A Multimodal LLM-Based Nutrition Label for Analyzing Social Media Feed Exposure

Tim Gollub, Armin Heidari, Cem Ertürkan, Benno Stein
Bauhaus-Universität Weimar, Germany
{tim.gollub, armin.heidari, cem.ertuerkan, benno.stein}@uni-weimar.de

Abstract

Algorithmically curated social media feeds shape political exposure, commercial influence, and cultural consumption, yet they remain difficult to study systematically due to limited data access and opaque recommendation mechanisms. We present a research-oriented framework that operationalizes feed-level exposure analysis using a browser extension combined with a server-side multimodal large language model (LLM). The system logs visible posts and their view time, performs zero-shot multimodal classification, and aggregates results into a customizable nutrition label summarizing exposure across analytical categories. It further supports retrieval-grounded conversational querying, dataset export and sharing, and human validation of LLM classifications. Designed as a methodological instrument for Social Sciences and Humanities, the framework enables both observational analysis and experimental research on transparency interventions, while critically examining epistemic, methodological, and ethical implications of LLM-based exposure analysis.

Keywords: Large Language Models; Multimodal Analysis; Social Media Research; Algorithmic Transparency; Exposure Measurement; Digital Humanities; Computational Social Science; Information Nutrition Label

1. Introduction

Algorithmically curated feeds increasingly mediate how individuals encounter political information, commercial messaging, and cultural content. For researchers in the Social Sciences and Humanities (SSH), understanding feed-level exposure is essential for studying algorithmic amplification, public discourse formation, and systemic risks.

However, empirical research on personalized feeds faces structural barriers. Platform APIs rarely provide access to individualized exposure streams, and recommender systems remain opaque. Regulatory developments such as the European Union's Digital Services Act (DSA) highlight this tension: Article 27 mandates transparency regarding recommender systems (European Union, 2022a), while Article 40 establishes data access provisions for vetted researchers studying systemic risks (European Union, 2022b). Yet practical methodological tools for operationalizing exposure-level analysis remain scarce.

We present a research-oriented framework that combines a browser extension with a multimodal LLM to log and analyze feed exposure. The system records visible posts and their view time, performs zero-shot classification across user-defined analytical categories, aggregates results into an exposure-oriented "nutrition label" (see Figure 1), and enables retrieval-grounded conversational querying. It further supports dataset export/import and human validation of automated classifications.

While inspired by the metaphor of consumer-facing nutrition labels, we primarily position the system as a methodological instrument for SSH research and as a testbed for studying transparency interventions.

2. Related Work

Our work builds on prior research tools for social media monitoring, most notably Zeeschuimer, a browser extension that allows users to collect data from social media feeds for research purposes (Peeters, 2025). While Zeeschuimer supports feed logging, it does not record post visibility duration and does not analyze or summarize content.

Inspired by food nutrition labels, several researchers have proposed "information nutrition labels" to communicate properties such as credibility, bias, or sourcing (Fuhr et al., 2017; Gollub et al., 2018; Kevin et al., 2018; Willinsky and Pimentel, 2024). These approaches primarily focus on item-level or publisher-level information. In contrast, our work emphasizes personalized, feed-level aggregation weighted by view time.

Most closely related is the recommender system label proposed by Belli and Wisniak (2023), which aims to reveal parameters influencing post amplification. Whereas their proposal focuses on transparency for content creators, our framework supports content consumers and researchers by aggregating exposure across customizable analytical categories.

More broadly, our work connects to emerging SSH uses of LLMs for qualitative coding, discourse analysis, and multimodal interpretation, while explicitly addressing reproducibility and epistemic concerns.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 198–204
11 May 2026.

Figure 1: Feed-level nutrition label. Aggregate statistics are computed per category and can be displayed by post count or view time (see select box at the top right) to approximate exposure. Selecting a category value enables drill-down to the corresponding posts in the local database.

3. Research Motivation

Studying exposure requires moving beyond isolated content analysis toward temporally weighted feed composition. Researchers may ask which themes dominate exposure over time, how political and commercial content are interwoven, or what emotional tones characterize curated streams. Such questions require operationalizing exposure not merely as post frequency but as time-weighted visibility, since feed consumption typically involves rapid scrolling through numerous posts with only occasional sustained attention to particular items. Capturing view time therefore provides a more meaningful approximation of exposure intensity than simple post counts alone.

Beyond observational analysis, the framework enables intervention research. The nutrition label and conversational interface can serve as transparency treatments, allowing researchers to investigate how aggregated exposure information influences scrolling behavior, engagement patterns, or perceptions of algorithmic bias.

4. System Architecture

The system consists of two components: (1) a client-side browser extension responsible for feed monitoring, storage, retrieval, and user interaction; and (2) a server-side multimodal LLM that performs semantic analysis and reasoning tasks. The architecture is designed to keep exposure datasets locally under researcher control while delegating computationally intensive multimodal inference to dedicated hardware.

In the current implementation, we deploy the open-weight multimodal model `Qwen2.5-VL-7B-Instruct` (Alibaba Cloud, 2024). The backend processes requests transiently and does not persist user data. This setup enables researchers to run the system on their own infrastructure without relying on commercial APIs.

4.1. Client-Side Data Capture

The browser extension currently operates within the Instagram web interface, with support for additional platforms such as TikTok currently under development. A content script detects which post is visible in the viewport using intersection observers and timestamp tracking. For each post, the system initially records the following captured fields:

- Post identifier
- Caption text
- Media reference (image or video URL)
- Metadata (e.g. timestamp, interaction counts such as likes and comments)
- User interaction signals (e.g., whether the user liked, saved, or commented on the post)
- View time (milliseconds visible)

All captured data are stored locally using IndexedDB. Thousands of posts can be stored without performance degradation, with local browser storage constituting the primary scalability constraint. The database supports keyword search, field-specific filtering, Boolean logic (AND, OR, NOT), and numerical comparisons ($<$, $>$, $=$) for metadata fields.

The database can be explored in an extra tab provided by the browser extension. Besides a classical table view, a gallery and feed view are available for an image/video focused exploratory analysis (Figure 2).

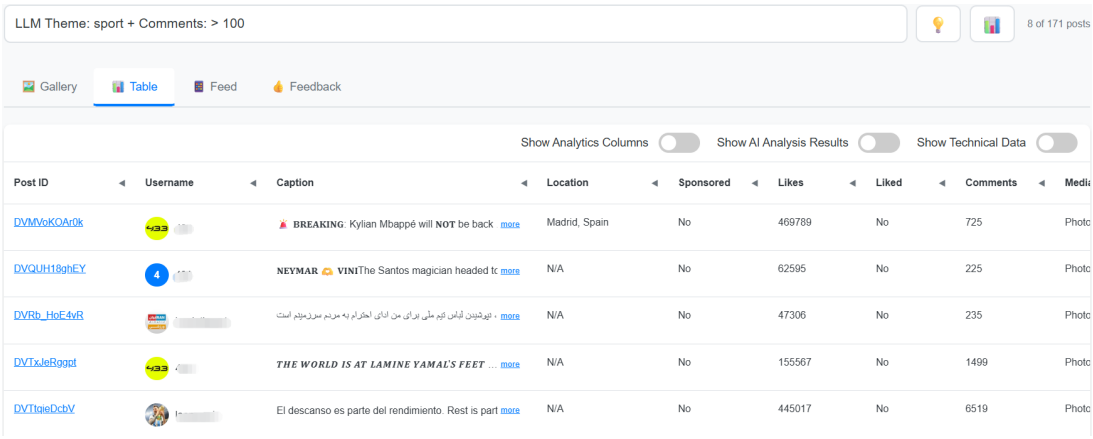

Figure 2: Database tab (table view). The extension stores posts locally and provides keyword search and filter controls (including nutrition label categories and post metadata) to retrieve subsets of the recorded feed for analysis.

4.2. Multimodal Description Generation

After initial storage, the system generates a textual description for the image or video (based on a video frame) of each post using the multimodal LLM. These descriptions are stored alongside the captured data as derived representations. Generating descriptions by default ensures that each post has a unified textual representation, facilitating consistent downstream classification and conversational reasoning.

4.3. Zero-Shot Category Classification

Classification is performed via structured zero-shot prompting. For each active nutrition label category, the prompt includes:

- Category name (required)
- Optional description
- Optional set of candidate values
- Whether the value set is open or closed

The LLM is guided to return valid JSON output conforming to a predefined schema. This design allows researchers to define new analytical categories dynamically without retraining the model. The resulting category assignments are appended to the corresponding post records in the local database.

Category definitions are managed through the extension settings interface, where researchers can specify names, optional descriptions, and candidate values (Figure 3).

For transparency at the item level, each post includes an inline overlay displaying the predicted category assignments (Figure 4).

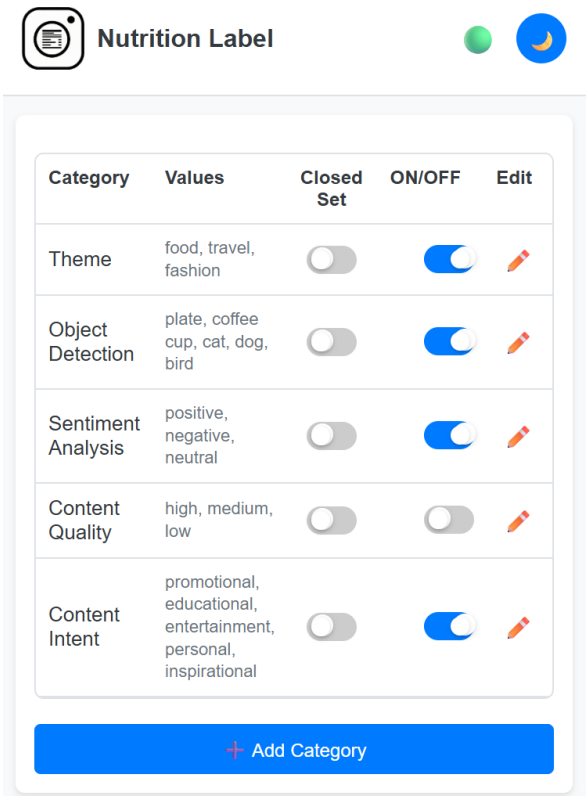

Figure 3: Extension settings. Researchers define nutrition label categories by specifying a name and optionally a description and candidate values. Categories are applied via zero-shot prompting without retraining.

4.4. Feed-Level Aggregation

Category assignments are aggregated across the recorded feed to generate a nutrition label summarizing exposure patterns. Aggregation can be computed either by post count or by cumulative

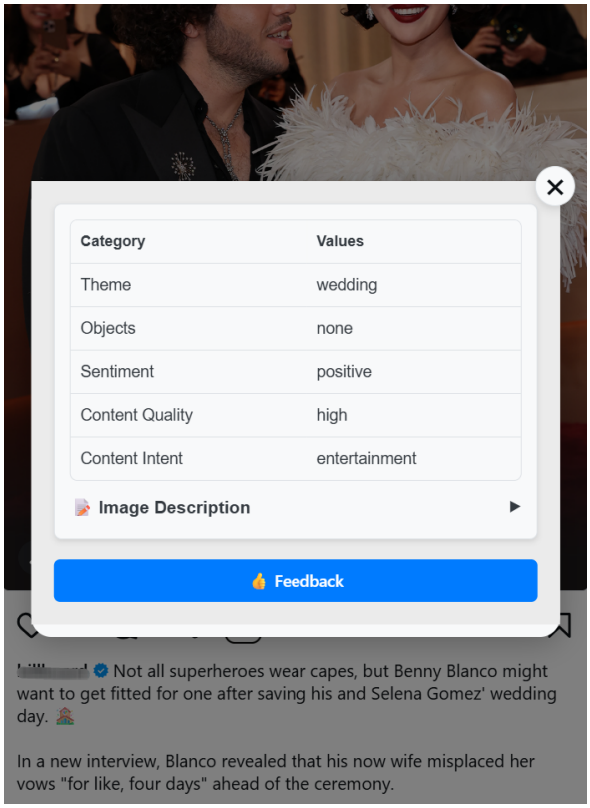

Figure 4: Post-level nutrition overlay. The LLM's category assignments for the selected post are displayed inline within the social media interface.

view time, thereby approximating exposure intensity under rapid scrolling conditions.

The aggregated results are presented in a feed-level nutrition label that supports drill-down interaction (see again Figure 1). Selecting a category value filters the local dataset and opens the corresponding subset in the database view.

4.5. Retrieval-Grounded Conversational Querying

The "Talk to your Feed" interface implements a two-step reasoning pipeline. Given a natural language question, the LLM first generates a structured query expressed in the system's custom search syntax. The extension executes this query locally against the IndexedDB dataset. Retrieved posts, including caption text, generated image/video descriptions, metadata, interaction signals, and category assignments, are then sent back to the LLM. The LLM produces an answer grounded explicitly in this retrieved content.

This separation between query generation and answer generation constrains reasoning to the recorded exposure dataset and reduces ungrounded outputs. The conversational interface exposes this retrieval-grounded reasoning process directly to researchers (Figure 5).

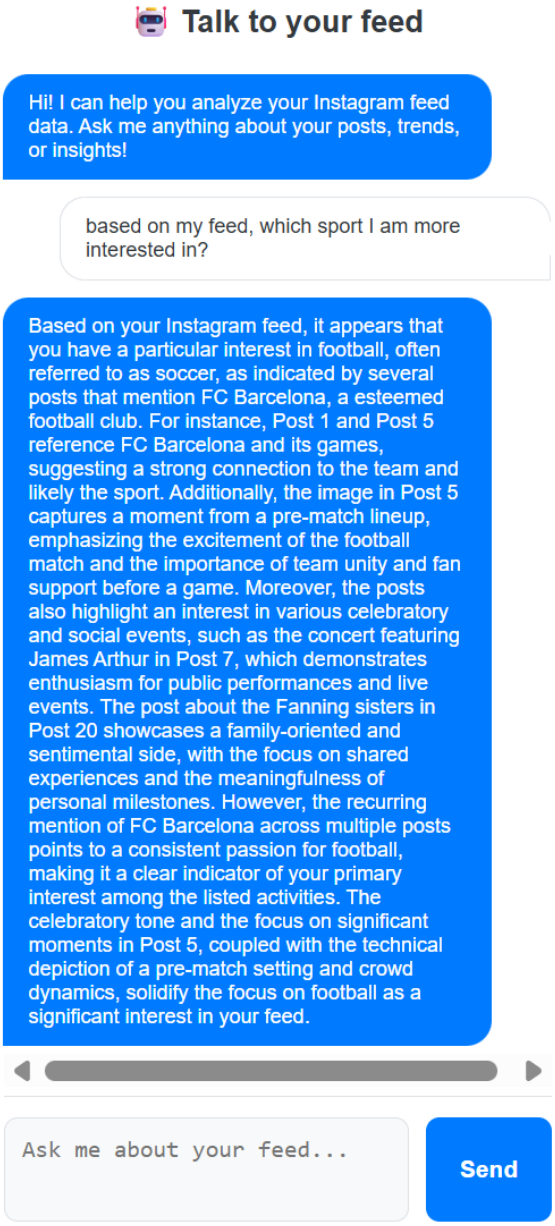

Figure 5: Retrieval-grounded conversational interface. The LLM first generates a structured query over the local database. Retrieved posts are then used to generate a grounded response.

4.6. Human-in-the-Loop Validation

To assess classification reliability, the system provides a review interface allowing researchers to navigate posts sequentially using keyboard controls and evaluate predicted category values. Validation feedback (correct/incorrect) is stored as additional fields associated with each post record.

Annotations can be exported to compute classification accuracy and inter-annotator agreement, enabling mixed-method research designs that combine automated coding with human validation (Figure 6).

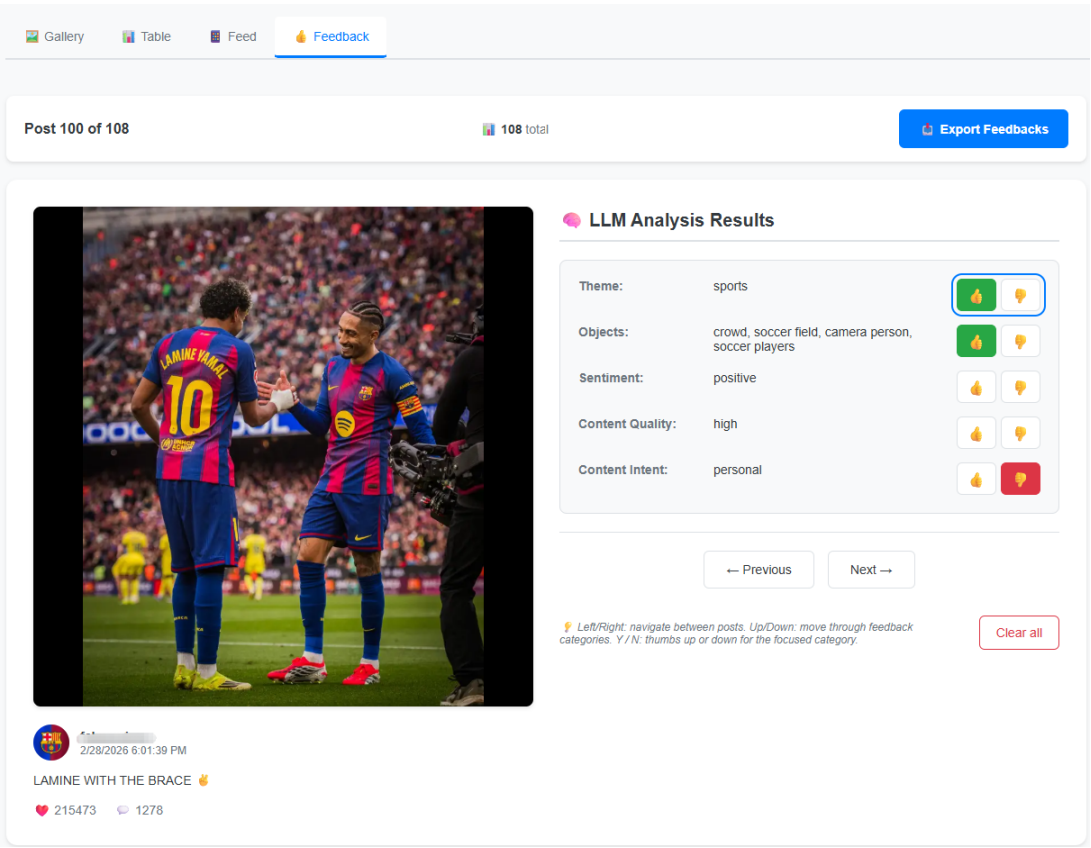

Figure 6: Review interface for validating LLM classifications. Researchers assess predicted category assignments and export annotations to compute accuracy and inter-annotator agreement.

4.7. Data Portability and Reproducibility

Exposure datasets can be exported and imported using a custom JSON format. Exported data include captured fields (metadata, interaction signals, view time), derived representations (image/video descriptions and category assignments), and human validation annotations. This functionality enables collaborative workflows in which multiple researchers analyze identical exposure datasets independently.

By combining local storage, open-weight LLM deployment, structured prompting, retrieval-grounded reasoning, and exportable datasets, the architecture supports flexible yet methodologically transparent exposure analysis.

5. Methodological, Epistemic, and Ethical Considerations

Using multimodal LLMs as analytical instruments for feed-level exposure analysis introduces methodological, epistemic, and ethical challenges that require careful consideration.

Zero-Shot Operationalization. Zero-shot classification enables researchers to define analytical categories dynamically without retraining. While this flexibility lowers barriers for exploratory research, it introduces prompt sensitivity and model-version dependency. Category definitions should therefore be understood as operational constructs rather than objective ground truth labels. Reproducibility depends on documenting prompts, model versions, and hardware configurations.

Multimodal Interpretation and Derived Representations. The system generates textual descriptions of images and video frames to create a unified representation for reasoning and retrieval. While this facilitates consistent analysis, multimodal LLMs may hallucinate details or overgeneralize visual cues. Researchers should treat generated descriptions and classifications as interpretive artifacts rather than direct observations.

Operationalizing Exposure Through View Time. Weighting aggregation by view time approximates attention allocation in scrolling environments. However, visibility duration does not fully capture cognitive engagement, affective response, or background exposure. The metric should therefore be interpreted as a proxy for temporal prominence rather than a direct measure of psychological impact.

Normativity and Bias. Certain analytical categories, such as "Content Quality" or "Intent", embed normative assumptions. LLM outputs may reflect biases present in training data, including cultural or linguistic biases. Human validation and inter-annotator agreement mechanisms are therefore essential components of responsible use.

Privacy and Data Governance. All exposure data are stored locally within the browser using IndexedDB. The backend LLM processes requests transiently and does not persist images or text. Images and video frames are transmitted for inference but not retained. The architecture is designed for deployment on researcher-controlled hardware rather than centralized service provision. Nonetheless, systematic feed logging raises broader ethical questions regarding surveillance normalization, contextual integrity, and informed consent in studies involving social media content.

Taken together, these considerations underscore that the system should be understood as a methodological instrument whose outputs require interpretive caution, documentation, and, where appropriate, human oversight.

6. Evaluation

We are currently implementing a two-part evaluation protocol addressing (1) the reliability of zero-shot LLM classifications and (2) the behavioral and perceptual effects of the nutrition label as a transparency intervention.

6.1. Classification Reliability and Agreement

To evaluate classification performance, we export recorded exposure datasets in the system's custom JSON format and distribute them to independent reviewers. Reviewers import the dataset into their local instance of the extension and use the built-in validation interface to annotate whether predicted category assignments are correct.

The exported annotation files are returned and aggregated using a dedicated analysis script. For each category, we compute:

- Inter-annotator agreement (e.g., Cohen's or Fleiss' κ)
- Majority-based ground truth labels
- LLM performance metrics with respect to majority assessments (e.g., accuracy, precision, recall)

This protocol enables systematic evaluation of zero-shot category definitions and provides insight into which types of categories (e.g., descriptive vs. normative) yield higher agreement and classification reliability.

6.2. User Perception and Intervention Effects

Beyond classification reliability, we plan to investigate the effects of the nutrition label as a transparency intervention. Participants will install the extension and use it with their own feed. The study will explore:

- How exposure summaries influence scrolling and engagement behavior
- Whether awareness of aggregated exposure alters perceptions of feed composition
- How accurately users can predict the distribution of their feed across categories prior to viewing the nutrition label

In particular, we are interested in comparing users' predicted category distributions with measured distributions derived from recorded exposure data. This comparison may reveal systematic misperceptions regarding feed composition and algorithmic influence.

Together, these evaluation components address both methodological validity of LLM-based classification and the societal implications of transparency interventions.

7. Discussion and Outlook

We presented a research-oriented framework for operationalizing feed-level exposure analysis using multimodal large language models. By combining view-time logging, structured zero-shot classification, retrieval-grounded conversational querying, and human validation mechanisms, the system provides an integrated environment for studying algorithmically curated media streams.

The framework contributes methodologically in three ways. First, it introduces time-weighted exposure as an operational construct for analyzing scrolling-based media consumption. Second, it demonstrates how open-weight multimodal LLMs can be embedded into reproducible research workflows through structured prompting, exportable datasets, and agreement-based evaluation. Third, it positions transparency interfaces not only as user-facing tools but as experimental instruments for studying the behavioral and perceptual effects of algorithmic disclosure.

The planned evaluation protocol will assess both classification reliability and the epistemic status

of zero-shot category definitions across descriptive and normative constructs. By combining inter-annotator agreement with majority-based performance metrics, we aim to clarify under which conditions LLM-based feed analysis yields stable and interpretable results.

Beyond methodological validation, the system enables empirical investigation of transparency interventions. Comparing users' predicted feed composition with measured exposure distributions may reveal systematic misperceptions about algorithmic influence and content balance. Such findings could inform debates in platform governance, digital literacy, and algorithmic accountability.

Future work includes extending support to additional platforms, incorporating richer video analysis, exploring multilingual category definitions, and formalizing reproducibility standards for LLM-based content analysis pipelines. As multimodal models continue to evolve, maintaining transparency in prompt design, model versioning, and hardware configuration will remain essential for ensuring methodological rigor.

By bridging exposure logging, multimodal reasoning, and human validation, the proposed framework contributes to emerging methodological toolkits for studying algorithmically curated information environments in the Social Sciences and Humanities.

References

Alibaba Cloud. 2024. Qwen2.5-vl-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct.

Luca Belli and Marlena Wisniak. 2023. What's in an algorithm? empowering users through nutrition labels for social media recommender systems. *Knight First Amendment Institute at Columbia University*.

European Union. 2022a. Article 27: Recommender system transparency, digital services act (eu) 2022/2065. Obliges online platforms to explain the main parameters of recommender systems and provide users with meaningful options to influence content recommendations.

European Union. 2022b. Article 40: Data access and scrutiny, digital services act (eu) 2022/2065. Providers of Very Large Online Platforms must provide access to data for vetted researchers to study systemic risks.

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2017. An information nutritional label for online documents. *SIGIR Forum*, 51(3):46–66.

Tim Gollub, Martin Potthast, and Benno Stein. 2018. Shaping the Information Nutrition Label. In *2nd International Workshop on Recent Trends in News Information Retrieval (NewsIR 2018) at ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 9–11.

Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Şahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker. 2018. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 28–33, Brussels, Belgium. Association for Computational Linguistics.

Stijn Peeters. 2025. Zeeschuimer.

John Willinsky and Daniel Pimentel. 2024. The publication facts label: A public and professional guide for research articles. *Learned Publishing*, 37(2):139–146.