Beyond Precision: Understanding the Impact of Algorithmic Accuracy and Transparency on User Perceptions in Keyword-Driven Contextual Advertising

Jingwen Cai
Department of Computing Science
Umeå University
Umeå, Sweden
jingwenc@cs.umu.se

Bart Piet Knijnenburg
School of Computing
Clemson University
Clemson, South Carolina, USA
bartk@clemson.edu

Johanna Björklund
Department of Computing Science
Umeå University
Umeå, Sweden
johanna@cs.umu.se

Sara Leckner
Department of Computer Science and Media Technology
Malmö University
Malmö, Sweden
sara.leckner@mau.se

Abstract

Algorithms frequently manage online advertising markets, aligning advertisements with article topics. Our work investigates how users perceive the relevance of ads to articles when ads are placed using different keyword extraction algorithms, including Large Language Models (LLMs), and how transparency about the placement procedure influences these perceptions and behavioral intentions. We conducted an online user experiment (N = 498) where ads are matched with news articles using the keyword extraction methods TF-IDF, KeyBERT, and DeepSeek. Results indicate that lightweight methods can match advanced LLMs in delivering high user-perceived ad-article relevance, which in turn fosters click and purchase intentions. However, providing explanations for the ad-article placements by displaying extracted keywords reduces ad interest and thereby weakens behavioral intentions, while simultaneously increasing perceived relevance and moderating algorithm effects. These findings highlight the complex impact of transparency-increasing explanations and suggest that algorithmic precision metrics must be complemented by user perception and intention measures.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Information systems** → **Evaluation of retrieval results**; **Online advertising**.

Keywords

Keyword Extraction, Human Factors, User Perception, Contextual Advertising, Transparency

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791240

ACM Reference Format:
Jingwen Cai, Bart Piet Knijnenburg, Johanna Björklund, and Sara Leckner. 2026. Beyond Precision: Understanding the Impact of Algorithmic Accuracy and Transparency on User Perceptions in Keyword-Driven Contextual Advertising. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3791240

1 Introduction

Online advertising is now a dominant economic and cultural force that generates vast revenues [22] and influences how people perceive organizations and brands. Behind the scenes, these ads are allocated through real-time auctions, where advertisers bid for available placements. However, increasing user concerns about privacy and ethical issues, such as tracking, manipulation and interruption, have contributed to negative attitudes toward online ads (e.g. [26, 74, 80]). These concerns have also prompted a broader global trend in advertising regulation and platform governance, with transparency requirements increasingly implemented worldwide, including the General Data Protection Regulation [1] and the California Consumer Privacy Act [2]. In turn, these shifts have fueled interest in *contextual advertising*, a privacy-aware approach that aligns ads with surrounding media content rather than individual user profiles. Given the scale and speed at which ad opportunities are traded [60], this alignment is typically achieved through algorithmic keyword extraction, whereby terms are identified within content and matched to pre-defined keyword sets to guide ad placement. Although this approach mitigates privacy risks, the integration of increasingly advanced techniques, such as Large Language Models (LLMs), raises new questions about user experiences. In particular, it remains unclear whether more sophisticated methods, while enabling finer-grained targeting, are consistent with user perceptions or intensify user privacy and ethical concerns. Understanding how users interpret and evaluate these evolving forms of contextual advertising is therefore essential.

Although modern contextual advertising systems may exploit richer semantic signals, keyword-based targeting remains central

and effective, particularly in privacy-sensitive and cold-start settings where profiling is infeasible. Its simplicity and interpretability make it a meaningful abstraction of contextual advertising for exploring user perceptions. Previous evaluations of keyword extraction have largely focused on benchmarking algorithmically extracted keywords against gold-standard datasets, emphasizing measures such as precision and accuracy (see, e.g., [28, 68]). However, it remains unclear whether improvements in algorithmic performance translate into meaningful gains in user experience or advertising effectiveness. For example, little is known about whether closer alignment with gold-standard keywords actually leads users to perceive ads as more relevant or to report stronger purchase intentions. This uncertainty reflects a broader misalignment between how keyword extraction systems are evaluated and what matters in real-world deployment: benchmark improvements at the keyword level may not correspond to the true objectives of contextual advertising. If this disconnect persists, increasingly complex models may improve benchmark scores but deliver little additional value to users. In such cases, they may prioritize sophistication over user benefits and produce ads that appear technically capable but fail to support positive engagement, trust needs, or informed decision-making for users.

The Digital Services Act [3] requires platforms to disclose "meaningful information about the main parameters" behind each targeted ad. Similar transparency mandates across jurisdictions reflect a broader regulatory shift toward explainability and accountability in advertising. For keyword-based targeting, this could involve revealing the specific keywords that triggered an ad's placement. Yet it remains unclear how such transparency affects users' perceptions or broader attitudes toward advertising. Without such understanding, transparency may resolve one set of concerns while inadvertently creating new conflicts between users and advertisers.

In this paper, we examine how different keyword-extraction algorithms and transparency about the matching process influence users' perceptions and behavioral intentions regarding ad-article recommendations. Reflecting these aims, our work is guided by the following research questions:

- **RQ1** How do different keyword extraction methods affect users' perceptions (perceived relevance, perceived congruence) and behavioral intentions (click intention, purchase intention) regarding keyword-based ad-article recommendations?
- **RQ2** What is the impact of providing explanations for recommendations, by displaying extracted keywords, on users' perceptions and intentions?
- **RQ3** How does providing explanations moderate the effects of keyword extraction methods on users' perceptions and intentions?
- **RQ4** To what extent do users' perceptions of relevance and congruence mediate the effects of keyword extraction methods on click and purchase intentions?

To answer these questions, we conducted an online user experiment, comparing the impact of ads matched with news articles using keywords extracted by three different algorithms (TF-IDF, KeyBERT and DeepSeek), against ads matched using a gold standard baseline, in terms of users' perceptions of relevance and congruence, and their click and purchase intentions.

The paper makes several contributions. First, while industry systems continually optimize complicated contextual matching strategies, empirical work focusing on user-centered perceptual mechanisms in such automated contexts remains limited. Our study provides evidence of a misalignment between algorithmic benchmarking performance and user perceptions. It highlights the ethical tension between users' right to transparency [45, 75] and advertisers' goal of maximizing purchase intentions, and reveals how additional complexity emerges at their intersection. Choosing algorithms with the highest quantitative performance does not guarantee positive user outcomes and may instead result in wasted time, money, and effort. Second, we show that explanations of ad placements play a complex role. They increase perceived relevance but simultaneously reduce ad liking, which in turn lowers behavioral intentions. Moreover, simpler algorithms may produce more negative user attitudes toward ads when their outputs are made explicit through explanations. Third, we offer practical implications for system design, showing that algorithm selection and explanation design require integrated, user-centered optimization.

The rest of the paper is organized as follows. We begin with the theoretical background, integrating perspectives on contextual advertising, keyword extraction, user perception and intention, and transparency in algorithmic ad placements. We then introduce our study design, present the findings, and conclude with a discussion of their implications for advertising practice, transparency policy, and algorithm design.

2 Theoretical background

2.1 Contextual Advertising and Keyword Extraction Methods

Contextual advertising refers to the strategic placement of ads within media environments deemed "relevant". In practice, relevance is commonly operationalized using keyword-based representations of both advertisements and media content. Advertisers compete for ad placements through automated auctions, which are typically resolved within approximately 100 milliseconds [60]. Even smaller platforms may conduct hundreds of thousands of such auctions per second, placing strict requirements on the efficiency and scalability of the underlying algorithms. For this reason, keywords are extracted from each article the first time it enters the advertising system. This process involves automatically identifying the most representative words or phrases in the document. These keywords summarize the article's content and serve as compact proxies for the full text, enabling efficient large-scale matching without repeated comparisons to complete documents.

Methods for keyword extraction range from simple statistical techniques to advanced LLM-based approaches. This study focuses on three representative methods:

- **TF-IDF** is one of the oldest yet still widely used keyword extraction methods [5, 70]. It is simple and computationally efficient, with applications in various domains including advertising [79]. However, it does not capture semantic meaning, which can lead to ambiguities for polysemous terms.

- **KeyBERT** [34] builds on the pre-trained BERT language model [23]. By leveraging embeddings, it captures semantic relationships and selects keywords accordingly. That said, its performance can vary depending on the characteristics of the domain and the degree of fine-tuning applied [64, 69].
- **DeepSeek V3** [21] is an open-sourced LLM with 671 billion parameters, released in 2024. Its performance is comparable to leading models such as ChatGPT-4, while being more cost-efficient to train and deploy. Despite these strengths, it has raised concerns similar to other LLMs, such as algorithmic discrimination [8, 77].

Previous research has shown that simple and lightweight extraction methods can sometimes outperform more advanced ones in user-centered tasks [15, 35]. Nevertheless, only a limited number of studies have examined this issue in depth, leaving a gap between method development and real-world application. Specifically, [15] compares keyword extraction methods using user perceptions but focuses solely on keyword quality in news articles, and does not consider ads or behavioral outcomes such as clicking intentions.

2.2 User Perception

Congruence is a central concept in contextual advertising, often framed as relatedness, similarity, or shared annotations [49, 58]. Similarly, perceived relevance captures how individuals judge whether a source connects to them or supports their needs, goals, or values [17]. While the concepts are sometimes used interchangeably, relevance can be seen as a sub-dimension of congruence [58].

Perceived relevance is a central pathway through which ad-context congruence shapes consumer attention, attitudes, and behavioral intentions [17, 42, 72, 76]. In contextual advertising, relevance is multi-layered. At a basic level, it reflects topical fit between an ad and adjacent content. More indirectly, relevance can emerge through priming, where media contexts activate associations that make the advertised product feel timely, useful, or meaningful. Priming [66, 78] describes how exposure to one stimulus influences the interpretation of subsequent information, often outside conscious awareness. Such associations can facilitate ad processing and recall, whereas weak or mismatched pairings may create cognitive conflict and sometimes generate negative evaluations. The Relevance-Accessibility Model [10] extends these dynamics by arguing that primed associations affect responses only when they are perceived as personally relevant to one's goals or needs, when there is interest in the ad. Advertising effects should therefore be strongest when contextual cues both heighten accessibility and align with consumers' perceived relevance.

Prior research generally links congruence and perceived relevance to positive consumer responses [19, 49, 67]. However, evidence is mixed. High congruence does not automatically improve effectiveness, and its benefits depend on conditions such as whether the ad captures undivided attention [38, 67]. In some contexts, incongruent ads may even outperform congruent ones.

2.3 Behavioral Intention

Click intention and purchase intention are both behavioral intentions that, in advertising, represent relatively low-cost actions but different depths of engagement. Click intention is driven by perceived contextual relevance and cues such as visual design, credibility signals, and emotional tone [25, 56]. Purchase intention reflects a stronger commitment. It is typically explained by attitudes toward the product and ad, subjective norms, and perceived behavioral control, which together form an intention to buy and can translate into actual purchases [6].

Dual Process Theory (DPT) and the Elaboration Likelihood Model (ELM) offer complementary accounts of how these intentions arise. DPT distinguishes between a fast, heuristic mode of thought (System 1) and a slower, analytic mode (System 2) [29, 43, 71]. ELM's peripheral and central routes can be understood as domain-specific instantiations of these systems [62]. Applied to contextual advertising, these suggest that clicks are more likely to emerge through System 1/peripheral processing when an ad feels immediately congruent with surrounding content. Purchase intention, by contrast, is more likely to be reinforced through System-2/central processing, when users have active goals and engage with substantive evaluative information. Users may therefore click to explore or compare options, and develop purchase intention if the message withstands more effortful scrutiny.

Consistent with this view, advertising research often treats clicks as an early step in an effects chain leading to purchase. Click intention may function as a precursor that increases the likelihood of later purchase intention, or as a mediator between initial perceptions and purchase intention [18, 81]. In contextual settings, perceived relevance or congruence tends to strengthen both click and purchase intentions, whereas perceived intrusiveness typically undermines them. Prior HCI research similarly shows that attitudes, decisions, and consumption behaviors reflect the interplay of these two processing modes, although debate remains about which mode dominates under different conditions [47, 48, 50, 54, 82].

2.4 Transparency in Contextual Advertising

Contextual advertising is typically generated by platform algorithms [27, 44]. Because these systems are often opaque, users lack insight into the logic behind contextual placements. This opacity can foster perceptions of manipulation, erode trust, and contribute to algorithm fatigue, ultimately increasing ad avoidance and reducing engagement [27, 77]. In response, scholars and practitioners have called for greater transparency and clearer explanations of ad delivery processes [11, 14]. Platforms have begun experimenting with such measures. For example, Meta's *Why Am I Seeing This Ad?* feature provides brief explanations of how prior interactions guide the extraction methods and ad delivery [14, 24].

However, transparency does not automatically resolve these concerns. On the one hand, consumers generally respond more positively when data collection is disclosed and explanations are clear [4, 41]. Such explanations can enhance trust by clarifying acceptable targeting logic, especially on trusted platforms [44]. On the other hand, disclosures may heighten awareness of persuasive intent, triggering more critical processing which can result in ad avoidance [13]. The impact of transparency also depends on explanation design. They can backfire if too detailed, too personal, or too vague [11, 26, 44]. Prior studies show that moderately detailed explanations typically yield more favorable evaluations than highly

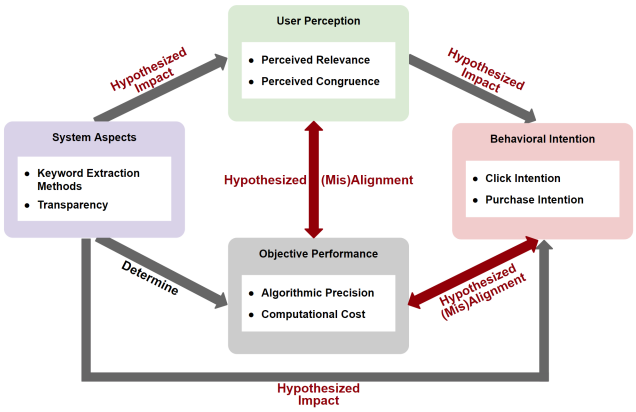


Figure 1: Conceptual model. It illustrates hypothesized impacts and potential (mis)alignments among the central components for keyword-based contextual advertising.

detailed ones [24], and that users tend to prefer interpretable, "non-creepy" explanations [26]. Thus, although transparency is widely advocated, its actual effects remain uncertain.

The Persuasion Knowledge Model (PKM) offers further theoretical insight into why transparency in advertising can both help and hurt. According to PKM, users interpret and respond to advertising attempts based on their understanding of persuasion tactics [30]. When explanations reveal targeting logic, users may infer advertisers' persuasive motives and activate their "persuasion knowledge", shaping perceptions of credibility, fairness and potential manipulation. This recognition can reduce ad effectiveness by triggering skepticism or resistance, especially when the persuasive intent is perceived as covert or misaligned with user goals [16]. In algorithmically delivered contextual advertising, explanations can become a key site of user sensemaking. Even simple disclosures can influence whether the system is viewed as relevant, trustworthy, or intrusive, making transparency ethically consequential.

2.5 The Conceptual Model

Based on the research questions and theoretical foundations, Figure 1 presents the conceptual model guiding our experiment. The model hypothesizes that different keyword extraction methods shape user-perceived relevance and congruence, which in turn mediates their click and purchase intentions. It further conceptualizes potential alignments or misalignments between algorithmic performance and user-centered evaluations. In addition, the model enables us to explore the potential moderating role of providing explanations for keyword-based ad placements, drawing on theory that highlights the complex nature of transparency in both activating users' persuasion knowledge and supporting their expectations for privacy. These hypothesized pathways structure our experimental design, guide the deployment of the stimulus conditions, and inform our analytic approach.

3 Method

We conducted a survey-based experiment to investigate participants' perceptions of keyword-based ad recommendations under different keyword extraction methods. This approach is well-suited to our research questions because perceptions of article–ad combinations are inherently subjective and therefore require self-reports rather than behavioral data. Moreover, real-world interaction with ads is relatively rare, making an observational experiment difficult to conduct. In the following section, we describe the participants, study design, stimuli, and the experimental procedure involved.

3.1 Participants

A total of 505 participants were recruited through the *Prolific* crowdsourcing platform (see Section 3.4 for additional details). This sample size was chosen to balance survey cost with statistical rigor. A power analysis for our exact experimental design is difficult to conduct, but an analogous, less powerful design produced an upper-bound on the minimum sample size of 430 participants to detect significant medium-sized within-subjects effects and within-between interactions with a statistical power of 0.80. Participation was restricted to US participants to reflect the origin of the news articles used in the experiment. To ensure anonymity, other than basic demographic information (see Table 1), no identifying information was collected. The study took an average of 17 minutes to complete, and participants were compensated at a rate of at least $8.00 per hour, in accordance with the compensation standards of Prolific. Participants were randomly assigned to experimental conditions, ensuring that any residual variation in ad attitudes or preferences was evenly distributed across groups.

3.2 Study Design

The participants were asked to evaluate a sequence of webpages, each containing a news article paired with an advertisement and a set of questions assessing ad-article similarity, perceived relevance, and purchase intentions (see Section 3.4.1). The experiment was manipulated along three major dimensions:

- **Advertisements (manipulated within-subjects)** — We designed five ads (see Figure A.1 in Appendix A) spanning different topics: vacation services, health supplements, automotive products, contraception, and laxatives. The final two categories were intentionally included to examine perceptions of sensitive ads in contrast to non-sensitive ones.
- **Keyword extraction methods (randomly assigned per ad)** — Four methods were used to extract keywords from the candidate articles, resulting in a *KPTimes Gold Standard* baseline, the frequency-based algorithm *TF-IDF*, the embedding-based algorithm *KeyBERT*, and the large language model *DeepSeek*. Out of these, TF-IDF is arguably the least complex and DeepSeek is the most complex. The extracted keywords were then used to identify the most suitable news articles to pair each ad with. We also included *Random* as a fifth method, where the articles and ads were matched entirely at random, without generating or referencing any keywords.
- **Explanation (manipulated between-subjects)** — For each ad–article recommendation, we varied the presence of explanations by displaying or hiding the keywords used to generate the recommendation. These explanations revealed only the extracted keywords, without disclosing the underlying

Table 1: Distribution of the age, gender, and education level reported by participants.

Age	Proportion	Gender	Proportion	Education Level	Proportion
< 18	0.2%	**Female**	49.0%	**Less than a high school diploma**	0.6%
18 - 29	19.0%	**Male**	49.2%	**High school diploma or equivalent**	21.4%
30 - 39	26.4%	**Non-binary**	1.2%	**College degree**	38.3%
40 - 49	22.6%	**Prefer to self-describe**	0.2%	**Graduate degree**	36.1%
50 - 59	17.1%	**Prefer not to disclose**	0.4%	**Other**	3.0%
60 - 69	9.1%			**Prefer not to disclose**	0.6%
> 69	5.4%				
Prefer not to disclose	0.2%				

Note: Demographic information was not available for one participant, and proportions are based on the remaining sample ($505 - 1$).

method, and were provided for all recommendations except those produced by the Random article selection method.

3.3 Stimuli

Ad–article recommendations were generated through a three-step process: first, creating keywords for each ad; second, extracting keywords for each candidate article using the specified keyword-extraction methods; and third, identifying the most relevant articles for each ad based on keyword similarity.

3.3.1 Ad keywords. To identify representative keywords for each stimulus, three independent evaluators reviewed printed versions of the five created advertisements used in the study. Each evaluator was asked to select ten keywords that, in their judgment, best captured the theme in terms of content and message of the respective ad. The resulting keywords were then compiled and duplicates deleted. The remaining keywords were used as measures of stimulus representation, with each ad having more than ten keywords and the number of keywords varied across ads.

3.3.2 News articles. We used the *KPTimes* dataset [31], which contains nearly 260,000 English-language news articles on diverse topics such as world news, published between 2006 and 2017 by the *New York Times* and *Japan Times*. To reduce participant fatigue, we excluded articles longer than 300 words. From the remaining corpus, we pre-selected 1,576 relevant articles based on ad-category keywords (i.e., vacation, health, automotive, contraception, and laxatives) to limit computational costs. This also avoids the problem of data sparsity in ad-article matching. We then added a random 20% sample from the remaining corpus, yielding a final dataset of 2,000 articles. The dataset provides metadata and pre-annotated keywords for each article, which served as the gold standards for evaluating the extraction methods in our study, as well as the keyword input for the KPTimes Gold Standard baseline condition. According to Gallina et al. [31], these gold standard keywords were first generated by algorithms and then revised by human editors.

3.3.3 Article keywords. Ten keywords were extracted from each news article in the final dataset using TF-IDF, KeyBERT and DeepSeek, respectively. To avoid introducing bias, text pre-processing was kept to a minimum and limited to the removal of stopwords, based on the *ROUGE* list [32]. In the case of KeyBERT, we used the pre-trained model `multi-qa-mpnet-base-dot-v1`[1] and extracted keywords under a 1-3 gram setting. This setting enables the extraction of not only unigram and bigram keywords but also trigram keyphrases, thereby providing the algorithm with a broader selection and allowing for more precise choices. The 1-3 gram setting was also used for TF-IDF. In the case of DeepSeek, we used the DeepSeek-V3[2] API to extract keywords for the articles, with a prompt scheme adapted from the *KeyLLM* project[3]. This incurred a cost of $1.23 for processing about 565,000 tokens. For a comparison of computing times across these algorithms, see Figure C.1 in Appendix C.

3.3.4 Ad-article matching. For each combination of ad and keyword-generation method, we iterated over the articles. For each article-method pair, we used *Word2Vec* [59] to compute cosine similarities between the ad keywords and the article keywords generated by the given method. The five articles with the highest similarity scores were then selected to form the ad–article recommendations for that particular ad and method. Under the Random method, five articles were selected entirely at random for each ad, without keyword extraction or similarity matching. Altogether, this process produced an initial pool of $5 \times 5 \times 5$ (five ads, five methods, and five articles) ad–article recommendations for the experiment. There are 33 duplicated articles, meaning that more than one method selected the same article for a given ad based on keyword similarities. These duplicates were retained in the pool because each method-ad-article combination represents a distinct matching instance.

3.4 Procedure

As previously described, we used an online questionnaire to survey participants' perceptions of the ad–article recommendations, aiming to examine the effects of the experimental manipulations and relevant covariates.

The questionnaires were designed using the *Potato* [61] annotation tool and consisted of three sections: a demographic questions page, the main ad-article recommendation rating pages (see Figure

[1]The model has good performance across various NLP tasks, see https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1
[2]Also known as DeepSeek Chat
[3]See https://maartengr.github.io/KeyBERT/guides/keyllm.html

B.1 in Appendix B for an example), and a feedback page. In the ad-article recommendation rating section, each participant was randomly assigned five ad-article recommendations from the initial pool, covering all ads. Each participant rated every ad once, but the method that produced the paired article for each ad was randomly selected from TF-IDF, KeyBERT, DeepSeek, Gold Standard, or Random. On each rating page, the ad and article were displayed side by side, followed by the survey questions covering the aspects described in Section 3.4.1. For simplicity, we refer to each ad-article recommendation together with the survey questions as an *experimental instance*.

We manipulated the presence of keyword-based explanations in a between-subjects manner, recruiting 278 participants who evaluated ad-article recommendations without seeing the keywords and 227 participants who evaluated ad-article pairs where keywords were presented below the ad and article. Participants were only allowed to take part in the experiment once, and the reduced participant sample size for the with-keywords condition was due to the fact that the Random method is not keywords-based. Hence, no keywords could be displayed for this method, meaning that 25 ad-article recommendations (five articles per ad) had to be excluded from this condition. Finally, seven participants were excluded because their responses showed invalid patterns. They provided identical ratings for more than 90% of the questions for more than one instance, suggesting low-effort or non-attentive responding. Valid ratings from 498 participants were included in the final analysis.

3.4.1 Dependent variables and covariates. The following subjective constructs were considered, each measured through one or more questionnaire items, rated on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree". Guided by theoretical foundations (e.g., [9, 27, 38, 49, 57]), the items were developed and subsequently refined through expert review. The complete questionnaire items for these constructs are provided in Appendix D.

- **Interest in the ad** was measured with the item: "I find the ad interesting".
- **Interest in the article** was measured with the item: "I find the article interesting".
- **Perceived relevance** was measured with three items evaluating how closely participants perceived the advertised products or services to be related to the news articles.
- **Perceived congruence** was measured with four items assessing participants' perceptions of topical similarity between ads and articles.
- **Click intention** was measured with three items evaluating how likely participants were to engage with the ad after viewing the ad–article pairing.
- **Purchase intention** was measured with three items assessing how likely participants were to consider purchasing the advertised products or services.

We conducted a Confirmatory Factor Analysis (CFA) to validate the robustness of these subjective constructs. The CFA was conducted using R's lavaan package, which uses a non-normality robust estimator (to account for Likert scale responses) and handles the within-subjects study design. By modeling each group of survey questions as a latent variable, confirmatory factor analysis assesses how well the question captures its corresponding factor (i.e., the loadings)[7, 46]. The results revealed a lack of discriminant validity due to very strong correlations between *perceived relevance* and *perceived congruence* ($r = .99, p < .001$), as well as between *click intention* and *purchase intention* ($r > .99, p < .001$). Such high correlations suggest that future studies may benefit from employing a more in-depth and fine-grained way of designing the survey questions. Nonetheless, participants in this study responded with highly similar patterns across these closely related constructs. Therefore, to mitigate the potential risk of multicollinearity, we combined the perceived relevance and perceived congruence into a single latent factor *perceived relevance*, and click intention and purchase intention into another latent factor *behavioral intention*. The resulting factor model had an adequate[4] fit ($\chi^2(88) = 1406.91, p < .001$, CFI = 0.966, TLI = 0.960, RMSEA = 0.078, 90% CI: [0.074, 0.081]) and displayed both convergent validity (Average Variance Extracted (AVE) > 0.5) and discriminant validity ($\sqrt{AVE}$ > the correlation between the factors, which is 0.331). Table D.1 in Appendix D shows the final factor loadings and AVEs.

Two additional observed covariates were included in the final analysis: *ad selection* and *matched interests*. For each ad–article recommendation, participants were asked how they thought the ad was selected for the article. This single-choice question ("Selected by humans", "I don't know", or "Selected by algorithms") was designed to capture the effect of perceived recommendation source on the latent variables. Participants also self-reported their product interests via a multi-select question spanning several categories of products and services, including "Travel & Experiences", "Health & Wellness", and "Automotive". This information, collected prior to the ad–article recommendation rating task and independent of the ads shown, allowed us to assess whether participants' personal interests aligned with any of the ads they evaluated, and to use this alignment as a covariate in our statistical analysis.

4 Results

In this section, we first present a Structural Equation Model (SEM) to statistically examine the effects of different keyword extraction methods on participants' perceptions and intentions (RQ1), as well as the mediation effects between them (RQ4). The model will also be used to evaluate the influence of providing explanations on participants' perceptions and intentions (RQ2) and their moderation effects (RQ3). We then provide more targeted interpretations for each research question based on the model results. Finally, we present the algorithmic performance following the common benchmarking evaluation procedure.

4.1 The Structural Equation Model

We construct the structural equation model by combining the latent variables from our confirmatory factor analysis results with the observed covariates and the three main experimental conditions described in Section 3.2, into a structural model. The initial structure of the model was guided by our research questions, and iteratively updated by removing non-significant effects to produce

[4] Theoretically, a well-fitting model is not statistically different from the fully specified model (i.e., the p-value of the χ^2 test should be > 0.05), but this statistic is commonly regarded as too sensitive [12]. As such, Hu and Bentler proposed cut-off values for the alternative fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90% CI falling below 0.10 based on extensive simulations [37].

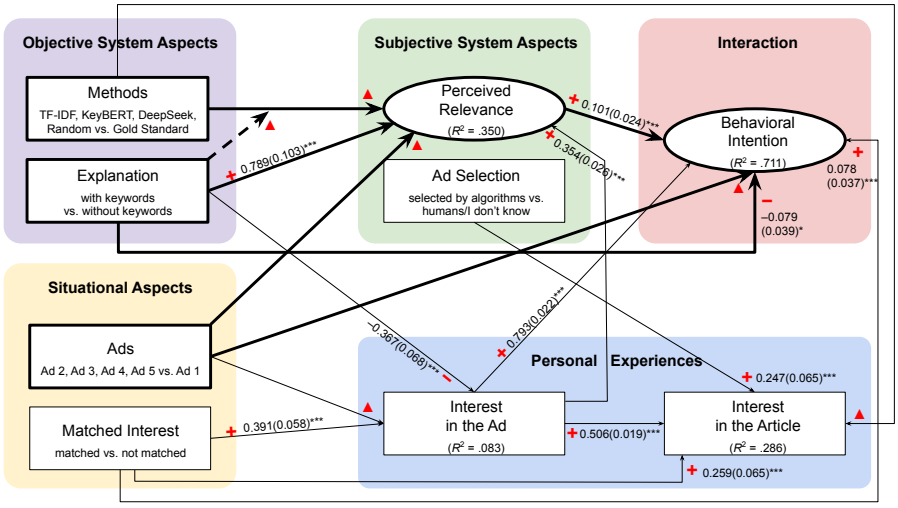


Figure 2: The structural equation model for the experimental data. Arrows represent causal effects, with "+" indicating positive effects and "-" the negative effects compared to the baselines. Numbers on the arrows represent the β coefficients (and standard errors) of the effects. "▲" denotes corresponding values given in Table 2. Significance levels: $***p < .001, **p < .01, *p < .05, n.s.\ p > .05$. R^2 is the proportion of variance explained by the model.

Table 2: Detailed estimates (β coefficients, standard errors and significance levels) corresponding to the "▲" indicators in Figure 2.

		Perceived Relevance		Behavioral Intention		Interest in the Ad		Interest in the Article	
		β (SE)	Sig.	β (SE)	Sig.	β (SE)	Sig.	β (SE)	Sig.
Methods (vs. Gold Standard)	TF-IDF	0.725 (0.130)	***					−0.087 (0.065)	n.s.
	KeyBERT	0.768 (0.115)	***					−0.196 (0.064)	**
	DeepSeek	0.357 (0.138)	**					−0.141 (0.065)	*
	Random	−0.105 (0.117)	n.s.					−0.233 (0.088)	**
Ads (vs. Ad 1)	Ad 2	0.182 (0.044)	***	−0.026 (0.040)	n.s.	−0.296 (0.044)	***		
	Ad 3	0.403 (0.051)	***	0.120 (0.037)	***	−0.019 (0.048)	n.s.		
	Ad 4	0.512 (0.052)	***	−0.131 (0.043)	***	−0.188 (0.048)	***		
	Ad 5	−0.115 (0.049)	*	0.040 (0.041)	n.s.	−0.491 (0.048)	***		

Note: Interaction effects between Explanation and Methods on perceived relevance are: $-0.648(0.132)^{***}$ for TF-IDF, $-0.751(0.132)^{***}$ for KeyBERT, and $0.078(0.156)$ *n.s.* for DeepSeek, compared to Gold Standard without explanation (See Figure 3a).

a clear and explainable model. The structural equation modeling analysis was also conducted using the lavaan package in R, which uses a non-normality robust estimator to account for Likert scale responses and handles the within-subjects study design. The final SEM model, shown in Figure 2, demonstrates good[4] fit to the experimental data: $\chi^2(335) = 1442.32, p < .001, \text{CFI} = 0.957, \text{TLI} = 0.950, \text{RMSEA} = 0.037, 90\%\ CI: [0.035, 0.038]$. Based on the structural equation modeling results, the marginal effects across keyword extraction methods and explanations on the dependent variables are reported in Figure 3. For clarity, the effects of the Gold Standard without explanation are normalized to zero, serving as a baseline.

4.2 The Effect of Keyword Extraction Methods on Participants' Perceptions and Behavioral Intentions (RQ1)

The SEM results show that keyword extraction methods significantly impact participants' perceptions of ad-article recommendations ($\chi^2(4) = 80.966, p < .001$). In the absence of keyword-based explanations, recommendations generated by all algorithmic methods used in our experiments are perceived as more relevant than those using gold-standard keywords. Specifically, Figure 3a shows that in the "no keywords" condition, KeyBERT ($b = 0.768, p < 0.001$) and TF-IDF ($b = 0.725, p < 0.001$) achieve the highest relevance ratings, followed by DeepSeek ($b = 0.357, p = 0.012$). In contrast, recommendations based on the Random method are perceived as

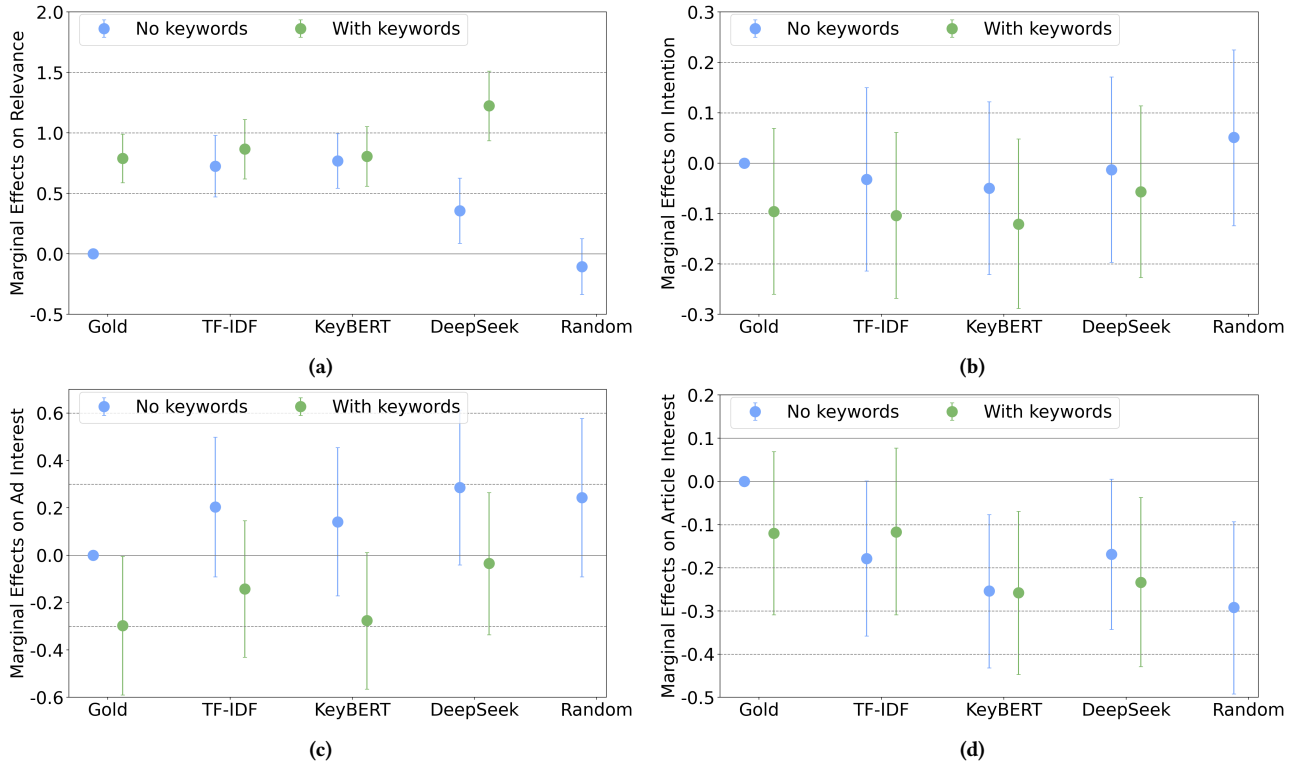


Figure 3: Marginal effects of keyword extraction methods and explanations on (a) perceived relevance, (b) behavioral intention, (c) interests in the ad and (d) interests in the article. The effect of Gold Standard without showing keywords as explanations on relevance is set to zero, and the error bars indicate 95% confidence intervals around the estimated marginal effects.

less relevant than the gold-standard baseline ($b = -0.105, p = 0.373$). These findings indicate that participants' perceptions of relevance are influenced by the choice of keyword extraction methods.

Interestingly, the model also shows a negative relationship between keyword methods and participants' interest in the recommended news articles ($\chi^2(4) = 12.764, p = .01$). While the Random method ($b = -0.233, p = 0.008$) results in the largest decrease in article interest compared to the gold-standard baseline, both KeyBERT ($b = -0.196, p = 0.002$) and DeepSeek ($b = -0.141, p = 0.029$) also result in a significantly lower article interest (TF-IDF has a non-significant negative effect: $b = -0.162, p = 0.183$). Figure 3d provides their marginal effects on the article interest. Furthermore, no significant direct causal relationship is found between participants' perception and their interest in the recommended news articles. This might suggest that the impact of different keyword extraction methods on perceived relevance is largely independent of their impact on article interests.

4.3 The Effect of Providing Explanations on Participants' Perceptions and Behavioral Intentions (RQ2)

In the Gold Standard baseline condition, providing explanations has a strong positive effect on participants' perceived relevance ($b = 0.789, p < 0.001$). When participants are shown the gold-standard keywords that are used to generate the specific recommendation, they evaluate the ad-article pair as more relevant than when no such information is provided, arguably because these explanations offer an explicit and credible source that helps participants perceive the relevance between the ad and news article.

In contrast, the model reveals that providing explanations negatively affects participants' click and purchase intentions ($b = -0.079, p = 0.043$). Beyond this direct effect, there is also a mediated effect. Providing explanations reduces their interest in the ad ($b = -0.367, p < 0.001$), which in turn has a strong effect on their behavioral intentions ($b = 0.793, p < 0.001$). These direct and indirect pathways are sufficiently strong that the mediating role of perceived relevance in positively linking explanations to behavioral intentions becomes secondary—the total effect of explanation on behavioral intention is $b = -0.303$.

4.4 The Moderation Role of Providing Explanations in the Effects of Keyword Extraction Methods on Perceptions and Behavioral Intentions (RQ3)

The SEM results also show a significant interaction effect between extraction methods and keyword-based explanations ($\chi^2(3) = 53.693, p < .001$). Whereas in the absence of keyword-based explanations, the algorithmic methods perform better than the Gold

Standard in shaping participants' perceived relevance (see Section 4.2), the advantage of these methods on perceptions diminishes to varying degrees when explanations are provided. Figure 3a shows that while providing explanations strongly enhances perceived relevance for the Gold Standard (see Section 4.3), this benefit is substantially weaker for TF-IDF (difference: $b = -0.648, p < 0.001$)[5] and KeyBERT (difference: $b = -0.751, p < 0.001$). In these conditions, the effect of keyword-based explanations is negligible, and the performance of these algorithms also does not exceed the Gold Standard baseline with keyword-based explanations. Only for the DeepSeek method do keyword-based explanations show a significant positive enhancement comparable to the effect of explanations on the baseline method (difference: $b = 0.078, p = 0.619$), allowing this method to outperform the baseline even when keyword-based explanations are shown to the user. These findings suggest that some extracted keywords that perform well algorithmically in matching ad-article pairs do not always translate into higher perceived relevance when directly observed by participants.

By contrast, there is no significant evidence that providing explanations moderates the effects of methods on behavioral intentions. As reported in Sections 4.2 and 4.3, explanations primarily influence intentions directly or through participants' interests in ads, largely independent of the method employed.

4.5 The Mediation Role of Perceived Relevance in the Effects of Keyword Extraction Methods on Behavioral Intentions (RQ4)

Although the model does not reveal a significant direct effect of keyword extraction methods on participants' click intention and purchase intention, participants' behavioral intentions are found to be positively influenced by their perceived relevance ($b = 0.101, p < 0.001$). This suggests that more relevant ad-article recommendations significantly increase participants' intentions to click on or buy the advertised products or services. Correspondingly, the keyword extraction methods indirectly shape intentions through their influence on perceived relevance.

4.6 The Effect of Situational Aspects on Participants' Perceptions and Behavioral Intentions

Beyond keyword extraction methods and explanations, the SEM analysis reveals that participants' perceived relevance and their behavioral intentions are also influenced by the type of ads shown. Part of these effects is mediated by relative differences in participants' interest in the ad, which has a significant effect on both perceived relevance ($b = 0.354, p < .001$) and on behavioral intentions ($b = 0.793, p < 0.001$). As these results are not germane to our main research questions, we cover their overall implications below:

Interest in the ad: Participants are most interested in the vacation services ad (Ad 1) and the automotive products ad (Ad 3), followed by the contraception ad (Ad 4) and the health supplements ad (Ad 2). They are least interested in the laxatives ad (Ad 5).

Perceived relevance of the ad to the article: Combining the direct and indirect effects, we find that participants see the best article matches for the automotive products ad (Ad 3) and the contraception ad (Ad 4), followed by the vacation services ad (Ad 1) and the health supplements ad (Ad 2). The laxatives ad (Ad 5) has the least relevant article matches.

Behavioral intention towards the ad: Again combining the direct and indirect effects, we find that participants have the highest behavioral intentions toward the automotive ad (Ad 3), followed by the vacation ad (Ad 1), then the health supplements ad (Ad 2), and finally the contraception ad (Ad 4) and the laxatives ad (Ad 5).

The SEM analysis also shows that an apparent match between the topic of the ad and the participant's self-reported interests has an effect on their interest in the ad (as expected), their behavioral intention toward the ad (as expected), and their interest in the article. When the ad matches participants' self-reported product or service interests, they express a significantly higher interest in the ad ($b = 0.391, p < 0.001$), their intentions toward the ad increase significantly ($b = 0.078, p = 0.034$), and so does their interest in the news article ($b = 0.259, p < 0.001$). The latter two effects are further strengthened by a mediation effect via ad interest, which has a significant impact on both behavioral intention ($b = 0.793, p < 0.001$) and article interest ($b = 0.506, p < 0.001$). The fact that matched ad interest has a significant effect on article interest is an important finding to justify the core principle behind contextual advertising: ads are placed with congruent articles with the assumption that users who are interested in a certain product or service are more likely to read articles about topics congruent with those products and services.

Finally, participants' perceptions of *how* the ad is selected for the article have a significant impact on their interest in the article ($b = 0.247, p < 0.001$). Specifically, participants perceived the articles as more interesting when they believed the articles had been matched to ads by algorithms, rather than when they believed the match was a non-algorithmic selection.

These findings reflect the fact that the extraction methods and explanation mechanisms only have a small effect on participants' interest in and behavior toward them. The ad itself and whether its topic matches the participants' interests are much more salient factors. These effects, as well as the effect of the ad's topic on perceived relevance, should be interpreted in a broader context rather than being tied to the specific ads and article dataset used in this study. For instance, our ad about laxatives is emblematic of a product category that faces inherent challenges in finding closely matching articles in the first place, and the user base that would be interested in such ads is also smaller.

Importantly, though, we find that the effect of ads is *independent* from the rest of the model, meaning that the effects of selection method and explanation apply across ad domains. Therefore, in the following sections, we focus more on the other two manipulations (i.e., keyword extraction methods and providing explanations).

[5]Note that these estimates are interaction effects between providing explanation and specific keyword extraction methods. It means that the effect of explanation on TF-IDF is lower than it is on the Gold Standard. A more direct marginal effect comparison is illustrated in Figure 3.

Table 3: Comparisons of keywords generated by TF-IDF, KeyBERT, and DeepSeek against gold-standard keywords. Metric scores for each method are averaged across the combination of their matched articles.

Keywords	Prec.@5	Recall@5	F1@5	Prec.@10	Recall@10	F1@10	Cosine Sim.	MRR@10	MAP@10
TF-IDF	0.091	0.120	0.101	0.065	0.169	0.092	0.874	0.254	0.082
KeyBERT	0.015	0.020	0.017	0.010	0.025	0.014	0.888	0.035	0.009
DeepSeek	**0.135**	**0.171**	**0.144**	**0.108**	**0.261**	**0.147**	**0.900**	**0.377**	**0.135**

Keywords	ROUGE 1			ROUGE 2			ROUGE L		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
TF-IDF	0.157	0.241	0.178	0.022	0.035	0.025	0.125	0.199	0.144
KeyBERT	0.114	0.321	0.161	0.016	0.043	0.022	0.089	0.257	0.126
DeepSeek	**0.239**	**0.442**	**0.295**	**0.079**	**0.147**	**0.097**	**0.187**	**0.349**	**0.232**

4.7 Objective Performance

In this section, we report algorithmic performance relative to the gold standard (Table 3), following the benchmarking procedure. For a meaningful comparison with participant assessments, the evaluation was based on the 125 articles retrieved by the Gold Standard, TF-IDF, KeyBERT and DeepSeek methods in the experiments (see Section 3.3.4), reduced to 92 after removing duplicates. Several evaluation metrics were used for the comparison [28, 36]. Results are provided for *Precision@K, Recall@K* and *F-score@K*, where K denotes the number of top-k ranked keywords considered in the evaluation. In our case, since all algorithms extracted ten keywords per article but the gold-standard sets vary in size, with the majority having fewer than ten elements, we evaluate $K = 5$ and $K = 10$. In addition, *ROUGE 1/2/L* [51] are reported to enable a wider comparison by measuring overlaps between the extracted and gold-standard keywords at different n-gram levels. Specifically, *ROUGE 1* and *ROUGE 2* measure the overlap of unigrams and bigrams between the evaluated and gold-standard keyword sets, whereas *ROUGE L* focuses on the longest common subsequence. To complement these metrics based on exact matches, *cosine similarity* is included to capture contextual similarities between the baseline and compared keywords. We also report *Mean Reciprocal Rank* (MRR) [53] and *Mean Average Precision* (MAP) [40] to account for ranking in the algorithmic keyword extraction, where the former gives higher scores to an algorithm that places the first correct keyword higher in the ranking, while MAP captures overall ranking quality across all correctly extracted keywords.

Results show that DeepSeek consistently outperforms TF-IDF and KeyBERT across all metrics, although the performance gap is less pronounced in terms of cosine similarity. Its performance on MRR and MAP further indicates that DeepSeek not only excels in exact and context matching, but also in ranking the most relevant keywords earlier in the set. In contrast, KeyBERT achieves the lowest scores across most metrics except cosine similarity, suggesting that it may be better at capturing semantically related keywords. TF-IDF, although not as good as DeepSeek, has a better performance than KeyBERT. All three algorithms perform comparably to the gold standard in terms of cosine similarity, yet show significant gaps on the other metrics. This indicates that while they are relatively effective at capturing context-level information, they struggle to reproduce the specific lexical choices found in the pre-annotated gold standards. However, in real-world applications, such as contextual advertising, capturing context is more valuable than word-level correspondence, as the primary goal is to match relevant content.

5 Discussion

In line with our research questions and the conceptual model, the SEM results confirm that keyword extraction methods affect participants' perceived relevance of ad-article pairs and that such relevance subsequently influences their behavioral intentions. Going beyond our research questions, we note that ad interest is another key factor in driving participants' behavioral intentions toward the ads. We discuss a theory-driven explanation of this process in Section 5.1. However, we find that the keyword extraction method leading to the highest perceived relevance depends on whether keywords are shown. Without keyword-based explanations, simpler extraction methods (TF-IDF and KeyBERT) outperform the more complex method (DeepSeek). With keyword-based explanations, DeepSeek outperforms all methods. In Section 5.2, we contextualize these results with algorithmic performance metrics reported in Section 4.7. Moreover, while providing keyword-based explanations generally benefits relevance judgments, it harms participants' intention to engage with the ad, an effect that is largely mediated by the ad interest. This naturally raises the question of what triggers the decline in ad interests and whether the underlying mechanism differs across keyword extraction methods. This issue will be addressed in Section 5.3. Finally, based on the findings and discussions, we provide insights into practical implications for users, future research, and practice in Section 5.4.

5.1 Effective Ads are Both Relevant and Interesting

Our results suggest that users' intention to engage with an ad is highest when contextual relevance activates ad-related associations (i.e., making them more accessible) and these associations align with the user's intrinsic interest. In this way, perceived relevance and overall interest jointly drive engagement. When users perceive that an ad conceptually aligns with the article, they experience smoother processing and higher perceived ad effectiveness. In addition, an interesting and well-aligned ad can stimulate curiosity and prompt further exploration, as reflected by one participant who noted: *"The ad made me be curious to read the article"*. Rather than ignoring or

resisting the ad, users may instead interpret it as an extension of the content flow, motivating them to seek more information.

This observation aligns with prior research on cognitive priming, which indicates that contextual stimuli can influence perceptions without conscious awareness when they are perceived as conceptually related [66, 78]. However, our findings further reveal that relevance alone does not determine users' behavioral intentions. Users' interest in the ad also plays a critical role in shaping purchase and click intentions. This suggests that effective advertisements operate through two complementary mechanisms: analytical relevance priming and affective interest attractions, which are discussed further in Section 5.3. Therefore, optimizing both dimensions is essential for fostering positive responses.

5.2 Misalignment of Algorithmic Performance and User Perception

Precision-based evaluation results in Table 3 indicate that DeepSeek consistently outperforms TF-IDF and KeyBERT across automated benchmark metrics. However, this superiority does not uniformly translate into user perceptions. Under conditions where explanations are not provided, participants rate simpler methods as generating more relevant ad-article matches than DeepSeek. This misalignment reveals that lightweight methods may be more effective for user-oriented tasks without explanations, and further suggests that algorithm evaluation should incorporate user perception rather than relying exclusively on technical benchmarks. At the same time, DeepSeek's benchmarking advantage is less pronounced on cosine similarity than on the other metrics. For this metric, TF-IDF and KeyBERT achieve performance comparable to the larger-scaled model. This context-level effectiveness of the algorithms is supported by participants' ratings. Figure 3a compares different methods against the Gold Standard in terms of participants' perceived ad-article relevance. It also contrasts conditions with and without showing extracted keywords as explanations. Relative to the baseline, all methods except Random improve perceived relevance to varying degrees. This pattern also indicates that the Gold Standard may not align with participants' perceptual judgments, complicating straightforward interpretations of algorithmic superiority. Interestingly, without explanation, TF-IDF and KeyBERT yield increases that are both higher than DeepSeek, despite DeepSeek's better performance on our benchmark evaluation metrics.

The pattern shifts when explanations are introduced. When keywords are presented, DeepSeek yields the largest increase in perceived relevance, while TF-IDF and KeyBERT achieve only marginal benefits relative to their no-explanation baselines. These findings reinforce the observation that while all methods are capable of capturing contextual alignment, they differ in their ability to communicate or make that alignment perceptible to end users. More importantly, the presence of explanations substantially complicates interpretations of method effectiveness and user preferences, revealing dynamics that extend far beyond what benchmark evaluations alone can capture. In the absence of explicit keyword explanations, the simplest method, TF-IDF, slightly outperforms DeepSeek on participants' perceptions, highlighting its practical value (see Figure C.1 in Appendix C) despite lower benchmark scores. When the extracted keywords are presented, however, DeepSeek shows a marked increase in perceived relevance, whereas TF-IDF and KeyBERT exhibit no comparable gains. A plausible explanation is that simpler methods generate coarser, more mechanically derived keywords that appear less natural or human-like to users than those produced by LLMs, even though they effectively capture contextual information. These findings underscore how lightweight methods can perform well in modeling conceptual relevance but fall short in producing extraction-level representation. Therefore, rather than weakening our moderation findings, the limited perceptual validity of the Gold Standard clarifies their underlying logic: explanations amplify relevance that users already recognize, rather than uniformly creating it across methods.

While all the algorithms demonstrate strong context-level effectiveness across varying model scales and complexities, the inclusion of explanations shapes how users perceive their outputs. The next section examines the multifaceted role of explanations in detail.

5.3 The Role of Explanations as Dual Activators

The SEM analysis shows that the inclusion of explanations has opposing effects on perceived relevance and behavioral intention (see Section 4.4). Figure 3a, Figure 3b and Figure 3c show marginal effects across keyword extraction methods and the presence versus absence of explanations for perceived relevance, behavioral intention, and interests in the ad. Providing explanations strengthens the perceived connection between the ad and article (i.e., relevance), which is typically expected to enhance behavioral intentions (see Section 4.5), yet it simultaneously reduces liking for the ad, ultimately lowering click and purchase intentions.

The DPT and ELM theories [29, 43, 62, 71] can help explain these contrasting findings. Consistent with prior HCI research [47, 48, 50, 54, 82], our findings indicate that both perceived relevance and behavioral intention reflect contributions from dual-process mechanisms. Via the System 1 route, participants' ad interest—an affective, heuristic-driven response—emerges as a strong driver of their intention to engage with an ad. In contrast, judgments of how relevant an ad is to the accompanying article appear to rely more on System 2 processing, involving deliberate, analytical evaluation of the fit between ad and context. In our case, we argue that providing explanations functions as a dual activator of System 2 processing: (1) explanations offer participants a concrete reference point to support relevance-evaluation goals, and (2) explanations interfere with the intuitive affect-driven basis of ad interest.

The latter may be explained by the fact that explanations activate participants' awareness of AI intervention. Figure 4 demonstrates that although the majority of the participants already think ads are algorithmically selected, this belief becomes more prevalent when explanations are available. In other words, explanations make the (presumed) algorithmic nature of the ad-article match more apparent. These findings can also be interpreted through the PKM. Based on the PKM, such explanations prompt users to "turn on" their persuasion knowledge. A "change of meaning" occurs when they perceive the explanation as a persuasion tactic and this triggers them to reflect on why the tactic is being used, thereby engaging more deliberate, System 2 processing. This deeper processing then shapes their beliefs and attitudes toward both the advertised product and the advertiser. When users detect a persuasion tactic but

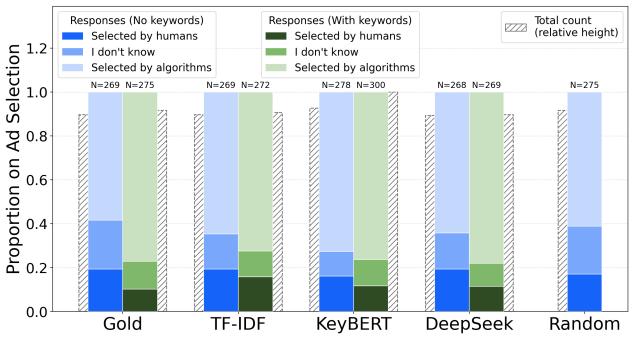


Figure 4: Proportion of participants' responses on Ad Selection. Shadow bars show the total count of all responses under the specific condition (methods × with/without keywords), with the relative height to the maximum count ($N = 300$).

judge it to be ineffective, they may question the professionalism of the advertiser and the quality of the product. In our case, explanations made participants less interested in the ad, especially when they felt the keywords or the recommendation were poor. In addition, as the PKM suggests, users develop beliefs about the appropriateness, such as fairness and manipulativeness, of particular persuasive tactics. In online advertising contexts, this persuasion knowledge may lead users to suspect that their personal data has been collected and analyzed to generate these recommendations, making them feel unsafe or manipulated, and reducing their ad interests. Similarly, when particular persuasion tactics are recognized, users may perceive a threat to their freedom of choice, which further harms their attitudes toward the ad. Arguably, this increased salience of the algorithm may distract users from the impulsive, affective path and redirect them towards more deliberate reasoning.

Along similar lines, several studies [20, 65] have shown that revealing information about the decision-making system may activate users' situation awareness, which results in poorer outcomes. While providing explanations is often expected to enhance transparency, users' trust and perceptions are shaped by more complex mechanisms. As noted before, when participants are informed through explanations, a larger share infer that the ad is selected by algorithms (see Figure 4). This suggests that such explanations may increase participants' skepticism and feelings of being manipulated, an effect consistent with prior research [52, 63].

Although negative effects exist, the role of explanations in fostering long-term trust remains under-explored. In addition, prior work has encouraged the design of more user-centric and comprehensive explanations [14, 55]. In our study, revealing keywords as explanations is closely related to the recommendation content, thereby playing a positive and authoritative role in helping participants evaluate relevance. However, because the explanations do not explicitly clarify that recommendations are independent of personal information, participants may have perceived them as personalized targeting, which could have undermined their intention. Since contextual advertising is inherently advantageous in protecting privacy, more research is needed to determine how explanations can be designed to positively shape user perceptions.

5.4 Practical Implications

For users, our findings highlight the importance of developing a more reflective awareness of contextual advertising. First, perceived relevance should not automatically be interpreted as evidence of personal profiling. Substantial efforts have been devoted to developing solutions that address privacy and ethical concerns. However, this does not mean that users should uncritically accept such ads. Instead, understanding how contextual alignment works allows users to make more deliberate judgments about whether an ad truly meets their needs or whether its appeal is simply a byproduct of contextual relevance, thereby potentially helping mitigate impulsive purchases in such situations. Second, enhancing reflective awareness benefits both individuals and society. When opportunities for feedback are available, user responses are critical in shaping responsible advertising practices and holding industry parties accountable. While advertisers must provide transparent and accessible choices, users' proactive engagement helps guide the ecosystem to greater accountability and alignment with public values.

For researchers, the misalignment observed in our study between precision-based benchmarking results and user perceptions underscores the limitations of relying solely on gold standard benchmarks to evaluate algorithms. More importantly, the quality and provenance of the gold standard itself may further exacerbate this misalignment. Although we employed an open-source and well-known dataset as the Gold Standard, our findings indicate that it may exhibit shortcomings. Its broad keyword associations may generate matches that appear arbitrary or weakly related from a user's perspective. Some participants expressed dissatisfaction with the ad-article matches generated by the Gold Standard in the free-text feedback box: *"They picked up on very broad keywords"* (with explanations); *"Poorly written algorithms, disreputable advertising companies"*; and *"I believe that the ads were chosen at random"*.

Gold standards are typically constructed using human expert annotations, including those applied in this study, but the user feedback above indicates that they can still lack contextual relevance from the user's perspective or fail to reflect real interaction dynamics with users. This phenomenon has also been reported in previous studies [15, 33]. Specifically, in [15], users were also found to evaluate keywords chosen by other users negatively. Consequently, evaluating algorithms exclusively against such benchmarks risks widening the gap between offline performance and user perspectives. Therefore, beyond constructing gold standard keyword datasets that better capture semantic nuance, contextual meaning, and user expectations, it is also essential to integrate complementary evaluation dimensions when assessing algorithms. These include user-centered perception and intention measures, qualitative feedback, and objective performance indicators, thereby enabling a more comprehensive understanding of both algorithms and users in AI-driven, user-facing tasks.

For practitioners and advertisers, and in line with similar research [39, 73], engaging advertisements and appealing products are key drivers of purchase intention, and relevance can further amplify this effect. Nonetheless, precise targeting should not be viewed as the sole determinant. The creative idea, visual appeal, and inherent product attractiveness embedded within ads are equally

essential. As one participant wrote: *"The ads were visually appealing, even if I had no interest, I would still want to check them out".*

More importantly, users' privacy and ethical concerns should be respected. For example, one participant expressed their concerns about online ads: *"In today's world, more and more people are learning about the dangers of clicking on random ads. They could lead you to malicious websites or even track your information...I don't trust ads...".* To address such concerns, advertisers should, for instance, provide notifications about data collection or offer explanations regarding the ad placement, thereby enhancing transparency and trust. However, such explanations must be used cautiously. As observed in our findings, transparency can be beneficial, but may simultaneously reduce user interest in ads and behavioral intentions by activating their persuasion knowledge. Therefore, when constructing explanations, advertisers should ensure they are simple, clear, and neutral, disclosing neither excessive behind-the-scenes technical details that may trigger skepticism, nor ambiguous information that undermines credibility. They must clarify whether personal data is being collected or analyzed, aligning with regulatory requirements and improving user experience in the long run.

Furthermore, since explanations may shift users from intuitive System 1 processing to more deliberate System 2 reasoning, explanations should foster emotional connection rather than disrupt intuitive engagement. At the same time, such emotional connections must be perceived as positive and beneficial to users, as they may otherwise trigger persuasion awareness and elicit defensive or hostile responses [24, 26]. This requires reducing uncertainty through understandable, contextually relevant rationales; emphasizing added value and user-oriented intentions rather than merely signaling compliance; and balancing transparency with persuasion sensitivity, so as to support user engagement, strengthen trust, and ultimately contribute to a more satisfying advertising experience.

In addition, when selecting keyword extraction methods for ad placement, advertisers should adopt a holistic perspective that integrates the objective aspects of algorithms, resource constraints, user perceptions, and the effects of transparency mechanisms. Our findings show that although TF-IDF and KeyBERT perform worse than DeepSeek on benchmark evaluations, they nevertheless yield higher perceived relevance in user judgments when explanations are not provided. Moreover, their computational efficiency and ease of deployment and maintenance further strengthen their practical appeal in real-world settings. However, when keyword-based explanations are introduced, DeepSeek demonstrates clear advantages in shaping user perceptions, possibly because it generates more relevant, contextually natural, and human-like keyword representations. This pattern suggests that lightweight methods can be highly effective when explanations are absent or when explanations do not explicitly expose the model's internal reasoning. In such cases, simpler approaches are sufficient for delivering effective ad-article matches, as they are not required to interact with users through their surface-level outputs. By contrast, when explanations are required or when they become central to system design, more sophisticated models are beneficial in retaining user engagement with ads and mitigating the negative perceptual effects of explanations.

6 Limitations and Future Work

Our study focuses on keyword-based contextual advertising, abstracting away more complex advertising systems. However, our main goal was not to replicate real-world environments or demonstrate system-level performance. Instead, we employed this effective and widely used strategy as a means of exploring how algorithms and transparency shape user perceptions and intentions, and to illustrate the gap between conventional benchmarking and user-centered evaluation. Likewise, the keyword-based explanations used in this study also serve as representative illustrations rather than full system implementations. Future research could incorporate more ecologically grounded architectures, commercial pipelines, and diverse explanation formats.

To focus annotation resources and minimize task fatigue, this study is limited to the three mainstream algorithms TF-IDF, KeyBERT and DeepSeek, which may limit the generalizability of the findings. Nonetheless, the methods represent key techniques across different algorithmic complexities, and we hope that the preliminary findings of this study will stimulate further research attention towards practical user experiences in this domain. Although DeepSeek is trained on a multilingual corpus, it has demonstrated performance comparable to the state-of-the-art models and performed well in our study. While we relied on its API, it remains feasible for future research or practices to run the model locally, enabling greater transparency and reproducibility. Nonetheless, with the rapid development of LLMs, an ever-expanding range of models is becoming available. Incorporating more models and comparing their impacts would be a promising direction for future work, although such exploration was beyond the scope of this study.

Similarly, we restrict to five ads to maintain participant attention. However, the ads cover diverse categories and yield consistent outcomes. Given that our emphasis is on keyword extraction methods and explanations, the ads serve as proxies for broader advertising content rather than an exhaustive sample.

Participants were recruited from the U.S. to match the origin of news articles. While our manipulations are not inherently culture-specific, future work could incorporate cross-cultural samples to assess external validity. We did not control for participants' prior attitudes towards ads, which may have affected their judgments, although random assignment should have mitigated systematic impact. Future work could consider such pre-existing tendencies.

Finally, participants' ratings of ad-article relevance without explanations show that, not only do algorithms surpass the Gold Standard, but even the Random outperforms the KPTimes Gold Standard (see Figure 3a), raising questions about the quality of such gold standards. One possible reason is that many sets of the Gold Standard contain fewer than ten keywords, which can constrain ad-article matching. However, when explanations are available, the perceived relevance of the gold-standard condition increased, suggesting participants value the displayed keywords. In addition, since our algorithms are not trained on these gold-standard keywords, the main effect we observe on perceptions and intentions, and the impact of explanations, appear robust to the quality of the gold standard. More broadly, these highlight the gap discussed in the paper, that users' perceptions may not align with typical benchmark evaluations against pre-annotated gold standards, particularly

when the source is unclear. Future work could compare alternative gold standards to gain deeper insights.

7 Conclusion

This study investigates how keyword extraction algorithms and ad placement transparency shape user perceptions and intentions within a controlled contextual advertising scenario. Rather than replicating industrial ad-tech pipelines, we abstract the problem to a tractable setting. It zooms in on the perceptual mechanisms that underpin how users evaluate ad-article relevance and respond to explanatory cues, allowing us to surface user-centered dynamics that may otherwise be obscured in more complex environments.

Results indicate that keyword extraction algorithms significantly impact users' perceived relevance of the ad to the article, and this perception in turn positively affects intention. Notably, even simple algorithms can generate ad-article matches that users perceive as sufficiently relevant, although they perform worse in precision-oriented benchmarking evaluations. These findings highlight the long-overlooked gap between automatic benchmarking and user-centered assessments, suggesting that lightweight methods remain attractive in practice for their computational efficiency, ease of deployment, and sufficiently good user experience. Meanwhile, transparency introduces complex outcomes. Providing explanations strengthens perceived relevance, with DeepSeek showing particular robustness when explanations are available. However, the presence of explanations can also reduce ad interest and intention, reflecting their complex role in shaping user engagement.

While our study does not mirror the full complexity of commercial ad-tech, the perceptual processes revealed here offer insights and implications relevant across contexts where users interpret contextual ads and explanations. The findings point to the need for future interdisciplinary research into different explanation designs and their downstream effects on user engagement. Ultimately, algorithmic performance and transparency in real-world advertising contexts are not just about precision and accountability, but more about aligning with what users truly value in practice.

Acknowledgments

This work was funded by the Marianne and Marcus Wallenberg Foundation and the Wallenberg AI under grant number 2020.0095, and supported by the Autonomous Systems and Software Program.

References

[1] 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L 119, pp. 1–88. http://data.europa.eu/eli/reg/2016/679/oj Entered into application on 25 May 2018.

[2] 2018. California Consumer Privacy Act of 2018 (Assembly Bill No. 375). California Civil Code §§ 1798.100–1798.199. https://oag.ca.gov/privacy/ccpa Enacted June 28, 2018; amended by California Privacy Rights Act (CPRA) in 2020.

[3] 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act). Official Journal of the European Union, L 277, 27 October 2022, pp. 1–102. http://data.europa.eu/eli/reg/2022/2065/oj In force since 17 February 2024; preliminary provisions applied from 16 November 2022.

[4] Elizabeth Aguirre, Dominik Mahr, Dhruv Grewal, Ko de Ruyter, and Martin Wetzels. 2015. Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing* 91, 1 (2015), 34–49. https://doi.org/10.1016/j.jretai.2014.09.005

[5] Akiko Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing & Management* 39, 1 (2003), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

[6] Icek Ajzen. 1991. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes* 50, 2 (1991), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[7] Heba Aly, Matias Volonte, Kaileigh Angela Byrne, and Bart P. Knijnenburg. 2025. Bridging the Trust Gap: Investigating the Role of Trust Transfer in the Adoption of AI Instructors for Digital Privacy Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 22 pages. https://doi.org/10.1145/3706598.3713570

[8] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Bangkok, Thailand, 386–397. https://aclanthology.org/2024.acl-short.37/

[9] A. Aribarg and E. M. Schwartz. 2019. Native Advertising in Online News: Trade-Offs Among Clicks, Brand Recognition, and Website Trustworthiness. *Journal of Marketing Research* 57, 1 (2019), 20–34. https://doi.org/10.1177/0022243719881112

[10] Wayne E. Baker and Richard J. Lutz. 2000. An Empirical Test of an Updated Relevance-Accessibility Model of Advertising Effectiveness. *Journal of Advertising* 29, 1 (2000), 1–14. https://doi.org/10.1080/00913367.2000.10673599

[11] Natã M. Barbosa, Gang Wang, Blase Ur, and Yang Wang. 2021. Who Am I? A Design Probe Exploring Real-Time Transparency about Online and Offline User Profiling Underlying Targeted Ads. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (Sept. 2021), 32 pages. https://doi.org/10.1145/3478122

[12] Peter M. Bentler and Douglas G. Bonett. 1980. Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin* 88, 3 (1980), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

[13] Sophie C. Boerman, Lotte M. Willemsen, and Eva P. Van Der Aa. 2017. "This Post Is Sponsored": Effects of Sponsorship Disclosure on Persuasion Knowledge and Electronic Word of Mouth in the Context of Facebook. *Journal of Interactive Marketing* 38 (2017), 82–92. https://doi.org/10.1016/j.intmar.2016.12.002

[14] Jean Burgess, Nicholas Carah, Daniel Angus, Abdul Obeid, and Mark Andrejevic. 2024. Why Am I Seeing This Ad? The affordances and limits of automated user-level explanation in Meta's advertising system. *New Media & Society* 26, 9 (2024), 5130–5149. https://doi.org/10.1177/14614448241251796

[15] Jingwen Cai, Sara Leckner, and Johanna Björklund. 2026. From precision to perception: Human-in-the-loop evaluation of keyword extraction for internet-scale contextual advertising. *Information Systems* 138 (2026), 102665. https://doi.org/10.1016/j.is.2025.102665

[16] Margret C. Campbell and Amna Kirmani. 2000. Consumers' Use of Persuasion Knowledge: The Effects of Accessibility and CognitiveCapacity on Perceptions of an Influence Agent. *Journal of Consumer Research* 27, June (2000), 69–83. https://doi.org/10.1086/314309

[17] Richard L. Celsi and Jerry C. Olson. 1988. The role of involvement in attention and comprehension processes. *Journal of Consumer Research* 15, 2 (1988), 210–224. https://doi.org/10.1086/209169

[18] Wen-Kuo Chen, Chia-Ju Ling, and Chien-Wen Chen. 2023. What affects users to click social media ads and purchase intention? The roles of advertising value, emotional appeal and credibility. *Asia Pacific Journal of Marketing and Logistics* 35, 8 (2023), 1900–1916. https://doi.org/10.1108/APJML-01-2022-0084

[19] Kwang Yeun Chun, Ji Hee Song, Candice R. Hollenbeck, and Jong-Ho Lee. 2014. Are contextual advertisements effective? *International Journal of Advertising* 33, 2 (2014), 351–371. https://doi.org/10.2501/IJA-33-2-351-371

[20] Mary L. Cummings. 2004. *Automation Bias in Intelligent Time Critical Decision Support Systems*. Vol. 2. 557–562. https://arc.aiaa.org/doi/abs/10.2514/6.2004-6313

[21] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[22] Dentsu. 2025. Ad Spend Forecast To Grow By 4.9% In 2025, Despite A Reduced Economic Outlook. Online. https://www.dentsu.com/news-releases/ad-spend-forecast-to-grow-by-four-point-nine-percent-in-2025-despite-a-reduced-economic-outlook

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://aclanthology.org/N19-1423

[24] Leyla Dogruel. 2019. Too much information!? Examining the impact of different levels of transparency on consumers' evaluations of targeted advertising. *Communication Research Reports* 36, 5 (2019), 383–392. https://doi.org/10.1080/08824096.2019.1671486

[25] Robert H. Ducoffe. 1996. Advertising Value and Advertising on the Web. *Journal of Advertising Research* 36, 5 (1996), 21–35. https://doi.org/10.1080/00218499.1996.12466626

[26] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174006

[27] Wenfang Fan, Bingjia Shao, and Yong Zhang. 2025. The more congruent, the better? The role of recommendation content congruence on consumers' click-through intention in in-feed advertising. *Journal of Retailing and Consumer Services* 87 (2025), 1–17. https://doi.org/10.1016/j.jretconser.2024.103831

[28] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering* 26, 3 (2020), 259–291. https://doi.org/10.1017/S1351324919000457

[29] Keith Frankish and Jonathan St. B. T. Evans. 2009. The duality of mind: An historical perspective. In *In two minds: Dual processes and beyond*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0001

[30] Marian Friestad and Peter Wright. 1994. The Persuasion Knowledge Model: How People Cope with Persuasion Attempts. *Journal of Consumer Research* 21, 1 (1994), 1–31. https://www.jstor.org/stable/2489738

[31] Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 130–135. https://doi.org/10.18653/v1/W19-8617

[32] Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint* (2018). https://doi.org/10.48550/arXiv.1803.01937

[33] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 28 pages. https://doi.org/10.1145/3555088

[34] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. v0.3.0 [software], Zenodo. https://doi.org/10.5281/zenodo.4461265

[35] Lovisa Hagström, Youna Kim, Haeun Yu, Sang goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2025. CUB: Benchmarking Context Utilisation Techniques for Language Models. arXiv:2505.16518 [cs.CL] https://arxiv.org/abs/2505.16518

[36] Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, 1262–1273. https://doi.org/10.3115/v1/P14-1119

[37] Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 6 (1999), 1–55. https://api.semanticscholar.org/CorpusID:123504887

[38] Wim Janssens, Patrick De Pelsmacker, and Maggie Geuens. 2012. Online advertising and congruency effects. *International Journal of Advertising* 31, 3 (2012), 579–604. https://doi.org/10.2501/IJA-31-3-579-604

[39] Hui Jiang, Paul R. Messinger, Yifei Liu, Zhibin Lu, Shuiqing Yang, and Gang Li. 2024. Divergent Versus Relevant Ads: How Creative Ads Affect Purchase Intention for New Products. *Journal of Marketing Research* 61, 2 (2024), 271–289. https://doi.org/10.1177/00222437231187630

[40] Xin Jiang, Yunhua Hu, and Hang Li. 2009. A ranking approach to keyphrase extraction. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 756–757. https://doi.org/10.1145/1571941.1572113

[41] Yucheng Jin, Karsten Seipp, Erik Duval, and Katrien Verbert. 2016. Go With the Flow: Effects of Transparency and User Control on Targeted Advertising Using Flow Charts. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16)*. Association for Computing Machinery, New York, NY, USA, 68–75. https://doi.org/10.1145/2909132.2909269

[42] A-Reum Jung. 2017. The influence of perceived ad relevance on social media advertising: An empirical examination of a mediating role of privacy concern. *Computers in Human Behavior* 70 (2017), 303–309. https://doi.org/10.1111/10.1016/j.chb.2017.01.008

[43] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[44] Tami Kim, Kate Barasz, and Leslie K. John. 2019. Why am I seeing this Ad? The effect of Ad transparency on Ad effectiveness. *Journal of Consumer Research* 45, 5 (2019), 906–932. https://doi.org/10.1093/jcr/ucz027

[45] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[46] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the Sixth ACM Conference on Recommender Systems* (Dublin, Ireland) *(RecSys '12)*. Association for Computing Machinery, New York, NY, USA, 43–50. https://doi.org/10.1145/2365952.2365966

[47] Bart P. Knijnenburg and Burcu Bulgurcu. 2023. Designing Alternative Form-Autocompletion Tools to Enhance Privacy Decision-making and Prevent Unintended Disclosure. *ACM Trans. Comput.-Hum. Interact.* 30, 6 (Sept. 2023), 42 pages. https://doi.org/10.1145/3610366

[48] Alfred Kobsa, Hichang Cho, and Bart P. Knijnenburg. 2016. The effect of personalization provider characteristics on privacy attitudes and behaviors: An Elaboration Likelihood Model approach. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2587–2606. https://doi.org/10.1002/asi.23629

[49] Anastasia Kononova, Wonkyung Kim, Eunsin Joo, and Kristen Lynch. 2020. Click, click, ad: the proportion of relevant (vs. irrelevant) ads matters when advertising within paginated online content. *International Journal of Advertising* 39, 7 (2020), 1031–1058. https://doi.org/10.1080/02650487.2019.1694561

[50] Chia-Ying Li. 2013. Persuasive messages on information system acceptance: A theoretical extension of elaboration likelihood model and social influence theory. *Computers in Human Behavior* 29, 1 (2013), 264–275. https://doi.org/10.1016/j.chb.2012.09.003

[51] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[52] Dewen Liu, Haoding Wang, and Youping Zhu. 2025. You plan to manipulate me: A persuasion knowledge perspective for understanding the effects of AI-assisted selling. *Journal of Business Research* 200 (2025), 115598. https://doi.org/10.1016/j.jbusres.2025.115598

[53] Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, 366–376. https://aclanthology.org/D10-1036/

[54] Paul Benjamin Lowry, Greg Moody, Anthony Vance, Matthew Jensen, Jeff Jenkins, and Taylor Wells. 2012. Using an elaboration likelihood approach to better understand the persuasiveness of website privacy assurance cues for online consumers. *Journal of the American Society for Information Science and Technology* 63, 4 (2012), 755–776. https://doi.org/10.1002/asi.21705

[55] Hongyu Lu, Weizhi Ma, Yifan Wang, Min Zhang, Xiang Wang, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2023. User Perception of Recommendation Explanation: Are Your Explanations What Users Need? *ACM Transactions on Information Systems* 41, 2 (Jan. 2023), 31 pages. https://doi.org/10.1145/3565480

[56] Henk Lütjens, Maik Eisenbeiss, Maximilian Fiedler, and Tammo Bijmolt. 2022. Determinants of consumers' attitudes towards digital advertising – A meta-analytic comparison across time and touchpoints. *Journal of Business Research* 153 (2022), 445–466. https://doi.org/10.1016/j.jbusres.2022.07.039

[57] Scott B MacKenzie, Richard J Lutz, and George E Belch. 1986. The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of marketing research* 23, 2 (1986), 130–143. https://doi.org/10.1177/002224378602300205

[58] Virginie Maille and Nathalie Fleck. 2011. Perceived congruence and incongruence: toward a clarification of the concept, its formation and measure. *Recherche et Applications en Marketing (English Edition)* 26, 2 (2011), 77–113. https://doi.org/10.1177/205157071102600204

[59] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint* (2013). https://doi.org/10.48550/arXiv.1301.3781

[60] Weitong Ou, Bo Chen, Xinyi Dai, Weinan Zhang, Weiwen Liu, Ruiming Tang, and Yong Yu. 2023. A Survey on Bid Optimization in Real-Time Bidding Display Advertising. *ACM Transactions on Knowledge Discovery from Data* 18, 3 (2023), 1–31. https://doi.org/10.1145/3628603

[61] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Abu Dhabi, UAE, 327–337. https://doi.org/10.18653/v1/2022.emnlp-demos.33

[62] Richard E. Petty and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag, New York, NY. https://doi.org/10.1007/978-1-4612-4964-1

[63] Xiaodong Qiu, Ya Wang, Yuruo Zeng, and Rong Cong. 2025. Artificial Intelligence Disclosure in Cause-Related Marketing: A Persuasion Knowledge Perspective. *Journal of Theoretical and Applied Electronic Commerce Research* 20, 3 (2025). https://doi.org/10.3390/jtaer20030193

[64] Jill Sammet and Ralf Krestel. 2023. Domain-Specific Keyword Extraction using BERT. In *Proceedings of the 4th Conference on Language, Data and Knowledge*. NOVA CLUNL, Vienna, Austria, 659–665. https://aclanthology.org/2023.ldk-1.72

[65] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*

(Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[66] Dietram A. Scheufele and Tewksbury David. 2007. Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *Journal of Communication* 57, 1 (2007), 9–20. https://doi.org/10.1111/j.0021-9916.2007.00326.x

[67] Sigal Segev, Weirui Wang, and Juliana Fernandes. 2014. The effects of ad–context congruency on responses to advertising in blogs: exploring the role of issue involvement. *International Journal of Advertising* 33, 1 (2014), 17–36. https://doi.org/10.2501/IJA-33-1-017-036

[68] Jiasheng Sheng, Zelalem Gero, and Joyce C. Ho. 2022. PubMed Author-assigned Keyword Extraction (PubMedAKE) Benchmark. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 4470–4474. https://doi.org/10.1145/3511808.3557675

[69] Hunsik Shin, Hye Jin Lee, and Sungzoon Cho. 2023. General-use unsupervised keyword extraction model for keyword analysis. *Expert Systems with Applications* 233 (2023), 120889. https://doi.org/10.1016/j.eswa.2023.120889

[70] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21. https://doi.org/10.1108/eb026526

[71] Keith E. Stanovich and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23, 5 (2000), 645–665. https://doi.org/10.1017/S0140525X00003435

[72] Debora Trampe, Diederik A. Stapel, Frans W. Siero, and Henk Mulder. 2010. Beauty as a tool: The effect of model attractiveness, product relevance, and elaboration likelihood on advertising effectiveness. *Psychology & Marketing* 27, 12 (2010), 1101–1121. https://doi.org/10.1002/mar.20362

[73] Yuqi Wang, Guohua Liu, Liangjie Zhu, Kun Nie, and Qiong Wang. 2025. Matching is believing: the effect of congruency on purchase intention in live streaming shopping context. *Current Psychology* 44 (2025), 216–230. https://doi.org/10.1007/s12144-024-07155-2

[74] Yang Wang, Huichuan Xia, and Yun Huang. 2016. Examining American and Chinese Internet Users' Contextual Privacy Preferences of Behavioral Advertising. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 539–552. https://doi.org/10.1145/2818048.2819941

[75] Daricia Wilkinson, Moses Namara, Karishma Patil, Lijie Guo, Apoorva Manda, and Bart P. Knijnenburg. 2021. The Pursuit of Transparency and Control: A Classification of Ad Explanations in Social Media.. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. Kauai, Hawaii, USA, 1–10. https://doi.org/10.24251/HICSS.2021.093

[76] Lihua Xia and Nathalie N. Bechwati. 2008. Word of mouse: The role of cognitive personalization in online consumer reviews. *Journal of Interactive Advertising* 9, 1 (2008), 3–13. https://doi.org/10.1080/15252019.2008.10722125

[77] Yu Yang, Jie Zhang, and Tong Gao. 2024. How do username and avatar affect people's engagement with native advertising on social media: from the self-disclosure perspective. *Psychology & Marketing* 41, 6 (2024), 1289–1317. https://doi.org/10.1002/mar.22053

[78] Youjae Yi. 1990. Cognitive and Affective Priming Effects of the Context for Print Advertisements. *Journal of Advertising* 19, 2 (1990), 40–48. http://www.jstor.org/stable/4188762

[79] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. 2006. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*. Association for Computing Machinery, Edinburgh, Scotland, 213–222. https://doi.org/10.1145/1135777.1135813

[80] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2021. What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 361, 24 pages. https://doi.org/10.1145/3411764.3445459

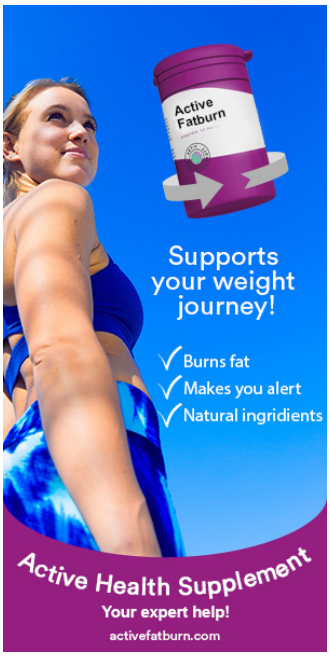
[81] Jing Zhang and En Mao. 2016. From online motivations to ad clicks and to behavioral intentions: An empirical study of consumer response to social media advertising. *Psychology & Marketing* 33, 3 (2016), 155–164. https://doi.org/10.1002/mar.20862

[82] Tao Zhou. 2012. Understanding users' initial trust in mobile banking: An elaboration likelihood perspective. *Computers in Human Behavior* 28, 4 (2012), 1518–1525. https://doi.org/10.1016/j.chb.2012.03.021

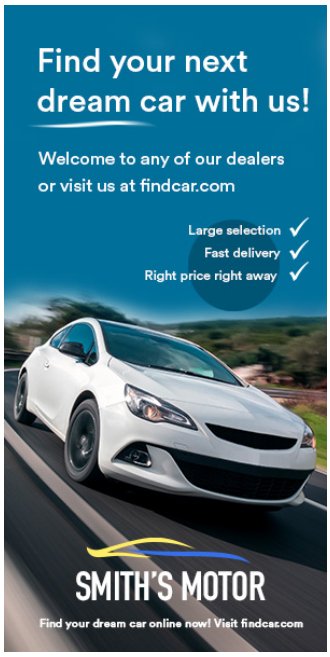
A Advertisements

The fictitious ads used in the experiments are presented below:

(a) Ad 1

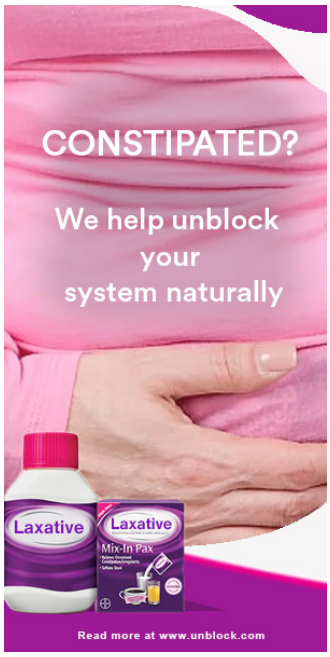


(b) Ad 2



(c) Ad 3

(d) Ad 4



(e) Ad 5

Figure A.1:

B An Example of Experimental Instance page

Here we provide an example of the experimental instance webpage used in our experiments.

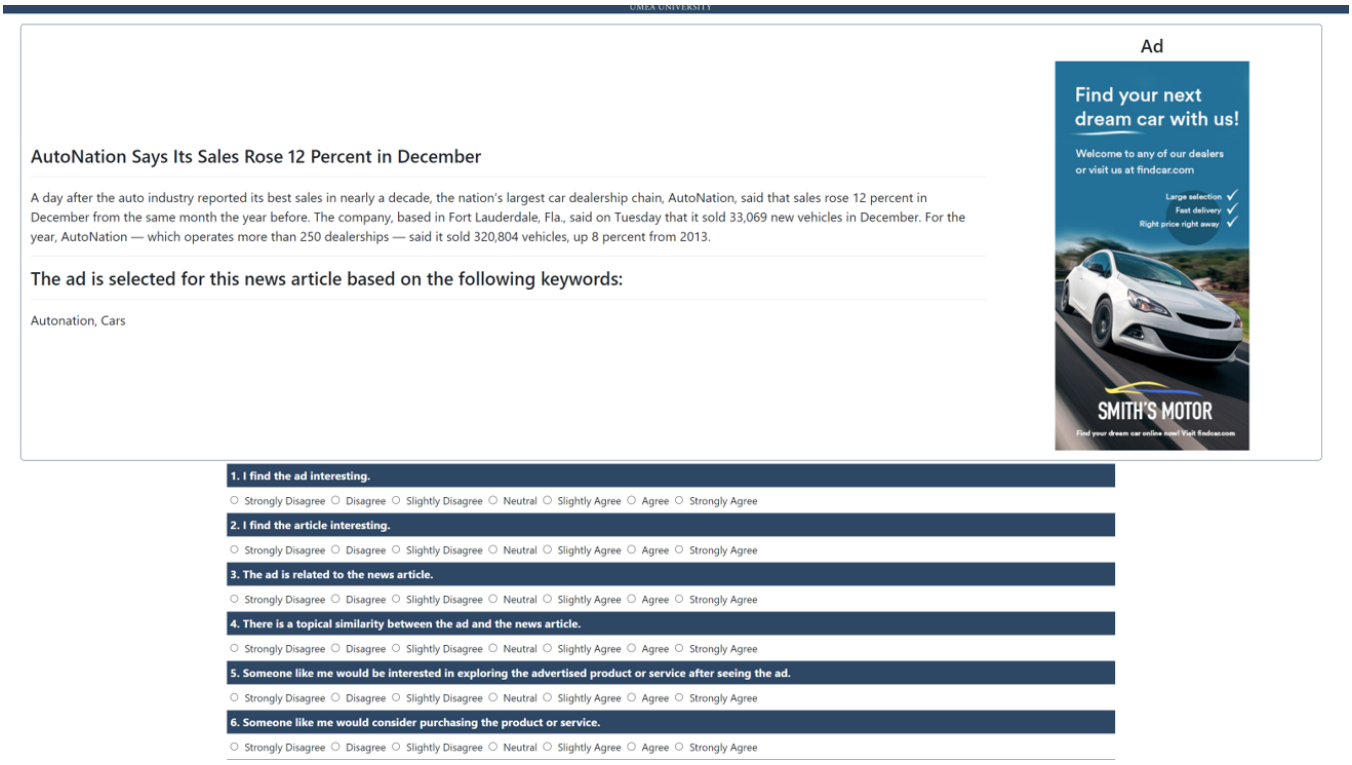


Figure B.1: An example of the instance page with the explanation.

C Computing Times

Here we report the computing times of extracting keywords using TF-IDF, KeyBERT and DeepSeek. To eliminate the influence of caching and I/O operations, each algorithm was prepared by five warm-up runs to ensure that the cache and compiler were fully initialized. In addition, all runtime measurements exclude document reading and result writing, recording only the core processing time. For DeepSeek, apart from the extraction prompt, no explicit instruction was given to discard previous conversation histories; instead, the prompt is regenerated for every keyword extraction to ensure a clean context. However, as all three algorithms are executed on a local device using a CPU, the processing times may be slower compared to those with GPUs or cloud-based services.

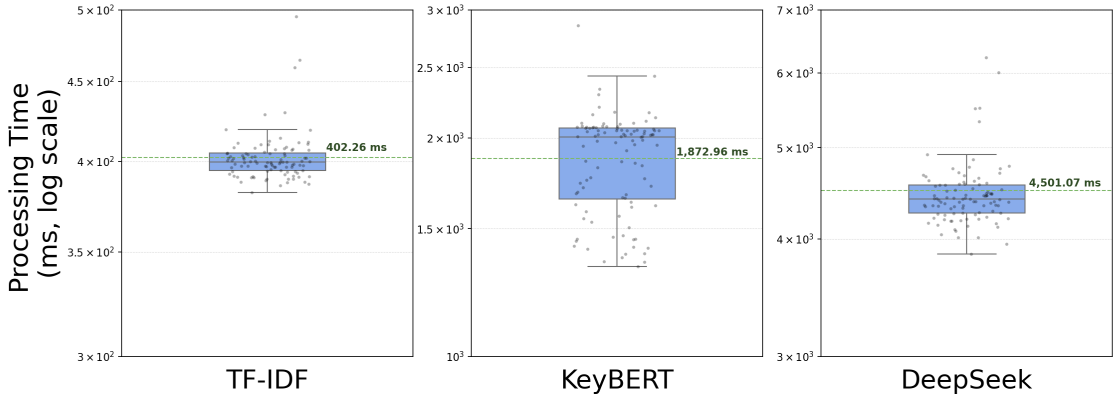


Figure C.1: Processing time of extracting keywords over 100 epochs, reported in milliseconds. Each grey dot corresponds to the time spent on extracting keywords for the same article per epoch. Horizontal orange lines show the average processing time across these 100 epochs and the error bars indicate standard deviations.

D The Results of the Confirmatory Factor Analysis

Table D.1: Standardized Loadings and Average Variance Extracted (AVE) of factors. Single-item constructs are not included in the final CFA. Items denoted with "(R)" represent reverse-worded questions, which load negatively on their respective factors.

Construct	Item	Loading
Interests of the Ad	I find the ad interesting.	-
Interests of the Article	I find the article interesting.	-
Perceived Relevance *AVE:* 0.793	The ad is related to the news article.	0.933
	The news article is related to the ad.	0.945
	The news article and the ad are unrelated. (R)	-0.860
	There is a topical similarity between the ad and the news article.	0.920
	The ad fits well with the article.	0.918
	The ad appears to be specifically chosen for this article.	0.878
	The ad appears to be randomly selected. (R)	-0.765
Behavioral Intention *AVE:* 0.746	Someone like me would be interested in exploring the advertised product or service after seeing the ad.	0.935
	Someone like me would click on the ad to learn more about the product or service.	0.921
	Someone like me is unlikely to click on the ad. (R)	-0.702
	Someone like me would consider purchasing the product or service.	0.911
	Someone like me would think the ad is persuasive.	0.784
	The ad appeals to someone like me.	0.902