Fact-Check Your Information (FYI): A Design Probe to Understand How People Actually Fact-Check Data-Driven Articles

Nguyen-Truong Thinh*, Yuxuan Du*, Phongsakon Mark Konrad, Arpit Narechania

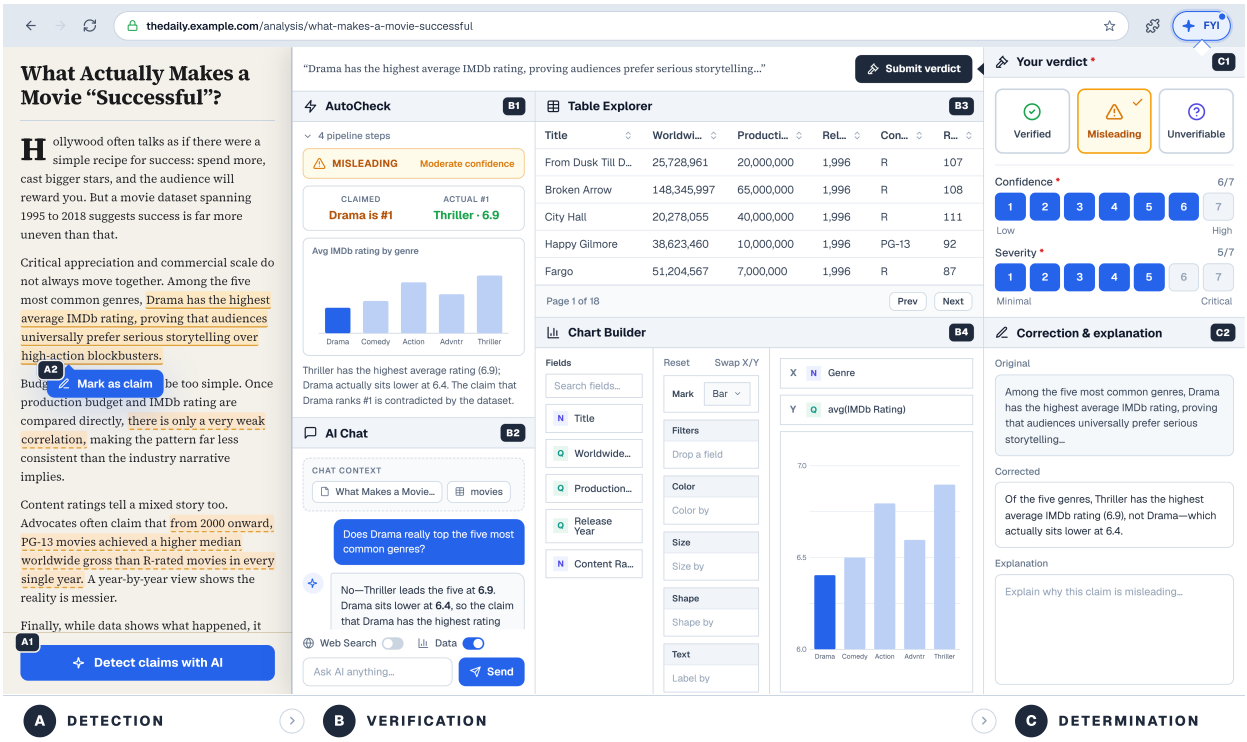

Fig. 1. The interface of FYI, a browser extension that enables in-situ data claim fact-checking directly alongside the article under review. The unified side panel supports a complete mixed-initiative pipeline: **(A) Detection** via (A1) AI detection or (A2) manual identification; **(B) Verification** through four complementary modalities: (B1) Auto Check, (B2) AI Chat, (B3) Table Explorer, and (B4) Chart Builder; and **(C) Determination** via (C1) user-authored verdicts and (C2) corrections and explanations.

Abstract—Data-driven journalism and policy reports frequently rely on statements grounded in statistical evidence, referred to as *data claims*. Verifying such a claim requires connecting it to the underlying structured dataset. However, existing systems typically isolate automated fact-checking from manual data exploration, leaving it unclear how readers coordinate AI assistance with manual inspection of the evidence in practice. We present FYI, a browser extension that embeds fact-checking in the reading environment, and use it as a design probe to study how people detect, verify, and determine the validity of data claims against the underlying dataset. FYI provides four complementary tools spanning the spectrum from full automation to manual data exploration. In an exploratory study ($N = 22$), participants used FYI to fact-check claims in a data-driven article. We find that participants adopted three distinct workflow archetypes—AI-first with manual confirmation, manual-first with AI supplement, and parallel co-review—with visualization serving as the primary mechanism for auditing AI conclusions. Trust in AI shifted dynamically, growing when multiple tools converged and eroding when AI outputs were inconsistent. These findings suggest that fact-checking systems should treat AI as a starting point that human verification complements rather than a definitive authority, elevate visualization as a core verification capability, and support flexible, user-driven workflows. We release FYI as open-source software for further research at https://github.com/DataVisards/FYI.

Index Terms—Data claims, fact-checking, human-AI interaction, trust calibration, visualization, large language models, design probe.

1 Introduction

Consider a reader encountering the following statistic in a public health report: *"The risk of death involving COVID-19 was consistently lower for people who had received two vaccinations compared to one or no vaccination"* [43]. This is a data claim, a statement grounded in quantitative evidence from structured datasets [18], routinely found in data journalism and policy reporting [63]. Unlike textual fact-checking, which relies on finding corroborating sources, verifying a data claim requires specialized tools to aggregate, compare, and interpret structured

- *Nguyen-Truong is with Holistics Software and is an incoming M.Phil. student at The Hong Kong University of Science and Technology. E-mail: truongthinh.nguyen03@gmail.com*
- *Yuxuan is an Independent Contributor. E-mail: yuxuan.du.sherry@gmail.com*
- *Mark is with the University of Southern Denmark. E-mail: phkon23@student.sdu.dk*
- *Arpit is with The Hong Kong University of Science and Technology. E-mail: arpit@ust.hk*

* = *Authors contributed equally to this work.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

data. This illusion of rigor can make misrepresented statistics appear credible, especially when readers lack the time or dedicated tools to inspect the underlying data.

Systematically addressing this risk requires supporting readers through the established fact-checking pipeline: claim *detection* (identifying check-worthy statements), *verification* (evaluating evidence against the dataset), and *determination* (synthesizing that evidence into a judgment) [20, 22]. Each stage combines rigorous data processing with subjective reasoning, so navigating the pipeline demands both computational assistance and manual inspection of the data.

Existing systems span the human-AI spectrum but remain fragmented. Fully automated pipelines are unreliable on precise numerical claims [46, 62], interactive analytics platforms impose a high cognitive burden [33], crowdsourced approaches lack analytical depth [37], and lightweight browser extensions provide only shallow credibility signals [6, 27] (see Sec. 2 for a detailed review). No unified environment lets users flexibly move between AI-generated outputs and direct data examination across the full fact-checking pipeline.

Mixed-initiative systems introduce a fundamental tension. As agency shifts towards AI, users may accept outputs without examining the data (automation bias [19, 34, 36]), while full human control risks cognitive overload [2, 9]. Users must actively decide how much agency to delegate at each step, constituting a form of trust calibration that can shift across phases and claims, and in response to prior AI reliability. How users calibrate this trust when both modalities are available remains poorly understood in data claim fact-checking.

These tensions point to an empirical gap in understanding how people actually behave when fact-checking data claims with AI-assisted and manual modalities simultaneously available. Studying these behaviors requires an integrated and instrumented artifact that makes cross-modal workflows observable.

To address this gap, we pursue the following research questions:

- **RQ1** (Detection): How do users detect check-worthy data claims?
- **RQ2** (Verification): How do users verify data claims?
- **RQ3** (Determination): How do users reach and communicate their fact-checking verdict?
- **RQ4** (Human-AI): How do users balance trust in AI against staying in control throughout the fact-checking process?

To answer these questions, we designed Fact-check Your Information (FYI), a browser extension that embeds the fact-checking workflow directly within the reading environment. FYI is scoped to dataset-grounded data-claim fact-checking for data-driven articles, whose authors are expected to disclose the underlying dataset as a downloadable file or linked source. It is intended for laypersons with basic data and visualization literacy who consume such articles online and can inspect tables or interpret charts, but who do not necessarily have access to professional fact-checking infrastructure. FYI supports the detection–verification–determination workflow: users can manually highlight claims or use AI detection; verify claims through a fully automated pipeline (Auto Check), a conversational AI agent (AI Chat), an interactive table (Table Explorer), and a visualization builder (Chart Builder); and submit verdicts with corrections where appropriate.

Using FYI as a design probe, we conducted an exploratory user study (N=22) in which participants freely combined these tools to fact-check data claims in a realistic data-driven article. By logging fine-grained interaction sequences, we turn the reading session itself into an observational window into users' fact-checking process. Our primary contribution is an exploratory design-probe study of mixed-initiative data-claim fact-checking. Specifically, we contribute:

1. Empirical characterization of how laypersons compose fact-checking strategies, move across AI-assisted and manual tools, and calibrate trust across modalities.
2. Design implications for future mixed-initiative fact-checking systems.
3. FYI, an open-source prototype contributed to the community as a testbed for future data fact-checking research (https://github.com/DataVisards/FYI).

2 Related Work

2.1 From Text Claims to Data Claims

The foundation of automated fact-checking was established primarily within the NLP community, focusing almost exclusively on unstructured text. The standard process is organized into a three-stage pipeline [22]: *detection* of check-worthy statements, *verification* against retrieved textual evidence, and *determination* of a final veracity label. Early benchmarks such as FEVER [54], LIAR [57], and SciFact [56] operated strictly within this framework. For text claims, detection typically involves identifying rhetorical markers or subjective versus objective framing to determine whether a statement is worthy of investigation [26]. Subsequently, the verification phase is treated fundamentally as a natural language inference (NLI) task, where the core challenge is semantic matching to determine whether a retrieved piece of text entails or refutes the detected claim [42].

However, *data claims*, statements grounded in quantitative evidence drawn from structured datasets, require fundamentally different approaches. Detecting a data claim requires identifying numerical assertions, statistical summaries, or comparative trends embedded within prose, and recognizing that these statements implicitly refer to an underlying dataset [55]. Once a claim is detected, verifying a data claim demands a combination of linguistic interpretation and precise analytical operations such as comparing, counting, and aggregating over table rows or visual marks [12] to derive the answer from the raw data. Thus, fact-checking of data claims has inherently higher computational and cognitive complexity than pure text processing.

2.2 Paradigms of Fact-Checking Systems

In this section, we review systems that support data-claim fact-checking or closely related evidence-checking tasks involving structured data, tables, or visualizations. To address the complexity of data claims, existing system designs have largely diverged into paradigms along the human-AI spectrum.

One is the fully automated paradigm, which aims to minimize user effort by delegating complex reasoning entirely to AI. For the detection phase, models like ClaimBuster [23] and subsequent LLM-based extractors [38] attempt to flag check-worthy statements automatically without user intervention. For verification, the field has shifted toward neural-symbolic decomposition strategies. Binder [14] parses claims into executable SQL or Python expressions, while DATER [61] decomposes tables into focused sub-tables for LLM reasoning, and Chain-of-Table [58] applies iterative table operations to evolve evidence through a reasoning chain. Recent end-to-end architectures, such as Aletheia [18], attempt to automate the entire lifecycle from semantic parsing to the generation of interactive data evidence representations, and multi-agent systems like Thucy [53] deploy specialized LLM agents to verify claims across relational databases with executable SQL evidence (achieving 94.3% on TabFact). Despite their computational power, these fully automated systems primarily operate without a human-in-the-loop mechanism. When automated reasoning produces an incorrect or hallucinated result, users are unable to contest the output [1]. Although Aletheia provides interactive widgets for overriding AI-inferred filters, its user agency remains limited to error correction rather than open-ended exploration.

In contrast, the interactive paradigm prioritizes direct data inspection. Some systems target verification problems adjacent to article-level fact-checking and demonstrate the role of visual evidence in checking data-related statements. For example, DataTales [52] supports LLM-assisted authoring of data-driven articles with integrated visual evidence, while EmphasisChecker [31] helps users construct analytical charts to check text-chart consistency. They both show how visual evidence can support inspection of data-related claims, even though their primary tasks differ. Meanwhile, visual overviews and raw tabular inspection serve complementary roles. Kim et al. [32] suggested that rigorous verification often requires users to drill down into specific table cells to resolve granular ambiguities. However, this paradigm assumes that the user has already detected the claim and formulated a verification strategy. Further, relying purely on interactive tools to navigate raw data and construct visual evidence imposes relatively high analytical burden and expertise requirements [9].

Table 1. Comparison of data fact-checking systems across four dimensions: Claim Detection (M1), Claim Verification (M2), Claim Determination (M3), and Human Agency (M4). The table compares systems relevant to data-claim fact-checking and adjacent evidence-checking tasks. Columns indicate which parts of the detection–verification–determination workflow each system supports, without implying that all systems are end-to-end fact-checking systems. Systems are grouped by interaction model—fully-automated NLP systems (top) and human-in-the-loop systems (bottom)—to reflect design intent instead of comparing their capability.

Y = Supported **P** = Partially Supported **N** = Not Supported

		Claim Detection		**Claim Verification**				**Claim Determination**		**Human Agency**	
		M1		*M2*				*M3*		*M4*	
System	**Venue**	*Automated Detection*	*Manual Selection*	*Automated Verification*	*NL Dialog*	*Tabular Inspection*	*Visual Analytics*	*Human Verdict*	*In-situ Annotation*	*AI Override*	*Multi-Tool Orchestration*
Fully-Automated NLP Systems											
ClaimBuster [23]	KDD '17	**Y**	**P**	**Y**	**N**	**N**	**N**	**P**	**N**	**N**	**N**
TabFact [12]	ICLR '20	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**N**	**N**
PASTA [21]	EMNLP '22	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**N**	**N**
DATER [61]	SIGIR '23	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**N**	**N**
Binder [14]	ICLR '23	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**P**	**N**
Chain-of-Table [58]	ICLR '24	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**N**	**N**
RePanda [11]	ACL '25	**N**	**N**	**Y**	**N**	**Y**	**N**	**N**	**N**	**N**	**N**
Human-in-the-Loop Systems											
Believe it or not [41]	UIST '18	**N**	**Y**	**Y**	**N**	**N**	**N**	**Y**	**N**	**Y**	**N**
Scrutinizer [30]	VLDB '20	**Y**	**P**	**P**	**P**	**Y**	**N**	**Y**	**N**	**Y**	**N**
StatCheck [4]	CIKM '22	**Y**	**P**	**Y**	**N**	**Y**	**N**	**N**	**N**	**P**	**N**
CrossData [13]	CHI '22	**Y**	**Y**	**P**	**N**	**Y**	**Y**	**N**	**Y**	**Y**	**Y**
DataTales [52]	VIS '23	**N**	**Y**	**P**	**N**	**P**	**Y**	**N**	**N**	**Y**	**Y**
DataDive [33]	IUI '24	**N**	**Y**	**P**	**P**	**Y**	**Y**	**N**	**N**	**Y**	**Y**
EmphasisChecker [31]	TVCG '24	**Y**	**Y**	**Y**	**N**	**Y**	**Y**	**N**	**Y**	**Y**	**Y**
Aletheia [18]	UIST '24	**Y**	**Y**	**Y**	**N**	**Y**	**Y**	**Y**	**Y**	**P**	**Y**
MisVisFix [15]	TVCG '25	**Y**	**Y**	**Y**	**Y**	**N**	**Y**	**P**	**Y**	**Y**	**Y**
T-REX [24]	ECML '25	**N**	**Y**	**Y**	**N**	**Y**	**Y**	**N**	**Y**	**N**	**Y**
FYI (ours)	—	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

Recently, mixed-initiative systems have sought to bridge the gap. At the organizational scale, Scrutinizer [30] demonstrated a structured mixed-initiative approach in which the system proposes SQL query fragments that domain experts validate or correct, with classifiers improving through active learning from accumulated human feedback, reducing verification time by over 50%. For individual users, StatCheck [4] prioritizes user agency by retrieving relevant statistical databases for journalists, leaving the final verification to the human. Similarly, real-time streaming interfaces like T-REX [24] highlight relevant table cells alongside LLM reasoning to make automated verification more transparent. Despite these advancements, the current systems for data claim fact-checking remain highly fragmented. Specifically, they are either fully automated, which strips user control, or highly manual, which overwhelms cognitive capacity. We currently lack unified environments that support a flexible workflow, where users can seamlessly transition between automated and human-in-the-loop fact-checking.

2.3 Human-AI Sensemaking and Trust Calibration

Bridging the gap between fully automated pipelines and high-effort manual analytics requires framing data fact-checking as a complex sensemaking process. Foundational models of sensemaking describe data analysis as an iterative cycle of information search, hypothesis establishment, and evidence synthesis [47]. In the data-claim fact-checking context, this loop spans both the detection and verification phases. It involves mapping claims to structured evidence and iteratively inspecting tables or visualizations to test interpretations.

However, introducing AI into the fact-checking process fundamentally alters user agency. A key challenge in AI-assisted analysis is that users over-rely on automated outputs instead of carefully seeking and checking information themselves [19]. In fact-checking scenarios, users might passively accept the AI verdict or an auto-generated chart without verifying its provenance, even when the underlying reasoning is flawed [10]. Conversely, if an AI fails to detect a claim or cannot explain its verification logic through transparent evidence, users may lose confidence and exhibit algorithmic aversion [28], entirely rejecting valid automated assistance and reverting to a fully manual check.

Therefore, effective mixed-initiative systems must be designed to support trust calibration, the process by which users align their reliance on the AI with the system's actual reliability [8]. This requires interactive override capabilities that allow users to recover from AI errors and re-exert agency [41]. Critically, the level of reliance may differ across the detection and verification phases [44]. For instance, AI claim detection may be accurate yet incomplete, while AI verdicts may be fluent but numerically incorrect. Users, therefore, need to calibrate trust based on the detailed scenarios. Furthermore, the benefits of explainability in automated fact-checking remain contested; Lim and Perrault [35] found that only one out of five tested XAI modalities marginally improved user performance, while graphical explanations may paradoxically anchor users to the AI's framing rather than encouraging independent verification. A recent review of human-AI decision support [50] further warns of a "fluency trap" where conversational AI interfaces inflate perceived understanding and trust without reliably improving decision quality, underscoring the need to evaluate whether multi-modal verification tools genuinely improve judgment or merely

shift reliance patterns.

Although existing research has studied the fact-checking workflows and tool use of professionals [29], empirical studies on laypeople remain exceedingly scarce. Specifically, because users rarely have access to a fully integrated toolset, there is a profound deficit in our understanding of how they actually behave. We lack empirical evidence on how users orchestrate strategies, divide agency, and calibrate trust when automated and interactive tools are simultaneously available. This critical empirical gap motivates the deployment of FYI as an in-situ design probe to observe and characterize these multi-modal behavioral dynamics and sensemaking strategies in real-world reading environments.

2.4 Design Gaps and Positioning

As summarized in Tab. 1, the comparison is structured along the detection–verification–determination pipeline extended with a human agency dimension, which is central to mixed-initiative fact-checking.

The upper part of the table shows that NLP systems typically provide limited support for human agency and determination except for Binder's partial AI override or ClaimBuster's limited manual inputs. Furthermore, they mostly treat claim detection and verification as an automated preprocessing step rather than a user-driven behavior. As a result, interactive operations, user-authored verdicts, and multi-tool orchestration often fall outside their design scope.

The lower part of the table shows a complementary pattern. Human-in-the-loop systems provide stronger support for user agency. For example, CrossData and EmphasisChecker successfully empower users with multi-tool orchestration, AI overrides, and in-situ annotation, while Aletheia provides multi-tool evidence and partial AI override through interactive filter correction. However, this agency often shifts more analytical work to users. For instance, DataTales and DataDive offer only partial support for AI-assisted fact-checking, so users still need to formulate analyses and inspect outputs. More critically, across both the NLP and human-in-the-loop paradigms, natural language dialog remains largely absent. While Scrutinizer offers structured query-based interaction and DataDive allows optional free-form text input, only MisVisFix [15] provides a multi-turn chat interface.

This comparison motivates FYI as an integrated design-probe prototype for our study. By combining natural language dialog (AI Chat), automated verification (Auto Check), tabular inspection (Table Explorer), and visual analytics (Chart Builder), FYI brings together modalities that are typically separated across prior systems, making it possible to empirically observe how users orchestrate these previously fragmented modalities to fact-check data claims.

3 FYI: A Design Probe

3.1 Design Goals

To observe how readers fact-check data claims when automated and manual tools are simultaneously available, FYI needed to keep the workflow within the reading context, expose multiple available verification modalities, and preserve user agency while recording the resulting process. Drawing on prior work in automated fact-checking [18, 22], interactive data exploration [25, 33], and analytical provenance [5, 39], we translate these requirements into three design goals.

DG_1: In-situ integration. Existing fact-checking tools often require users to leave their reading environment [51], incurring context-switching costs [49]. FYI should embed the full investigation workspace in the browser as a side panel, allowing users to verify claims while reading the article.

DG_2: Multi-modal verification. Data claim verification is rarely reducible to a single technique [18]: AI pipelines accelerate evidence retrieval but may hallucinate [46, 62], tabular inspection reveals precise values, visualizations expose distributional patterns, and web search provides external corroboration [3]. FYI should offer multiple complementary tools so that users can triangulate evidence across modalities.

DG_3: Human agency and provenance. Recent work highlights the risk of over-reliance on AI outputs during fact-checking [16, 34, 36]. FYI should keep the user as the primary decision-maker: AI provides evidence and suggestions, but verdicts are always user-submitted. The system should also log all interactions to enable post-hoc analysis of verification workflows and trust calibration.

Together, these goals shape FYI as a design probe that supports in-situ fact-checking, multi-modal verification, and user-authored determination while making users' interaction behaviors observable.

3.2 Prototype Overview

FYI is a browser extension that hosts its fact-checking workspace in a side panel beside the article under review. Targeting data-driven articles, FYI assumes the underlying dataset is available, either as a downloadable file or a linked source—a transparency norm in data journalism [63]. The reader uploads this dataset as grounding data, shared between the AI's analysis and the reader's own inspection. The reader can then confirm each AI suggestion against this data, viewing it as tables and charts, instead of accepting it on the AI's word.

The user workflow consists of three stages:

1. **Detection.** The user activates FYI on an article page, uploads CSV datasets as grounding data, and initiates claim detection either via the AI pipeline or by manually highlighting passages. Detected claims are highlighted in the article (Sec. 3.3).
2. **Verification.** The user selects a claim and investigates it using four complementary tools (Sec. 3.4), triangulating evidence across AI and manual modalities.
3. **Determination.** The user submits a judgment (verified, misleading, or unverifiable) with confidence and severity ratings, and optionally provides a textual correction for misleading claims (Sec. 3.5).

All AI-powered components (claim detection, Auto Check, and AI Chat) use GPT-4.1 [45], except web search, which uses Perplexity Sonar; all prompts are included in the supplemental materials.

3.3 Claim Detection

Claim detection supports both automated and manual pathways, spanning the human-AI spectrum. AI-powered detection offers speed and coverage, while manual curation preserves user agency and captures claims that automated methods may miss—such as implicit comparisons or domain-specific assertions. Providing both enables observation of how users divide detection labor between themselves and AI.

AI-powered detection. Data claims are often embedded in flowing prose in ways that readers may not immediately recognize as verifiable [18], so automated detection lowers the entry barrier and supports broad coverage. An LLM classifies whether each sentence is a potential data claim, which is defined as a natural-language statement whose veracity depends on a specific dataset. Detected claims are highlighted directly in the article and listed in the side panel for investigation.

Manual claim detection. AI detection can miss claims that depend on context, domain knowledge, or subjective judgment about what is worth checking, so users need to supplement and override. Users can select any text passage on the article page, triggering a floating toolbar with a "Mark as claim" option (DG_3). User-added claims always take priority: when a manual claim overlaps an AI-detected claim in the same text region, the AI claim is automatically suppressed. Users can also dismiss AI-detected claims they consider irrelevant, but AI cannot modify or remove user-added claims. This asymmetry ensures that human judgment always supersedes automated detection, while dismissal patterns provide a behavioral signal of disagreement with AI.

3.4 Verification Tools

The investigation workspace provides four complementary tools (DG_2), each accessible as a tab within the side panel. Users can switch between tools freely, and the system tracks tool-usage sequences and dwell time per claim (DG_3).

The four tools span a spectrum from fully automated (Auto Check) through user-directed AI (AI Chat) to fully manual (Table Explorer, Chart Builder), enabling observation of how users compose verification strategies.

3.4.1 Auto Check

Auto Check provides a zero-effort baseline, executing a four-step streaming pipeline: (1) resolving references to make the claim self-contained, (2) generating and executing verification code (Binder-inspired [14]), (3) producing an interactive Vega-Lite evidence chart, and (4) outputting a verdict badge with confidence, a claimed-vs-actual comparison, and natural-language reasoning. Steps are available in a collapsible view, and a "Follow up in AI Chat" button seeds deeper exploration in AI Chat.

3.4.2 AI Chat

Because Auto Check produces a single predetermined analysis, users often need to ask follow-up questions, explore alternative interpretations, or seek external corroboration—tasks that require flexible, user-directed AI assistance. AI Chat provides a multi-turn conversational interface for this open-ended verification. The system prompt includes the article content, the selected claim, and metadata for any uploaded datasets. Users can toggle two capabilities to control the scope of AI involvement:

- **Web Search** enables the model to query external sources for corroboration beyond the dataset; results are rendered as numbered inline citations with clickable source links.
- **Data Analysis** grants the model access to Python code execution via client-side Pyodide, with datasets pre-loaded as pandas DataFrames. This allows multi-step statistical analysis, aggregation, and computation within the conversation.

3.4.3 Table Explorer

AI-generated summaries can obscure the underlying evidence, so Table Explorer provides direct, unmediated access to the raw dataset, enabling users to inspect specific values rather than relying on AI interpretations. Users can sort by any column, apply categorical filters (checkbox multi-select) or quantitative filters (range sliders), and paginate through large datasets.

3.4.4 Chart Builder

Both Table Explorer and Chart Builder operate without AI, but Chart Builder demands the highest analytical effort: users must decide what to plot, how to encode it, and how to interpret the result. Following the shelf-based encoding paradigm of Voyager [59, 60], users build Vega-Lite charts by dragging dataset fields onto encoding shelves (x, y, color, size), selecting mark types (bar, line, point, area), and applying aggregation functions (e.g., average, median, count) and filters. Multiple chart tabs allow parallel exploration of different hypotheses for the same claim.

3.5 Verdict and Correction

The verdict form serves as both a decision endpoint and a data collection instrument. By capturing not only the judgment but also the reasoning and tool attribution, it enables post-hoc analysis of how different evidence sources map to different decision outcomes. Users submit a judgment (*verified*, *misleading*, or *unverifiable*), a 7-point confidence rating, and a ranked list of which tools were helpful during the investigation (ordered from most to least helpful). For misleading and unverifiable claims, users additionally provide a severity rating (1–7) and an explanation; misleading claims also support an inline correction where users edit the original claim text to reflect the accurate version.

3.6 Interaction Logging

To enable the design probe analysis ($\mathbf{DG_3}$), FYI records every user action as a timestamped event. The event schema covers 25 event types spanning the full pipeline: claim detection and dismissal, tool invocations and toggle changes, data exploration actions (sorting, filtering, chart encoding changes), chat exchanges (messages, responses, tool calls), and verdict submissions with all associated metadata. At session completion, the system exports a JSON artifact containing the full event log, all claims and verdicts, and session summary metrics. This granular provenance data enables reconstruction of each participant's complete verification workflow—which tools were used for which claims, in what order, and with what outcomes—supporting the behavioral analyses reported in Sec. 5.

4 Study

Using FYI as an instrumented design probe, we conducted an exploratory study to examine how participants fact-check data claims when multiple AI and manual verification tools are simultaneously available. The complete study materials, including the article, dataset, and post-study interview questions, are provided in the supplementary materials.

4.1 Participants

We recruited 24 participants through university mailing lists and research group networks. Two were excluded due to data loss (one missing session logs, one missing session recording), yielding a final sample of 22 (14 male, 8 female; 12 aged 18–24, 10 aged 25–34). Participants held Bachelor's (10), Master's (5), or PhD (7) degrees from disciplinary backgrounds including computer science, HCI, data science, and engineering. Self-reported prior experience with data visualization was high ($M = 5.68/7$, $SD = 1.09$), and daily use of generative AI tools was near the ceiling ($M = 6.18/7$, $SD = 1.14$). Participants reported moderate fact-checking experience ($M = 4.18/7$, $SD = 1.44$), while familiarity with the article's topic was mixed ($M = 3.68/7$, $SD = 1.94$). The study was conducted entirely at Holistics Software, where the lead author was employed at the time and where research of this kind did not require IRB approval. Informed consent was obtained from each participant before the study. Each participant was compensated USD 10 for their time.

4.2 Materials

We adopted a data-first approach, selecting the dataset before drafting the article to ensure experimental control.

Dataset. A movie industry dataset containing 1,724 films (1995–2018) with eight attributes (title, worldwide gross, production budget, release year, content rating, running time, genre, and IMDb rating) was constructed from publicly available sources [40]. The movie domain was chosen for broad accessibility without requiring specialized knowledge, while the attributes support diverse verification operations (filtering, aggregation, cross-group comparison, and correlation).

Article. Grounded in the dataset, we authored a movie industry analysis article covering box-office performance, ratings, and genre patterns, designed to resemble typical data-driven journalism. We embedded six data claims that span the three verification outcomes (3 verified, 2 misleading, and 1 unverifiable). We further designed the claims so that no single tool would suffice to check them: some can be settled by a direct data operation such as filtering or charting a value, others require multi-step analysis such as comparing groups or computing a correlation, and others demand judgment beyond the data, such as recognizing a statistically true but rhetorically overreaching statement or a claim the dataset cannot settle. This variation was intended to require participants to combine multiple verification strategies rather than rely on any single tool or pathway.

4.3 Procedure

Each session lasted approximately 60 minutes and was conducted remotely via Zoom. The study followed three phases.

Phase I: Onboarding and warm-up (20 min). Participants received a brief introduction to data claims and a guided walkthrough of FYI's interface, followed by a warm-up task to reach a baseline level of proficiency with each tool before beginning the main task. They were also told that the study focused on understanding their verification process rather than evaluating their speed or correctness.

Phase II: Main task (30 min). Participants were presented with the movie article and its grounding dataset. They were first instructed to read the article and identify data claims they considered worth checking, either by running AI detection, manually highlighting passages, or both. Then they freely investigated those claims using any combination of FYI's four tools. We intentionally did not prescribe a fixed claim order or tool order, because the design-probe study aimed to observe how

Table 2. Per-participant summary (P03 and P10 excluded due to data loss; see Sec. 4). Claims = highlighted claims; Verdicts = submitted verdicts; Confidence = mean self-reported confidence (1–7); Duration = session length in minutes.

ID	Claims	Verdicts	Verified	Misleading	Confidence	Duration
P01	5	5	3	0	5.20	21.7
P02	15	11	6	2	6.45	26.3
P04	12	6	4	0	6.17	28.4
P05	6	6	3	2	7.00	24.8
P06	8	5	4	1	6.40	25.8
P07	5	5	4	1	6.80	33.3
P08	4	4	4	0	7.00	27.1
P09	9	7	5	0	6.43	30.9
P11	6	6	6	0	6.67	22.0
P12	4	4	4	0	7.00	12.5
P13	5	5	3	2	6.80	23.8
P14	6	6	3	2	5.33	25.9
P15	14	9	7	1	6.33	22.5
P16	6	6	3	3	6.67	18.9
P17	7	6	4	1	4.67	24.9
P18	6	6	5	0	6.00	15.8
P19	6	6	5	1	5.83	27.7
P20	13	7	5	2	6.57	30.4
P21	7	7	5	1	6.00	22.0
P22	5	5	4	1	6.60	26.1
P23	7	6	4	2	6.83	17.6
P24	5	5	4	1	7.00	26.3
Mean	7.3	6.0	4.3	1.0	6.35	24.3

participants appropriated tools, formed workflows, and calibrated trust during verification. Because interaction logs alone cannot capture the reasoning behind tool choices or trust judgments, participants followed a concurrent think-aloud protocol [17], verbalizing their reasoning, hypotheses, and motivations for switching between tools throughout the task. For each investigated claim, they submitted a verdict (verified, misleading, or unverifiable) along with a confidence rating (1–7) and a ranked list of which tools they found helpful. For misleading or unverifiable claims, they additionally provided a severity rating and a short explanation or correction.

Phase III: Post-study questionnaire and interview (10 min). After the main task, participants completed a questionnaire covering demographics, prior experience, perceived helpfulness of each tool (7-point Likert), the NASA Task Load Index (NASA-TLX; 7-point scale, six dimensions), and the System Usability Scale (SUS; 10 items, 5-point scale). The session ended with a semi-structured interview probing four topics: claim detection strategy, verification strategy, tool switching triggers, and tool utility with AI trust preferences.

4.4 Data Collection and Analysis

FYI logged all user interactions as timestamped events, capturing 2,250 events across 25 event types. From these logs, we derived a per-participant metrics file (22 rows, 64 columns) aggregating session-level measures with questionnaire responses, and a master event table (2,250 rows) preserving all raw interactions.

For qualitative analysis, think-aloud protocols and interview transcripts were analyzed using thematic analysis [7]. Three researchers independently read all transcripts, produced per-participant memos, and iteratively developed themes through constant comparison across three coding passes. Themes were then discussed and consolidated to reach consensus. The resulting themes are integrated with quantitative findings in Sec. 5.

5 Results

The following analysis combines interaction logs, self-reported metrics, think-aloud protocols, and post-study interviews. We organize findings around the four research questions, preceded by a behavioral overview. Table 4 summarizes the qualitative themes identified through thematic analysis, which structure the narrative in each subsection.

5.1 Overview

Participants highlighted 161 claims across all sessions (88 AI-detected, 73 manually added). Five were dismissed, leaving 156 actively investigated claims. Across the 22 participants, 133 verdicts were submitted, 95 verified (71.4%), 23 misleading (17.3%), and 15 unverifiable (11.3%). These counts describe how broadly participants chose to investigate the article. They should not be read as evidence that finding more claims is always better, since exhaustive verification may increase workload in real reading contexts.

Table 3. Tool usage summary. Claims = claims where the tool was used at least once; Users = participants who used the tool; Most Helpful = percentage of verdicts selecting this tool as most helpful; Helpfulness = post-task rating (1–7 Likert).

Tool	Claims	Users	Most Helpful	Helpfulness
Auto Check	90 (57.7%)	18 (82%)	22.6%	5.59 (±1.05)
AI Chat	91 (58.3%)	21 (95%)	35.3%	5.14 (±1.55)
Chart Builder	105 (67.3%)	22 (100%)	30.1%	5.59 (±1.50)
Table Explorer	45 (28.8%)	21 (95%)	12.0%	5.27 (±1.28)

Session durations ranged from 12.5 to 33.3 minutes ($M = 24.3$, $SD = 5.0$), with participants highlighting 7.3 claims on average ($SD = 3.2$) and submitting 6.0 verdicts ($SD = 1.6$). Individual differences were pronounced. The number of verdicts ranged from 4 (P08, P12) to 11 (P02), and P08 and P12 investigated only AI-detected claims, while P02 and P15 manually added numerous claims (Tab. 2).

System usability was rated "good" (SUS $M = 74.1$, $SD = 13.4$) with moderate cognitive load (NASA-TLX $M = 3.1/7$, $SD = 0.7$), suggesting that the multi-tool design was feasible without overwhelming participants.

5.2 RQ1: How Do Users Detect Data Claims?

FYI offered two ways to surface claims: AI auto-detection and manual highlighting. Participants differed mainly in which pathway they began and why.

AI-first detection (15/22; Tab. 4). The majority of participants used AI detection as an efficient starting point, then supplemented it with manual scanning. As P01 noted, "*I was lazy, I was reluctant to pick the claims one by one.*" However, few treated AI detection as sufficient. For instance, P11 let AI find four claims, then manually added two more. Trust in AI detection was moderate even within this group. P09 observed that AI "*always detects a very, very long sentence*" rather than an accurate claim boundary, requiring manual adjustment.

Manual-first detection (7/22; Tab. 4). A smaller group systematically scanned for quantitative cues such as numbers, superlatives, and comparators before running AI as a check. P02 explained, "*I don't want to be too dependent on using the AI, so first I just do it manually.*" Data-savvy participants (P02, P04, P15) consistently preferred this approach, with P15 finding manual highlighting faster than waiting for AI to process the article.

Overall, AI-first detection was more common, but participants rarely treated AI detection as complete. Manual-first participants scanned by themselves to maintain control before optionally using AI as a check.

5.3 RQ2: How Do Users Verify Data Claims?

All four verification tools achieved broad adoption (Tab. 3). On average, participants used 3.7 out of 4 tools ($SD = 0.48$, range: 3–4). When measuring per-claim adoption (a tool being used at least once per investigated claim), Chart Builder was used most frequently (105 claims, 67.3%), followed by AI Chat (91 claims, 58.3%) and Auto Check (90 claims, 57.7%), while Table Explorer was used less frequently (45 claims, 28.8%). Despite the frequency difference, post-task helpfulness ratings were uniformly positive ($M > 5.0/7$ for all tools).

Think-aloud protocols and interaction logs suggested three recurring workflow orientations, though participants often shifted between them across claims instead of following a fixed strategy throughout the session (Fig. 2). The most common strategy was *AI-first, manual confirmation* (9/22; Tab. 4). These participants initiated their process with Auto Check or AI Chat to establish a baseline verdict, subsequently verifying the findings through Chart Builder or Table Explorer. As P12 explained, "*Auto Check and AI Chat have the same result… then I double-check with Chart Builder to make sure.*" Conversely, a second group favored *manual-first, AI as supplement* (6/22; Tab. 4). P24 described this sequence: "*after I see the chart, I can have 90% confidence*

Fig. 2. Tool-use sequences for every participant (across 22 participants, 131 claims). Each pill is one claim, tracing the ordered tools used to evaluate it from left to right and ending at the verdict (♦); claims appear in the order they were addressed. Pill shading marks whether the claim opened with an AI tool (Auto Check / AI Chat) or a manual tool (Chart Builder / Table Explorer).

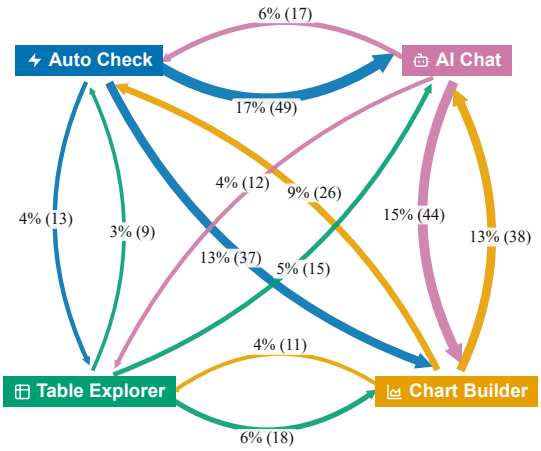

Fig. 3. Transitions between tools (289 transitions across 131 claims). Arrows run from the previously used tool to the next; line thickness encodes each transition's share of all transitions.

that I have the correct answer. Then I will use Auto Check to finish the rest 10%." Notably, P15 never used AI for verification, entirely relying on a manual approach. The third group engaged in *parallel co-review* (4/22; Tab. 4), running AI and manual tools simultaneously. P14 clarified this strategy as treating AI as a collaborator, "*set it on, like, a sub-agent kind of mode, and then do my own thing, and then converge to see if we come up with the same verdict.*"

More notably, these archetypes were not fixed types but tendencies. Participants adapted their approach across claims as system reliability and familiarity shifted, often moving between archetypes within a single session. For instance, P13 initially relied on an AI-first approach but pivoted to a Chart Builder-dominant approach after Auto Check gave an incorrect verdict. By contrast, P24 began with manual exploration but gradually incorporated Auto Check as a background process as their comfort level increased. The tool-transition graph (Fig. 3) reflects this fluidity. Rather than following a fixed order, participants moved frequently among tools. The most common single transition was Auto Check → AI Chat, while AI Chat and Chart Builder formed the busiest two-way exchange. Table Explorer was used least overall, as it provided limited support for aggregation-focused claims.

Overall, tool transitions were primarily driven by a desire for triangulation and confidence-building rather than dissatisfaction with the initial tool. As P21 noted, "*It's not abandoning, it's more about reinforcing the claim.*"

5.4 RQ3: How Do Users Reach Verdicts?

Participants reported consistently high confidence across all verdict types ($M = 6.35/7$, $SD = 1.20$), spending an average of 33.6 seconds ($SD = 15.2$) on each verdict. Confidence was comparably high for verified ($M = 6.6$) and misleading ($M = 6.3$) claims, but notably lower for unverifiable ones ($M = 4.8$, $SD = 1.78$), reflecting the inherent uncertainty of that judgment. Thematic analysis identified three mechanisms driving participants' final determinations.

Visual evidence as stopping criterion (10/22; Tab. 4). For these participants, a self-constructed chart in Chart Builder that clearly aligned with or debunked a claim was often decisive, at times overriding the other tools' conclusions. In 55 instances, users deliberately launched Chart Builder *after* Auto Check completed, using charts to audit AI conclusions. P02 noted, "*If the chart already supports or dismisses the claim, then I can confidently say that I've already verified it.*" However, this created a usability–confidence paradox. Participants with lower visualization literacy (e.g., P17, who found charting "*too much of a mental task,*" avoided Chart Builder entirely) were forced to rely on AI tools they trusted less.

Multi-tool convergence (10/22; Tab. 4). In contrast, these participants treated no single tool as sufficient. They stopped only once multiple independent tools *agreed*, so a divergence among tools meant continuing to investigate rather than concluding. P19 illustrated this layered confidence as "*If both these agents give me the correct answer, I'm mostly sure it is correct. If additionally I'm able to plot it on the chart by myself, then it is 100% correct.*"

Claim decomposition and statistical reasoning (15/22; Tab. 4). Reaching a verdict required substantive analytical reasoning beyond tool operation. Participants actively decomposed complex sentences into testable sub-claims (P17: "*there are two claims within the claim*") and debated statistical nuances like correlation versus causation or mean versus median. Furthermore, they successfully recognized when claims required external context beyond the provided dataset, correctly labeling them as "unverifiable."

5.5 RQ4: How Do Users Calibrate Trust between AI and Manual Tools?

To study how participants calibrated their trust, we first examine how accurate Auto Check's verdicts were. We compared the researchers' verdicts on the six embedded claims against outputs from Auto Check and the participants, and this analysis covers only the 79 verdict instances for which all three labels were available (Fig. 4). Overall, Auto Check and participants matched the researchers' labels at similar rates of 71% and 76%, respectively, yet performance varies by claim type. For *verified* claims, they both aligned with the researchers' verdicts nearly universally. Auto Check correctly assessed 44 of 45 verdicts and participants 42 of 45. While on the *misleading* claims, alignment was weaker for both: Auto Check matched the *misleading* label on 41% of these verdicts (12 of 29) and participants on 52% (15 of 29), with the rest mostly misclassified as *verified*. Most misjudgments came from one specific claim that takes a factual statistic that dramas hold the highest average IMDb rating and twists it into a misleading causal inference that audiences *universally prefer serious storytelling*.

The clearest disagreement appeared on the single *unverifiable* claim, which cannot be proven or disproven using the dataset alone. Across 79 claims, Auto Check returned an *unverifiable* verdict only once, but never on the unverifiable claim defined by researchers. Instead,

Table 4. Qualitative themes from thematic analysis of think-aloud protocols and post-study interviews ($N = 22$). Themes are not mutually exclusive; participants may appear in multiple themes. PIDs enable cross-referencing with Tab. 2.

RQ	ID	Theme	n	Participants
RQ1: Detection	T1	AI-first detection, manual supplement	15	P01, P05, P06, P07, P08, P09, P11, P12, P16, P17, P19, P21, P22, P23, P24
	T2	Manual-first detection, AI as recheck	7	P02, P04, P13, P14, P15, P18, P20
RQ2: Verification	T3	AI-first workflow, manual confirmation	9	P01, P08, P11, P12, P16, P17, P19, P21, P23
	T4	Manual-first workflow, AI as supplement	6	P02, P04, P15, P18, P20, P24
	T5	Parallel co-review	4	P07, P14, P19, P22
RQ3: Verdicts	T7	Visual evidence as stopping criterion	10	P01, P02, P05, P07, P13, P14, P19, P20, P22, P24
	T8	Multi-tool convergence as confidence signal	10	P05, P08, P12, P14, P17, P19, P20, P21, P22, P24
	T9	Claim decomposition and statistical reasoning	15	P02, P04, P05, P07, P09, P12, P13, P14, P17, P18, P19, P20, P22, P23, P24
RQ4: Trust	T10	Trust builds through convergence, erodes through inconsistency	12	P05, P07, P08, P12, P13, P14, P17, P19, P20, P21, P22, P24
	T11	Transparency modulates trust	6	P06, P17, P20, P21, P22, P24
	T12	AI initiates, human decides (dominant reliance model)	13	P01, P05, P07, P08, P09, P11, P12, P14, P17, P18, P20, P21, P22

it consistently categorized it as *verified* or *misleading*. By contrast, participants labeled this claim as *unverifiable* 3 times out of 5 total verdicts they gave. On the single unverifiable claim, Auto Check never matched the researchers' label, whereas participants did so in 3 of 5 cases. This pattern points to a possible blind spot: Auto Check was accurate on the straightforward, data-answerable claims but did not flag the unverifiable claim that required contextual judgment. Participants mitigated this limitation by using AI for initial analysis while retaining final judgments for themselves. We further analyze their trust-calibration pattern as follows.

Convergence builds trust, inconsistency erodes it (12/22; Tab. 4). Trust calibrated dynamically with experience. Alignment between AI and manual tools boosted confidence, while numerical inconsistencies, even with directional agreement, quickly eroded it. For example, P17 found different Pearson correlations from Auto Check and AI Chat: "*why are the numbers not consistent?*" Moreover, first impressions proved critical. P13 abandoned Auto Check for the rest of the session after an initial incorrect verdict. Interestingly, participants evaluated different output modalities independently, often trusting Auto Check's textual conclusions while disregarding the broken or incorrectly scaled charts it frequently generated.

Transparency modulates trust (6/22; Tab. 4). Participants trusted AI more when its reasoning process was visible. P20 preferred Auto Check because "*it shows the steps... so it feels more controllable*," while P22 deeply inspected AI Chat's Python code to verify its underlying logic. By exposing intermediate operations, the system could empower users to calibrate their trust based on comprehensible evidence rather than blind faith.

AI initiates, human decides (13/22; Tab. 4). The dominant reliance model positioned AI as an initial assessor, with humans retaining final decision-making authority. For instance, P18 said "*I prefer AI doing the initial. But not the end result.*" However, the strictness of this human-in-the-loop behavior appeared to vary with participants' data literacy and professional background. Individuals with high data literacy, such as AI/data professionals (P04, P12, P15), actively minimized their use of AI to maintain strict analytical control. In contrast, participants with lower data literacy (e.g., P16, P23) were far more willing to surrender this control, leaning toward full AI delegation rather than independent verification.

6 Discussion

Our findings suggest that navigating the detection–verification–determination pipeline is highly dynamic when multiple modalities are available simultaneously. We discuss what these behaviors reveal about multi-tool orchestration, human-AI complementarity, and trust dynamics, and distill these observations into design implications for future mixed-initiative fact-checking systems.

6.1 The Pipeline in Practice

Within our open-ended design-probe setting, participants rarely followed a strict detection–verification–determination sequence, and the boundaries between detection and verification largely dissolved. Participants frequently discovered additional check-worthy claims *during* the verification phase, prompting them to return to the text and manually highlight new statements. Verification itself proved deeply iterative. Participants actively decomposed complex claims, switched tools to test alternative interpretations, and occasionally revised earlier verdicts upon encountering contradictory evidence. This cyclical behavior aligns with sensemaking theory [48], suggesting that in-situ fact-checking involves fluid movement between information foraging and synthesis rather than a rigid forward progression.

FYI's role as an instrumented design probe helped capture this process. By embedding multiple tools within the reading context and logging tool transitions, FYI made it possible to observe how participants constructed workflows across claims and modalities. Instead of following a fixed tool order, participants adapted their workflows based on emerging evidence, the analytic operation required by the claim, and their evolving trust in the system.

6.2 Human-AI Complementarity and Visual Auditing

A central insight from our study is that participants did not treat AI and manual tools as substitutes, but as complements serving distinct epistemic roles. The emergence of diverse workflow archetypes—*AI-first with manual confirmation*, *manual-first with AI supplement*, and *parallel co-review*—reflects different strategies for distributing cognitive labor across the human-AI spectrum.

This multi-strategy behavior extends findings from interactive fact-checking systems such as WebSeek [25] and DataDive [33], where users combined extraction, tabular inspection, and chart construction. By introducing generative AI into the toolkit, FYI created a richer space for strategy composition. Notably, no single workflow dominated, and strategies fluidly shifted within individual sessions. This heterogeneity suggests that verification systems should support flexible composition rather than prescribing a fixed optimal process.

Crucially, visualization emerged as a distinct *verification modality*, not merely an auxiliary feature. Participants frequently launched Chart Builder *after* Auto Check completed, using self-built charts to audit AI conclusions. This pattern extends the emphasis-checking paradigm [31] by revealing a deeper cognitive preference. Users actively used visualization tools as instruments of oversight. They explicitly valued the agency of constructing their own evidence over passively consuming AI-produced summaries.

6.3 Trust Dynamics and the Usability-Confidence Paradox

Our results reveal trust as a dynamic process that shifts with accumulated experience rather than a stable individual trait [8]. Trust was built through cross-tool convergence, but was easily eroded when outputs were numerically inconsistent even if directionally correct, and in extreme cases collapsed permanently after a single AI error. These dynamics extend prior work on trust calibration by showing that in multi-tool environments, users evaluate output modalities independently, often accepting AI's textual conclusions while dismissing its flawed visual charts. A few participants chose to cross-validate one AI tool against another instead of referencing original data. They either used AI Chat to re-examine Auto Check (P11) or compared figures across the two AI tools (P17). However, such AI-on-AI checks could produce conflicting values that undermine user confidence in automated outputs, underscoring the value of auditing with self-built charts.

While high overall confidence and the similar ground-truth agreement patterns of Auto Check and participants (Fig. 4) raise valid con-

cerns regarding automation bias and anchoring effects, the *AI initiates, human decides* reliance model shows users actively attempting to maintain final authority. However, this resistance is gated by data literacy, resulting in a critical usability-confidence paradox. Participants who found chart authoring too cognitively demanding were forced to rely on automated AI tools that they otherwise trusted less. This suggests a vulnerability that limited visualization literacy may constrain a user's ability to audit algorithmic outputs. Furthermore, we found that process visibility, such as inspecting generated code, acted as a modulator of trust and was associated with more selective override behaviors.

6.4 Design Implications

Based on these empirical insights, we propose four implications for future mixed-initiative fact-checking systems.

DI_1: Design AI as an initial guide, not a definitive authority. The dominant *AI initiates, human decides* pattern suggests that AI is most useful when providing an investigative starting point. Systems should present AI outputs as provisional hypotheses that actively invite human verification, instead of definitive verdicts that require significant cognitive effort to override.

DI_2: Elevate visualization as a core auditing modality. Given its critical role in post-AI verification, user-constructed visualization confers a sense of agency and interpretive confidence that AI-generated evidence may not readily substitute. Fact-checking systems should integrate visualization authoring as a primary capability, and consider AI-assisted chart suggestion to lower the entry barrier for users with limited visualization literacy.

DI_3: Support flexible, cross-modal workflow composition. The fluidity of the three observed workflow archetypes argues against rigid, step-by-step verification wizards. Systems should provide independent, composable tools that users can freely arrange and interleave according to the type of evidence required, the amount of analysis involved, and their own expertise.

DI_4: Make AI reasoning and uncertainty visible. Because process transparency appears to modulate trust calibration, future systems must move beyond opaque final outputs. Systems should expose intermediate reasoning, data queries, and underlying computation steps, alongside confidence scores, to support informed calibration over blind acceptance or algorithmic aversion.

7 Limitations and Future Work

While our design probe provides rich behavioral insights, several factors constrain the scope of our findings.

Dataset availability and ecological scope. FYI assumes access to a relevant structured dataset, which allowed us to focus on fact-checking against inspectable data. This assumption fits data-driven articles that disclose their underlying data [63]. Yet it excludes articles that make quantitative claims without releasing the underlying data, or cases where the available data are incomplete, messy, or require substantial cleaning before verification. Future systems should support upstream data discovery, provenance assessment, and data preparation in addition to the dataset-grounded fact-checking workflow studied here.

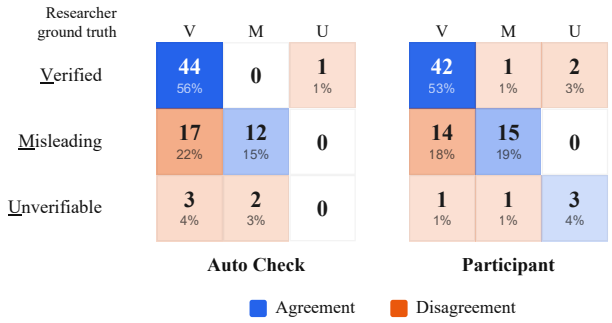

Fig. 4. Verdict agreement with researcher ground truth ($n = 79$). Rows give the researcher's ground-truth verdict; columns give the verdict from Auto Check (left) and the participant (right), where V = Verified, M = Misleading, U = Unverifiable. Diagonal cells (blue) are agreements and off-diagonal cells (orange) are disagreements; each cell reports the count and its share of all 79 verdict instances, and shade encodes magnitude.

Domain and sample. The study used a single article in the movie domain with one accompanying dataset. Verification strategies may differ substantially in domains such as finance or public health, where claims involve multi-table joins, real-time data, or specialized domain knowledge. Our university-affiliated participants reported high visualization literacy and AI familiarity, which may limit generalizability to readers with lower data or visualization literacy. The observed workflow diversity may therefore underestimate the challenges faced by less experienced users.

Accuracy and ground truth. Ground truth in our study was limited to the six embedded data claims, against which we report Auto Check's and participants' verdict accuracy (Sec. 5.5). The remaining 75 claims were surfaced by participants and lay outside this set; many are not data claims, and none carry researcher veracity labels, as they were never intended as study ground truth. Future studies could establish ground truth for every claim in an article and examine textual and data claims in conjunction, enabling a fuller assessment of verification accuracy across claim types.

System and model dependence. The current Table Explorer supports direct lookup and filtering, and future versions could add lightweight in-table analytical features, such as conditional formatting or simple grouped summaries, to reduce the need to switch tools for basic comparisons. In addition, all AI-powered components rely on GPT-4.1, whose outputs vary with model version, prompt design, and stochastic sampling. Known limitations in LLM numerical reasoning [62] and hallucination tendencies [46] affected participants' trust dynamics (e.g., Auto Check producing broken charts). The multi-tool design partially mitigates this by providing non-AI verification paths through Table Explorer and Chart Builder, but findings may not generalize to other model families or future model versions.

8 Conclusion

FYI is an in-situ browser extension which spans the human-AI spectrum from fully automated verification (Auto Check) through user-directed AI assistance (AI Chat) to unmediated data exploration (Table Explorer, Chart Builder). We used it as a design probe for an exploratory study ($N = 22$). Our findings revealed that participants did not follow a fixed linear pipeline. Instead, they constructed heterogeneous workflows, including AI-first with manual confirmation, manual-first with AI supplement, and parallel co-review. Visualization served as a primary auditing mechanism, as participants used self-built charts to inspect and challenge AI conclusions. Trust was calibrated dynamically based on cross-tool convergence, output inconsistency, and process transparency. These findings suggest that future fact-checking environments could use AI as an initial guide rather than a definitive authority, since participants tended to reach for it first and found it broadly accurate, while treating manual tools such as visualization as indispensable rather than secondary. This points toward keeping both humans and AI in the loop, prioritizing visualization as a core verification capability, and supporting flexible cross-modal workflows.

Supplemental Materials

To facilitate subsequent studies on fact-checking tools, we open-source the FYI software prototype at `https://github.com/DataVisards/FYI`. Supplemental material is available in the IEEE Xplore digital repository and includes (1) a video demonstration of FYI outlining its claim detection, multi-tool verification, and verdict determination workflows; (2) the study article and movie dataset used in the evaluation, along with the researcher ground-truth labels for the six embedded claims, the study protocol, the interview guide, and the questionnaire instrument; (3) all LLM prompts for claim detection, Auto Check, and AI Chat; (4) the raw per-participant interaction logs with a data dictionary; (5) the per-participant think-aloud transcripts; and (6) the post-task questionnaire responses.

Acknowledgments

The authors used Claude Code (Anthropic) to help write and debug portions of the FYI prototype's source code (Sec. 3); all AI-generated code was reviewed and tested by the authors. Generative AI was not used for the study design, data, or analysis. We thank the members of the DataVisards Lab at HKUST, user study participants, and anonymous reviewers for their feedback during various stages of this work.

References

[1] F. Aljamaan, M.-H. Temsah, I. Altamimi, A. Al-Eyadhy, A. Jamal, K. Alhasan et al. Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. *JMIR Med. Inform.*, 12, art. no. e54345, Jul. 2024. doi: 10.2196/54345 2

[2] P. M. Allen, J. A. Edwards, F. J. Snyder, K. A. Makinson, and D. M. Hamby. The effect of cognitive load on decision making with graphically displayed uncertainty information. *Risk Anal.*, 34(8):1495–1505, Aug. 2014. doi: 10.1111/risa.12161 2

[3] K. Aslett, Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625:548–556, Jan. 2024. doi: 10.1038/s41586-023-06883-y 4

[4] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu, Q. Massonnat, A. Deiana et al. Statistical claim checking: StatCheck in action. In *Proc. ACM Int. Conf. Inf. & Knowl. Manage. (CIKM)*, pp. 4798–4802. ACM, New York, Oct. 2022. doi: 10.1145/3511808.3557198 3

[5] J. E. Block, S. Esmaeili, E. D. Ragan, J. R. Goodall, and G. D. Richardson. The influence of visual provenance representations on strategies in a collaborative hand-off data analysis scenario. *IEEE Trans. Visual. Comput. Graphics*, 29(1):1113–1123, Jan. 2023. doi: 10.1109/TVCG.2022.3209495 4

[6] B. Botnevik, E. Sakariassen, and V. Setty. BRENDA: Browser extension for fake news detection. In *Proc. ACM Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, SIGIR '20, pp. 2117–2120. ACM, New York, Jul. 2020. doi: 10.1145/3397271.3401396 2

[7] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qual. Res. Psychol.*, 3(2):77–101, Jan. 2006. doi: 10.1191/1478088706qp063oa 6

[8] S. Cao and C.-M. Huang. Understanding user reliance on AI in assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), art. no. 471, 23 pp., Nov. 2022. doi: 10.1145/3555572 3, 8

[9] B. G. d. S. Cezar and A. C. G. Maçada. Cognitive overload, anxiety, cognitive fatigue, avoidance behavior and data literacy in big data environments. *Inf. Process. & Manage.*, 60(6), art. no. 103482, Nov. 2023. doi: 10.1016/j.ipm.2023.103482 2

[10] J. H. Chae and D. Tewksbury. Perceiving AI intervention does not compromise the persuasive effect of fact-checking. *New Media Soc.*, 28(1):191–211, Jan. 2026. doi: 10.1177/14614448241286881 3

[11] A. Chegini, K. Rezaei, H. Eghbalzadeh, and S. Feizi. RePanda: Pandas-powered tabular verification and reasoning. In *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 32200–32212. Association for Computational Linguistics, Vienna, Austria, Jul. 2025. doi: 10.18653/v1/2025.acl-long.1549 3

[12] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li et al. TabFact: A large-scale dataset for table-based fact verification. In *Proc. Int. Conf. Learn. Representations (ICLR)*, art. no. rkeJRhNYDH. OpenReview, Virtual Conference, Jun. 2020. doi: 10.48550/arXiv.1909.02164 2, 3

[13] Z.-T. Chen and H. Xia. CrossData: Leveraging text-data connections for authoring data documents. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '22, art. no. 95. ACM, New York, Apr. 2022. doi: 10.1145/3491102.3517485 3

[14] Z. Cheng, T. Xie, P. Shi, C. Li, R. Nadkarni, Y. Hu et al. Binding language models in symbolic languages. In *Proc. Int. Conf. Learn. Representations (ICLR)*, art. no. lH1PV42cbF. OpenReview, Kigali, Rwanda, May 2023. doi: 10.48550/arXiv.2210.02875 2, 3, 5

[15] A. K. Das and K. Mueller. MisVisFix: An interactive dashboard for detecting, explaining, and correcting misleading visualizations using large language models. *IEEE Trans. Visual. Comput. Graphics*, 32(1):134–144, Jan. 2026. doi: 10.1109/TVCG.2025.3633884 3, 4

[16] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl. Acad. Sci. USA*, 121(50), art. no. e2322823121, Dec. 2024. doi: 10.1073/pnas.2322823121 4

[17] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, rev. ed., 1993. doi: 10.7551/mitpress/5657.001.0001 6

[18] Y. Fu, S. Guo, J. Hoffswell, V. S. Bursztyn, R. Rossi, and J. Stasko. "The data says otherwise" — towards automated fact-checking and communication of data claims. In *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, UIST '24, art. no. 134, 20 pp. ACM, New York, Oct. 2024. doi: 10.1145/3654777.3676359 1, 2, 3, 4

[19] K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Amer. Med. Inform. Assoc.*, 19(1):121–127, Jan. 2012. doi: 10.1136/amiajnl-2011-000089 2, 3

[20] L. Graves. Understanding the promise and limits of automated fact-checking. Factsheet, Reuters Inst. Study Journalism, Univ. Oxford, Oxford, U.K., Feb. 2018. doi: 10.60625/risj-nqnx-bg89 2

[21] Z. Gu, J. Fan, N. Tang, P. Nakov, X. Zhao, and X. Du. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pp. 4971–4983. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, Dec. 2022. doi: 10.18653/v1/2022.emnlp-main.331 3

[22] Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206, Feb. 2022. doi: 10.1162/tacl_a_00454 2, 4

[23] N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, KDD '17, pp. 1803–1812. ACM, New York, Aug. 2017. doi: 10.1145/3097983.3098131 2, 3

[24] T. L. Horstmann, B. Geisenberger, and M. Alam. T-REX: Table – refute or entail explainer. In *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discovery Databases (ECML PKDD)*, vol. 16022 of *Lecture Notes in Computer Science*, pp. 470–474. Springer, Porto, Portugal, Sep. 2025. doi: 10.1007/978-3-032-06129-4_33 3

[25] Y. Huang and A. Narechania. Facilitating proactive and reactive guidance for decision making on the web: A design probe with WebSeek. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '26, art. no. 800. ACM, New York, Apr. 2026. doi: 10.1145/3772318.3791945 4, 8

[26] K. Hyland. Stance and engagement: A model of interaction in academic discourse. *Discourse Stud.*, 7(2):173–192, May 2005. doi: 10.1177/1461445605050365 2

[27] F. Jahanbakhsh and D. R. Karger. A browser extension for in-place signaling and assessment of misinformation. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '24, art. no. 946. ACM, New York, May 2024. doi: 10.1145/3613904.3642473 2

[28] S. M. Jones-Jang and Y. J. Park. How do people react to AI failure? automation bias, algorithmic aversion, and perceived controllability. *J. Comput.-Mediated Commun.*, 28(1), art. no. zmac029, Jan. 2023. doi: 10.1093/jcmc/zmac029 3

[29] P. Juneja and T. Mitra. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), art. no. 418, Nov. 2022. doi: 10.1145/3555143 4

[30] G. Karagiannis, M. Saeed, P. Papotti, and I. Trummer. Scrutinizer: a mixed-initiative approach to large-scale, data-driven claim verification. *Proc. VLDB Endowment*, 13(12):2508–2521, Aug. 2020. doi: 10.14778/3407790.3407841 3

[31] D. H. Kim, S. Choi, J. Kim, V. Setlur, and M. Agrawala. EmphasisChecker: A tool for guiding chart and caption emphasis. *IEEE Trans. Visual. Comput. Graphics*, 30(1):120–130, Jan. 2024. doi: 10.1109/TVCG.2023.3327150 2, 3, 8

[32] D. H. Kim, E. Hoque, J. Kim, and M. Agrawala. Facilitating document reading by linking text and tables. In *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, UIST '18, pp. 423–434. ACM, New York, Oct. 2018. doi: 10.1145/3242587.3242617 2

[33] H. Kim, K. D. Le, G. Lim, D. H. Kim, Y. J. Hong, and J. Kim. DataDive: Supporting readers' contextualization of statistical statements with data exploration. In *Proc. ACM Int. Conf. Intelligent User Interfaces (IUI)*, IUI '24, pp. 623–639. ACM, New York, Mar. 2024. doi: 10.1145/3640543.3645155 2, 3, 4, 8

[34] H.-P. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks et al. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '25, art. no. 1121. ACM, New York, Apr. 2025. doi: 10.1145/3706598.3713778 2, 4

[35] G. Lim and S. T. Perrault. XAI in automated fact-checking? The benefits are modest and there's no one-explanation-fits-all. In *Proc. Aust. Comput.-*

Human Interaction Conf. (OzCHI), OzCHI '23, pp. 624–638. ACM, New York, May 2024. doi: 10.1145/3638380.3638388 3

[36] C. Liu, Q. Zhou, X. Shen, X. B. Liu, T. Wu, and X. A. Chen. Behavioral indicators of overreliance during interaction with conversational language models. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '26, art. no. 790. ACM, New York, Apr. 2026. doi: 10.1145/3772318.3790332 2, 4

[37] T. Lloyd, T. Nguyen, K. Levy, and M. Naaman. Beyond community notes: A framework for understanding and building crowdsourced context systems for social media. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, CHI '26, art. no. 332. ACM, New York, Apr. 2026. doi: 10.1145/3772318.3791889 2

[38] D. Metropolitansky and J. Larson. Towards effective extraction and evaluation of factual claims. In *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 6996–7045. Association for Computational Linguistics, Vienna, Austria, Jul. 2025. doi: 10.18653/v1/2025.acl-long.348 2

[39] A. Narechania, K. Odak, M. El-Assady, and A. Endert. ProvenanceWidgets: A library of UI control elements to track and dynamically overlay analytic provenance. *IEEE Trans. Visual. Comput. Graphics*, 31(1):1235–1245, Jan. 2025. doi: 10.1109/TVCG.2024.3456144 4

[40] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Trans. Visual. Comput. Graphics*, 27(2):369–379, Feb. 2021. doi: 10.1109/TVCG.2020.3030378 5

[41] A. T. Nguyen, A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace et al. Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, pp. 189–199. ACM, New York, Oct. 2018. doi: 10.1145/3242587.3242666 3

[42] Y. Nie, H. Chen, and M. Bansal. Combining fact extraction and verification with neural semantic matching networks. *Proc. AAAI Conf. Artif. Intell.*, 33(1):6859–6866, Jul. 2019. doi: 10.1609/aaai.v33i01.33016859 2

[43] Office for National Statistics. Deaths involving COVID-19 by vaccination status, England: deaths occurring between 2 January and 2 July 2021. ONS Statistical Bulletin, 2021. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19byvaccinationstatusengland/deathsoccurringbetween2januaryand2july2021. Accessed: Mar. 31, 2026. 1

[44] K. Okamura and S. Yamada. Adaptive trust calibration for human-AI collaboration. *PLoS ONE*, 15(2), art. no. e0229132, Feb. 2020. doi: 10.1371/journal.pone.0229132 3

[45] OpenAI. GPT-4.1. OpenAI API model documentation, 2025. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-4.1. Accessed: Jul. 27, 2026. 4

[46] A. Pesaranghader and E. Li. Hallucination detection and mitigation in large language models, 2026. arXiv:2601.09929. [Online]. Available: https://arxiv.org/abs/2601.09929. doi: 10.48550/arXiv.2601.09929 2, 4, 9

[47] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. Int. Conf. Intelligence Analysis*, vol. 5, pp. 2–4. MITRE, McLean, May 2005. 3

[48] P. Pirolli and D. M. Russell. Introduction to this special issue on sensemaking. *Hum.–Comput. Interact.*, 26(1–2):1–8, Jan. 2011. doi: 10.1080/07370024.2011.556557 8

[49] J. S. Rubinstein, D. E. Meyer, and J. E. Evans. Executive control of cognitive processes in task switching. *J. Exp. Psychol.: Hum. Percept. Perform.*, 27(4):763–797, Aug. 2001. doi: 10.1037/0096-1523.27.4.763 4

[50] M. S. S. Samu, N. Khan, K. T. Elahi, T. B. Rahman, M. R. Islam, and F. Sadeque. AI as teammate or tool? a review of human–AI interaction in decision support, 2026. arXiv:2602.15865. [Online]. Available: https://arxiv.org/abs/2602.15865. doi: 10.48550/arXiv.2602.15865 3

[51] D. E. P. Schultz. Truth goggles: automatic incorporation of context and primary source for a critical media experience. S.M. thesis, Program Media Arts Sci., School Archit. Planning, Massachusetts Inst. Technol., Cambridge, 2012. [Online]. Available: http://hdl.handle.net/1721.1/76530. 4

[52] N. Sultanum and A. Srinivasan. DataTales: Investigating the use of large language models for authoring data-driven articles. In *Proc. IEEE Visualization Visual Analytics (VIS)*, pp. 231–235. IEEE, Melbourne, Australia, Oct. 2023. doi: 10.1109/VIS54172.2023.00055 2, 3

[53] M. Theologitis and D. Suciu. Thucy: An LLM-based multi-agent system for claim verification across relational databases, 2026. Presented at the AAAI Workshop on LLM-Based Multi-Agent Systems (LaMAS), Singapore, Jan. 2026. doi: 10.48550/arXiv.2512.03278 2

[54] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, pp. 809–819. Association for Computational Linguistics, New Orleans, Jun. 2018. doi: 10.18653/v1/N18-1074 2

[55] A. Vlachos and S. Riedel. Identification and verification of simple claims about statistical properties. In *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pp. 2596–2601. Association for Computational Linguistics, Lisbon, Portugal, Sep. 2015. doi: 10.18653/v1/D15-1312 2

[56] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan et al. Fact or fiction: Verifying scientific claims. In *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, Online, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.609 2

[57] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 422–426. Association for Computational Linguistics, Vancouver, Canada, Jul. 2017. doi: 10.18653/v1/P17-2067 2

[58] Z. Wang, H. Zhang, C.-L. Li, J. M. Eisenschlos, V. Perot, Z. Wang et al. Chain-of-Table: Evolving tables in the reasoning chain for table understanding. In *Proc. Int. Conf. Learn. Representations (ICLR)*, art. no. 4L0xnS4GQM. OpenReview, Vienna, Austria, May 2024. doi: 10.48550/arXiv.2401.04398 2, 3

[59] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans. Visual. Comput. Graphics*, 22(1):649–658, Jan. 2016. doi: 10.1109/TVCG.2015.2467191 5

[60] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand et al. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proc. ACM CHI Conf. Human Factors Comput. Syst. (CHI)*, pp. 2648–2659. ACM, New York, May 2017. doi: 10.1145/3025453.3025768 5

[61] Y. Ye, B. Hui, M. Yang, B. Li, F. Huang, and Y. Li. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proc. ACM Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, pp. 174–184. ACM, New York, Jul. 2023. doi: 10.1145/3539618.3591708 2, 3

[62] Z. Yuan, H. Yuan, C. Tan, W. Wang, and S. Huang. How well do large language models perform in arithmetic tasks?, 2023. arXiv:2304.02015. [Online]. Available: https://arxiv.org/abs/2304.02015. doi: 10.48550/arXiv.2304.02015 2, 4, 9

[63] R. Zamith. Transparency, interactivity, diversity, and information provenance in everyday data journalism. *Digit. Journalism*, 7(4):470–489, Jan 2019. doi: 10.1080/21670811.2018.1554409 1, 4, 9