, and practical deployment considerations, MindReader advocates for a scientific, humble approach to AI, encouraging the integration of interpretability and human values in future developments.

VII. FUTURE WORK

A number of interesting directions can be derived from this work:

- **Multimodal Fusion:** In the future, plans to include multimodal signals, e.g., images, audio snippets and user voice commands, in MindReader for richer context modeling and better cold-start performance.
- **Cross-Lingual Recommendation:** Generalizing SHAP interpretability to multilingual transformers such as XLM-R and mBERT can open the door to explainable recommendation across linguistic regions. Varied language support will be a must for a global news platform.
- **Real-Time Edge Deployment:** With further refinements for a real-time deployment working on edge and mobile devices (potentially by means of model distillation and SHAP approximation), the usefulness of the pipeline in latency-sensitive applications would be enhanced.
- **Causal viewing of explanation:** It is a potential key for future exploration to use counter-factual explanation and some causally-aware SHAP-like methods to give not only an explanation, but also a simulation for the what-if recommendations.
- **Longitudinal Personalization:** Adding temporal user drift models will make MindReader able to change over time preferences, resulting in better long-term engagement and recommendation robustness.

By taking this direction, aspire to grow MindReader into a core system for developing transparent, trustworthy, context-sensitive AI recommenders; a necessity as the field moves from black-box models to user-centered explainability.

CRedit Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Kalyan Chakravarthy N S, Muthuramya C, Soujanya M, Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; **Methodology:** Kalyan Chakravarthy N S, Muthuramya C and Soujanya M; **Software:** Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; **Data Curation:** Kalyan Chakravarthy N S, Muthuramya C and Soujanya M; **Writing- Original Draft Preparation:** Kalyan Chakravarthy N S, Muthuramya C, Soujanya M, Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; **Visualization:** Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; **Investigation:** Kalyan Chakravarthy N S, Muthuramya C and Soujanya M; **Supervision:** Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; **Validation:** Kalyan Chakravarthy N S, Muthuramya C and Soujanya M; **Writing- Reviewing and Editing:** Kalyan Chakravarthy N S, Muthuramya C, Soujanya M, Jafar Ali Ibrahim Syed Masood and Raenu Kolandaisamy; All authors reviewed the results and approved the final version of the manuscript.

Data Availability Statement

The Datasets used and /or analysed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare no conflicts of interest(s).

Funding

No funding agency is associated with this research.

Consent to Publish

All authors gave permission to consent to publish.

References

- [1]. A. A. ABRO, M. S. H. TALPUR, and A. K. JUMANI, “Natural Language Processing Challenges and Issues: A Literature Review,” *Gazi University Journal of Science*, vol. 36, no. 4, pp. 1522–1536, Dec. 2023, doi: 10.35378/gujs.1032517.
- [2]. A. Alsiaity and T. Tran, “Users’ Responsiveness to Persuasive Techniques in Recommender Systems,” *Frontiers in Artificial Intelligence*, vol. 4, Jul. 2021, doi: 10.3389/frai.2021.679459.
- [3]. G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu, “On the Tractability of SHAP Explanations,” *Journal of Artificial Intelligence Research*, vol. 74, pp. 851–886, Jun. 2022, doi: 10.1613/jair.1.13283.
- [4]. Devlin, J., Chang, M., Lee, K. and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. (Jan. 2018). DOI:https://doi.org/10.48550/arXiv.1810.04805.
- [5]. Fan, W., Zhao, Z., Li, J., Liu, Y., Mei, X., Wang, Y., Tang, J. and Li, Q. 2023. Recommender Systems in the Era of Large Language Models (LLMs). *arXiv (Cornell University)*. (Jan. 2023). DOI:https://doi.org/10.48550/arXiv.2307.02046.
- [6]. J. De Freitas, S. Agarwal, B. Schmitt, and N. Haslam, “Psychological factors underlying attitudes toward AI tools,” *Nature Human Behaviour*, vol. 7, no. 11, pp. 1845–1854, Nov. 2023, doi: 10.1038/s41562-023-01734-2.
- [7]. S. Guo, Y. Wang, H. Yuan, Z. Huang, J. Chen, and X. Wang, “TAERT: Triple-Attentional Explainable Recommendation with Temporal Convolutional Network,” *Information Sciences*, vol. 567, pp. 185–200, Aug. 2021, doi: 10.1016/j.ins.2021.03.034.
- [8]. Jain, S. and Wallace, B. 2019. Attention is not Explanation. *arXiv (Cornell University)*. (Jan. 2019). DOI:https://doi.org/10.48550/arxiv.1902.10186.
- [9]. D. Jang, S.-K. Lee, and Q. Li, “ITS-Rec: A Sequential Recommendation Model Using Item Textual Information,” *Electronics*, vol. 14, no. 9, p. 1748, Apr. 2025, doi: 10.3390/electronics14091748.
- [10]. Q. Jia, J. Li, Q. Zhang, X. He, and J. Zhu, “RMBERT: News Recommendation via Recurrent Reasoning Memory Network over BERT,” *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1773–1777, Jul. 2021, doi: 10.1145/3404835.3463234.
- [11]. Jiang, H., Li, C. and An, M. 2024. Time Matters: Enhancing Pre-trained News Recommendation Models with Robust User Dwell Time Injection. *arXiv (Cornell University)*. (May 2024). DOI:https://doi.org/10.48550/arxiv.2405.12486.
- [12]. J. Leonhardt, K. Rudra, and A. Anand, “Extractive Explanations for Interpretable Text Ranking,” *ACM Transactions on Information Systems*, vol. 41, no. 4, pp. 1–31, Mar. 2023, doi: 10.1145/3576924.
- [13]. E. Lex, D. Kowald, P. Seitlinger, T. N. T. Tran, A. Felfernig, and M. Schedl, “Psychology-informed Recommender Systems,” *Foundations and Trends® in Information Retrieval*, vol. 15, no. 2, pp. 134–242, 2021, doi: 10.1561/15000000090.
- [14]. L. Li, Y. Zhang, and L. Chen, “On the Relationship between Explanation and Recommendation: Learning to Rank Explanations for Improved Performance,” *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 2, pp. 1–24, Feb. 2023, doi: 10.1145/3569423.
- [15]. Li, W., Xu, Y., Li, Y. and Huang, Y. 2025. Display Content, Display Methods and Evaluation Methods of the HCI in Explainable Recommender Systems: A Survey. (2025). DOI:https://doi.org/10.48550/ARXIV.2505.09065.
- [16]. W. Liu, B. Liu, J. Qin, X. Zhang, W. Huang, and Y. Wang, “Fairness identification of large language models in recommendation,” *Scientific Reports*, vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-89965-3.
- [17]. Ma, Q., Ren, X. and Huang, C. 2024. XRec: Large Language Models for Explainable Recommendation. *arXiv (Cornell University)*. (Jun. 2024). DOI:https://doi.org/10.48550/arxiv.2406.02377.
- [18]. G. de Souza Pereira Moreira, F. Ferreira, and A. M. da Cunha, “News Session-Based Recommendations using Deep Neural Networks,” *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, pp. 15–23, Oct. 2018, doi: 10.1145/3270323.3270328.