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

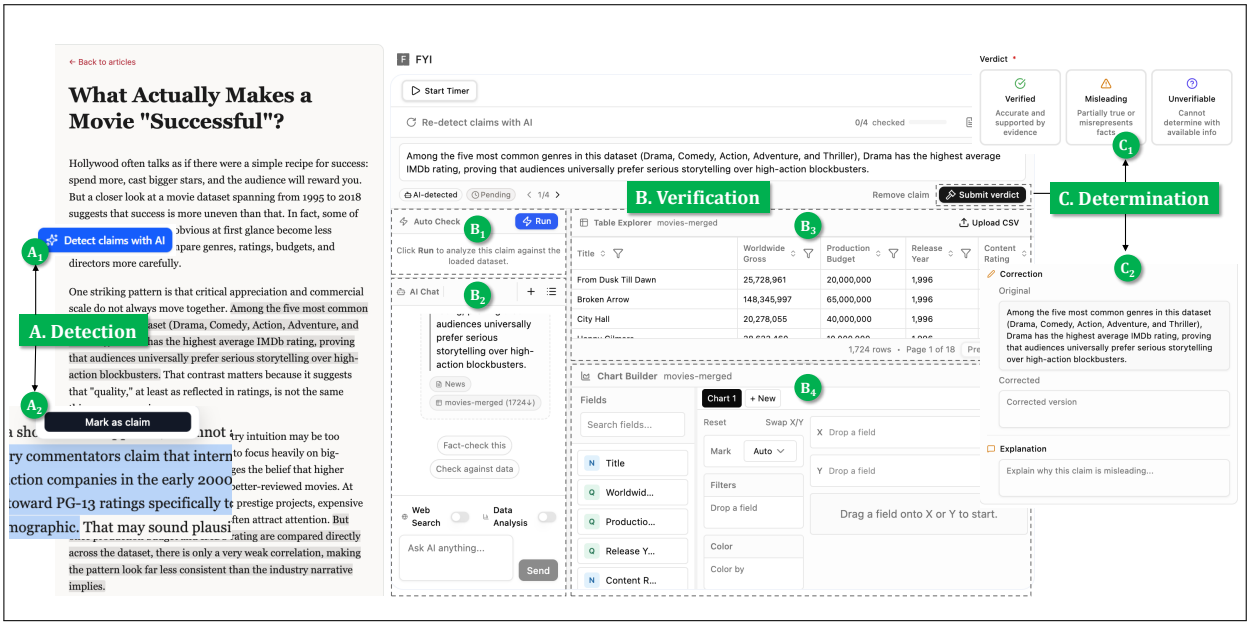

Fig. 1: The interface of FYI, a browser extension that enables in-situ data claim fact-checking directly alongside the article under review. The unified side panel supports a complete mixed-initiative pipeline: **(A) Detection** via (A1) AI detection or (A2) manual identification; **(B) Verification** 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 claims—statements grounded in numbers and statistics—are common in journalism and policy reports, yet verifying them requires significant analytical effort that most readers cannot undertake alone. Existing tools either fully automate verification, risking blind trust in AI, or relying entirely on manual data exploration with a high cognitive burden. We present FYI, a browser extension that bridges this gap through four complementary tools within a unified side panel, spanning the spectrum from automation to user-driven exploration. Using FYI as a design probe, we conducted an exploratory study ($N = 22$) in which participants verified claims in a data-driven article while thinking aloud. 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 was not static but shifted dynamically. Specifically, it grew when multiple tools converged on the same answer and eroded when AI outputs were inconsistent. These findings suggest that fact-checking systems should treat AI as a starting point rather than a definitive authority, elevate visualization as a core verification capability, and support flexible, user-driven workflows. We contribute FYI as open-source at https://github.com/DataVisards/FYI.

Index Terms—Fact-checking, data claims, human-AI interaction, design probe, browser extension, visualization.

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 [16], routinely found in data journalism and policy reporting [62]. Unlike textual fact-checking, which relies on finding corroborating sources, verifying a data claim requires specialized tools to aggregate, compare, and interpret structured data. This illusion of rigor makes misrepresented statistics easily accepted by readers who lack the time or expertise to analyze them.

Systematically addressing this requires a well-established pipeline: claim *detection* (identifying check-worthy statements), *verification* (evaluating evidence), and *determination* (reaching a judgment) [18,20]. Detection and verification constitute the analytical core of data claim fact-checking and serve as our primary focus. Determination, the final synthesis of evidence into a user-authored judgment, is also considered.

Existing systems span the human-AI spectrum but remain fragmented: fully automated pipelines are unreliable on precise numerical claims [45,61], interactive analytics platforms impose high cognitive

- *Nguyen-Truong Thinh* and Arpit Narechania are with The Hong Kong University of Science and Technology. E-mails: truongthinh.nguyen03@gmail.com, arpit@ust.hk*
- *Yuxuan Du* is an Independent Contributor. E-mail: yuxuan.du.sherry@gmail.com*
- *Phongsakon Mark Konrad is a research collaborator from SDU Centre for Industrial Software, University of Southern Denmark. E-mail: phkon23@student.sdu.dk*

**These 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

burden [33], crowdsourced approaches lack analytical depth [38], and lightweight browser extensions provide only shallow credibility signals [5,26] (see section 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 toward AI, users may accept outputs without examining the data (automation bias [17,35,37]), while full human control risks cognitive overload [4, 34]. Users must actively decide how much agency to delegate at each step, constituting a form of trust calibration that may shift across phases, 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.

There is a significant gap in understanding how people actually behave when fact-checking data claims with multiple modalities simultaneously available. What is needed is not another standalone tool but a *design probe*—an integrated environment covering the human-AI spectrum that enables observation of naturalistic fact-checking behaviors [47].

To address these gaps, 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 people balance trusting an AI and staying in control throughout the fact-checking process?

To answer these questions, we designed and developed a prototype Fact-check Your Information (FYI), a browser extension that embeds the fact-checking workflow directly within the reading environment via a unified side panel. For detection, users can manually highlight claims or use AI detection; for verification, four tools span the automation spectrum: a fully automated pipeline (Auto Check), a conversational AI agent (AI Chat), an interactive table (Table Explorer), and a visualization builder (Chart Builder). Users conclude by submitting a verdict and annotating 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, including tool transitions, query reformulations, and verdict revisions, the reading session itself serves as an observational window into users' fact-checking process. To contextualize these system logs with qualitative insights, we concurrently captured think-aloud protocols and conducted post-study interviews. Our primary contributions include:

1. Empirical characterization of user sensemaking behaviors, revealing how individuals compose fact-checking strategies and calibrate confidence across multiple modalities.
2. Concrete design implications for future mixed-initiative fact-checking systems.
3. An open-source, multi-modal browser extension, FYI (`https://github.com/DataVisards/FYI`), designed as an in-situ design probe to observe data claim fact-checking.

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 verification process is organized into a three-stage pipeline [20], namely detecting check-worthy statements, retrieving textual evidence, and predicting a final veracity label. Early benchmarks such as FEVER [53], LIAR [55] and SciFact [54] operated strictly within this framework. For text claims, detection typically involves identifying rhetorical markers or subjective versus objective framing to determine if a statement is worthy of investigating [25]. 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 [21].

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 [29]. Once 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 [10] to derive the answer from the raw data. Thus, fact-checking of data claims has inherent higher computational and cognitive complexity compared to pure text processing.

2.2 Paradigms of Fact-Checking Systems

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 [22] and subsequent LLM-based extractors [39] attempt to automatically flag check-worthy statements without user intervention. For verification, the field has shifted toward neural-symbolic decomposition strategies. Binder [12] parses claims into executable SQL or Python expressions, while DATER [59] decomposes tables into focused sub-tables for LLM reasoning, and Chain-of-Table [56] applies iterative table operations to evolve evidence through a reasoning chain. Recent end-to-end architectures, such as Aletheia [16], attempt to automate the entire lifecycle from semantic parsing to generating interactive data evidence representations, and multi-agent systems like Thucy [52] 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 [7]. Although Aletheia provides interactive widgets for overriding AI-inferred filters, its user agency remains limited to error correction rather than open-ended exploration.

On the contrary, the interactive paradigm prioritizes direct data inspection. DataTales [51] supports LLM-assisted authoring of data-driven articles with integrated visual evidence, while Emphasis-Checker [31] empowers users to construct analytical charts to independently verify text-chart consistency. Both treat visualization as a robust modality for active hypothesis testing rather than passive consumption. 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 has a premise 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 a relatively high analytical burden and expertise requirements [4].

Recently, mixed-initiative systems have been intended to bridge the gap. At the organizational scale, Scrutinizer [30] demonstrated a structured mixed-initiative approach where 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 [2] 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 [23] 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 emphasize manual analytics that overwhelm 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

Table 1: Comparison of data fact-checking systems across four dimensions: Claim Detection (M1), Claim Verification (M2), Claim Determination (M3), and Human Agency (M4).

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*
ClaimBuster [22]	KDD '17	Y	P	Y	N	N	N	P	N	N	N
TabFact [10]	ICLR '20	N	N	Y	N	Y	N	N	N	N	N
PASTA [19]	EMNLP '22	N	N	Y	N	Y	N	N	N	N	N
DATER [59]	SIGIR '23	N	N	Y	N	Y	N	N	N	N	N
Binder [12]	ICLR '23	N	N	Y	N	Y	N	N	N	P	N
Chain-of-Table [56]	ICLR '24	N	N	Y	N	Y	N	N	N	N	N
RePanda [9]	ACL '25	N	N	Y	N	Y	N	N	N	N	N
Believe it or not [42]	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 [2]	CIKM '22	Y	P	Y	N	Y	N	N	N	P	N
CrossData [11]	CHI '22	Y	Y	P	N	Y	Y	N	Y	Y	Y
DataTales [51]	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 [16]	UIST '24	Y	Y	Y	N	Y	Y	Y	Y	P	Y
MisVisFix [13]	TVCG '25	Y	Y	Y	Y	N	Y	P	Y	Y	Y
T-REX [23]	ECML '25	N	Y	Y	N	Y	Y	N	Y	N	Y
FYI (ours)	—	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

sensemaking process. Foundational models of sensemaking describe data analysis as an iterative cycle of information searching, hypotheses establishing, and evidence synthesizing [46]. In the fact-checking context, this loop spans both the detection and verification phases. Users must first go through the text to detect implicit data claims, and then verify the underlying structured evidence. When interactive visual tools are combined with raw tabular inspection, they support this sensemaking loop by making users actively interrogate data.

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 [17]. 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 [8]. 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 [27], 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 [60]. This requires interactive override capabilities that allow users to recover from AI errors and re-exert agency [42]. Critically, the reliance level may differ across the detection and verification phases [63]. 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 [36] 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 [49] 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.

While existing research studied the fact-checking workflows and tool usages of professionals [28], empirical studies on normal readers 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 around our detection–verification–determination pipeline extended with a human agency dimension, which collectively motivated FYI's design.

From the top half of the table, the NLP systems lack agency and determination. The top half of the table highlights a structural limitation of automated NLP pipelines. With very few exceptions (such as Binder's partial AI override or ClaimBuster's limited manual inputs), these systems largely lack support for Claim Determination (M3) and Human Agency (M4). Furthermore, they largely ignore manual claim selection, treating detection as an automated preprocessing step rather than a user-driven behavior. Because these models are evaluated against static benchmark inputs rather than interactive use cases, interactive

operations and multi-tool orchestration fall entirely outside their design scope.

From the bottom half of the table, the HCI systems lack conversational and automated depth. These visual and mixed-initiative systems excel at providing human agency. 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 comes at the cost of automated depth. For instance, DataTales and DataDive offer only partial support for AI-assisted fact-checking, leaving a relatively high analytical burden for users. More critically, across both the NLP and HCI paradigms, natural language dialog remains largely absent. While Scrutinizer offers structured query-based interaction and DataDive allows optional free-form text input, only MisVisFix [13] provides a multi-turn chat interface — though it targets misleading visualization correction rather than data claim verification against structured evidence.

This matrix underscores the unique positioning of FYI. It is the only system designed to provide wide coverage across the entire human-AI spectrum. By introducing natural language dialog(AI Chat) alongside automated verification(Auto Check), tabular inspection(Table Explorer), and visual analytics(Chart Builder), FYI bridges the gap between the computational power of NLP pipelines and the user-centered agency of HCI tools. This unified architecture provides the necessary foundation for our exploratory user study, allowing us to empirically observe how users orchestrate these previously fragmented modalities to fact-check data claims.

3 System Design

FYI is a Chrome extension that embeds the complete investigation workspace in a side panel alongside the article under review.

3.1 Design Goals

Drawing on prior work in automated fact-checking [16, 20], interactive data exploration [24, 33], and analytical provenance [3, 40], we identified three design goals.

DG_1: In-situ integration. Existing fact-checking tools often require users to leave their reading environment [50], incurring context-switching costs [48]. 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 [16]: AI pipelines accelerate evidence retrieval but may hallucinate [45, 61], tabular inspection reveals precise values, visualizations expose distributional patterns, and web search provides external corroboration [1]. 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 [14, 35, 37]. 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: DG_1 embeds the workflow in context, DG_2 spans detection and verification modalities, and DG_3 captures interaction data for analyzing trust calibration.

3.2 System Overview

FYI is implemented as a Chrome extension with a side panel that hosts the investigation workspace alongside the article under review, communicating with a backend API for LLM-powered claim detection and verification.

The user workflow maps onto the detection–verification–determination pipeline:

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 (subsection 3.3).
2. **Verification.** The user selects a claim and investigates it using four complementary tools (subsection 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 (subsection 3.5).

All AI-powered components (claim detection, Auto Check , and AI Chat) use GPT-4.1 [44] via OpenRouter, except web search which uses Perplexity Sonar; all prompts are included in the supplemental materials (Sec. 8).

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 casual readers may not recognize as verifiable [16], so automated detection lowers the entry barrier and ensures broad coverage. An LLM classifies each sentence as a potential data claim, 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 curation. 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 ($\mathbf{DG_3}$). 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. Conversely, users can 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 ($\mathbf{DG_2}$), 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 ($\mathbf{DG_3}$).

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 [12]), (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

While 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 [57,58], 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 section 5.

4 Study

We conducted an exploratory study to examine how participants fact-check data claims when multiple AI and manual tools are simultaneously available. Designed as a design probe evaluation [24], the study addresses our four research questions by observing detection strategies (RQ1), verification workflows (RQ2), verdict formation (RQ3), and trust calibration between human and AI modalities (RQ4).

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 ($M = 3.68/7$, $SD = 1.94$). All participants provided informed consent, and we handled data in accordance with standard ethical practices for low-risk, non-clinical human-subjects research. Each participant was compensated HKD $75 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 (1996–2018) with eight attributes (title, worldwide gross, production budget, release year, content rating, running time, genre, and IMDb rating) was constructed from publicly available sources [41]. 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). The full dataset is included in the supplemental materials (Sec. 8).

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. The article embedded claims that varied in both verification outcome (verified, misleading, and unverifiable) and operational complexity: easy claims could be checked through straightforward retrieval or sorting, medium claims required cross-group comparison or aggregation, and hard claims involved interpreting broader patterns or recognizing that the available data were insufficient. This design was intended to trigger a range of verification strategies rather than a single pathway. Beyond these embedded claims, participants were free to detect and investigate additional claims using AI detection or manual highlighting. The article is included in the supplemental materials (Sec. 8).

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.

Phase II: Main task (30 min). Participants were presented with the movie article and its grounding dataset. They first identified data claims—either by running AI detection, manually highlighting passages, or both—and then freely used any combination of FYI's four tools to verify those claims. Because interaction logs alone cannot capture the reasoning behind tool choices or trust judgments, participants followed a concurrent think-aloud protocol [15], 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), a severity rating, and a ranked list of which tools they found helpful.

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. Interview questions are included in the supplemental materials (Sec. 8).

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

Table 2: Per-participant summary (P03 and P10 excluded due to data loss; see section 4). Claims = total claims investigated; 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

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)

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 [6]. 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 section 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). 5 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%).

Session durations ranged from 12.5 to 33.3 minutes ($M = 24.3$, $SD = 5.0$), with participants investigating 7.3 claims on average ($SD = 3.2$) and submitting 6.0 verdicts ($SD = 1.6$). Individual differences were pronounced. The number of verdict 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 $M = 3.1/7$, $SD = 0.7$), confirming that the multi-tool design was feasible without overwhelming participants.

5.2 RQ1: How Do Users Detect Data Claims?

Thematic analysis of interview responses identified two primary detection strategies.

AI-first detection (15/22; Tab. 4). The majority of participants used AI detection as an efficient starting point, then supplemented 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 detect a very, very long sentence" rather than 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 AI's processing time.

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, AI Chat, and Auto Check saw comparable frequencies (90–105 claims each), while Table Explorer was used less frequently (45 claims). Despite the frequency difference, post-task helpfulness ratings were uniformly positive ($M > 5.0/7$ for all tools).

Three workflow archetypes. Think-aloud protocols and interviews indicated that users' tool usage reflected distinct verification strategies, instead of random switching (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, "AutoCheck and AIChat 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 that "after I see the chart, I can have 90% to make sure I'm having the correct answer. Then I will use AutoCheck to finish the rest 10%." Notably, P15 never used AI for verification, entirely relying on manual approach. A 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."

Importantly, these archetypes were not fixed traits. Users adapted their strategies based on system reliability and growing familiarity. For instance, P13 initially relied on an AI-first approach but pivoted to Chart Builder-dominant after Auto Check gave an incorrect verdict. By contrast, P24 began with manual exploration but gradually incorporated Auto Check as a background process as his comfort level increased. The sequential tool flow (Fig. 3b) confirms this fluidity. Auto Check dominated the initial verification step (65/131 sequences) but yielded to AI Chat and Chart Builder in Step 2–3. Table Explorer was used less frequently early on, 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. Just as P21 mentioned, "It's not abandoning, it's more about reinforcing the claim."

Claim complexity drives tool selection (8/22; Tab. 4). While workflow archetypes describe the macro-level sequencing of verification, participants actively tailored their specific tool choices to the analytical demands of each claim. Simple lookups (rankings, specific values) were typically directed to Table Explorer, while comparative claims involving averages or trends naturally mapped to Chart Builder. More

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
	T6	Claim complexity drives tool selection	8	P02, P04, P05, P08, P13, P17, P18, P23
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

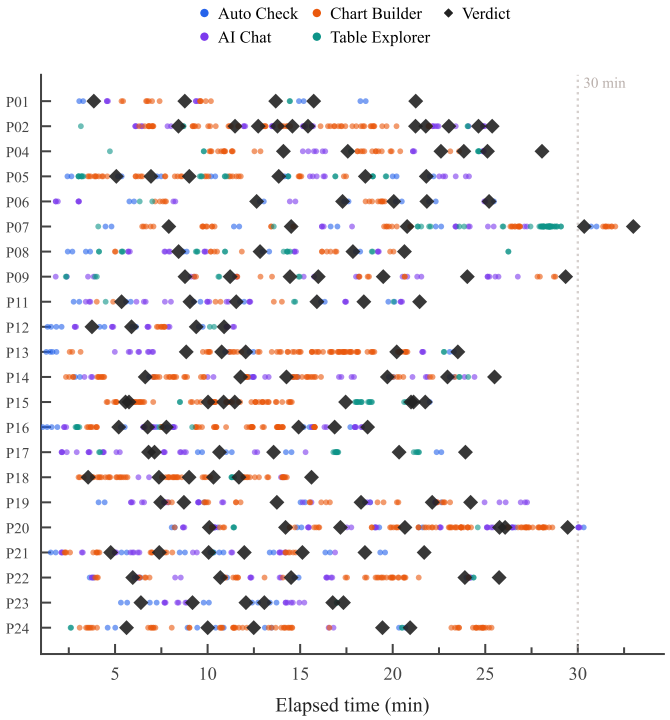

Fig. 2: Tool interaction timeline for each participant. Each dot represents one tool event; diamonds mark verdict submissions. Diverse temporal patterns emerge: some participants interleave tools throughout (P007, P020), while others cluster tool use in distinct phases (P12, P23).

complex statistical claims (e.g., correlations) frequently led to a hybrid approach. Users relied on AI Chat for the underlying computations, followed by Chart Builder for visual confirmation. Claims requiring external context beyond the provided dataset were either routed to AI Chat for web searches or simply deemed unverifiable. Notably, Table Explorer was rated least useful by nearly half of the participants (10/22). This was primarily due to its lack of aggregation capabilities, while "most of the claims involve aggregation" (P13). Thus, Table Explorer limits its role to initial schema exploration and simple lookups.

5.4 RQ3: How Do Users Reach Verdicts?

Participants reported consistently high confidence across all verdict types ($M = 6.31/7$, $SD = 1.20$), spending an average of 33.6 seconds ($SD = 15.2$) for each verdict. As shown in Fig. 3a, 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). Visual clarity from Chart Builder frequently served as the ultimate verification endpoint, with participants finding self-constructed charts inherently more convincing than AI-generated outputs. In 55 instances, users deliberately launched Chart Builder *after* Auto Check completed, using charts to audit AI conclusions. P02 noted, "If the chart already supported, or dismiss 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 entirely on AI tools they trusted less.

Multi-tool convergence (10/22; Tab. 4). Alternatively, participants ceased investigation when multiple tools aligned. P19 illustrated this layered confidence as "If both these agents give me 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?

Across the 88 claims where Auto Check produced a suggestion and the participant also submitted a verdict, the AI suggested 70 verified, 17 misleading, and 1 unverifiable. Participants' final verdicts on the same claims were broadly aligned (63 verified, 20 misleading, 5 unverifiable; Fig. 3a), suggesting that Auto Check's framing influenced decisions. However, participants overrode the AI toward unverifiable in 5 cases where it suggested only 1, indicating selective override when evidence was insufficient. Thematic analysis revealed three trust-related themes.

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 permanently abandoned Auto Check after an initial incorrect verdict. Interestingly, participants evaluated different output modalities independently, often trusting Auto Check's textual conclusions while disregarding its frequently generated broken or incorrectly scaled charts.

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. With the exposition of intermediate opera-

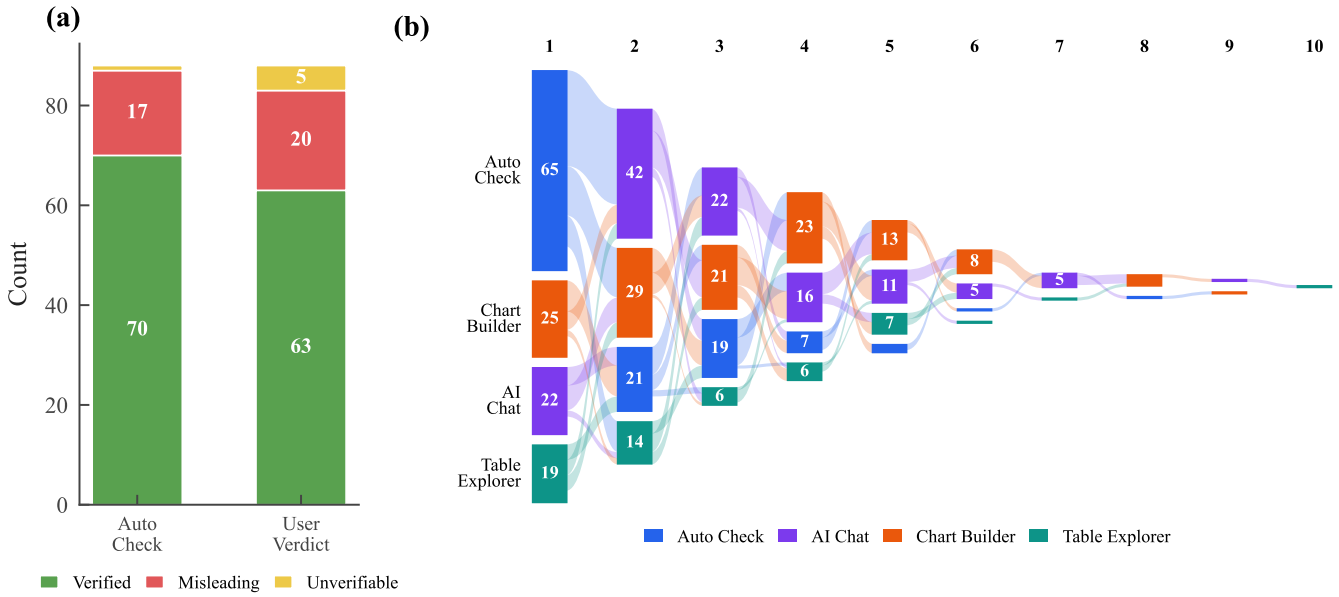

Fig. 3: AI agreement and sequential tool transitions. (a) Comparison of Auto Check 's suggested verdicts with participants' final verdicts on the same 88 claims where both exist. Close alignment for verified and misleading suggests anchoring; divergence in unverifiable indicates selective override. (b) Sankey diagram of sequential tool usage across 131 claim-investigation sequences. Auto Check dominates Step 1; AI Chat and Chart Builder rise from Step 2 onward; most sequences conclude by Step 5–6.

tions, the process visibility 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 human 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 varied significantly with reliance, which was heavily influenced by participants' professional backgrounds. 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 a highly dynamic process when multiple modalities are simultaneously available. 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

FYI was designed around a linear detection–verification–determination pipeline ($\mathbf{DG_1}$–$\mathbf{DG_3}$), yet participants rarely followed this strict sequence. In practice, the boundaries between detection and verification largely dissolved. Participants frequently discovered additional check-worthy claims *during* verification phase, making them 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 [47], demonstrating that in-situ fact-checking relies on fluid alternation between information foraging and synthesis rather than a rigid forward progression.

The design probe methodology was essential for capturing this complexity. Because FYI embedded all tools within the reading context ($\mathbf{DG_1}$), we could observe transitions that would be invisible in systems offering only a single verification modality. The interaction logs ($\mathbf{DG_3}$) enable us to observe that users actively resist prescribed procedural orders, while they construct distinct workflows dynamically shaped by the inherent complexity of 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 [24] 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 strongly argues 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 utilized visualization tools as an instrument 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 [60]. Trust built through cross-tool convergence, but 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.

While high overall confidence and close alignment with AI suggestions (Fig. 3a) raise valid concerns 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 construction too cognitively demanding were forced to rely on automated AI tools that they inherently trusted less. This highlights a significant vulnerability that a lack of visualization literacy directly compromises a user's ability to safely audit algorithmic outputs. Furthermore, we found that process visibility such as inspecting generated code acted as a powerful trust modulator, enabling more selective and appropriate 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 effective 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 provides a unique agency and interpretive confidence that AI-generated evidence cannot 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 observed three 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 specific complexity of the claim and personal expertise.

DI_4: Make AI reasoning and uncertainty visible. Because process transparency directly modulated trust calibration, future systems must move beyond opaque final outputs. Exposing 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.

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 ($M = 5.68/7$) and AI familiarity ($M = 6.18/7$), which may not represent casual readers or populations with lower data literacy. The observed workflow diversity may therefore underestimate the challenges faced by less experienced users.

Accuracy and ground truth. We embedded claims of known veracity in the article, but 75 additional user-detected claims were drawn from factually accurate content and were not designed to include errors. The 71.4% verified rate and high confidence ($M = 6.31/7$) could reflect genuine accuracy or systematic overconfidence—possibly amplified by anchoring on Auto Check's suggestions (Fig. 3a). Without ground-truth labels for all investigated claims, we cannot disentangle these possibilities. Future studies should embed claims with known ground truth at controlled difficulty levels.

System and model dependence. 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 [61] and hallucination [45] 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

We presented FYI, an in-situ browser extension, spanning 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). Using FYI as a design probe, we conducted exploratory study ($N = 22$) that revealed that fact-checking is not a linear pipeline but an iterative sensemaking process. Instead of passively accepting AI verdicts, participants constructed heterogeneous workflows, including AI-first with manual confirmation, manual-first with AI supplement, and parallel co-review. Visualization proved as a primary auditing mechanism, with users actively building charts to verify AI conclusions. Trust calibrated dynamically based on cross-tool convergence and process transparency. These findings indicate that future fact-checking environments should avoid rigid, single-tool designs. By treating AI as an initial guide, prioritizing visual authoring, and supporting adaptable workflows, systems can effectively balance automated efficiency with the depth of human judgment. We contribute FYI as open-source at https://github.com/DataVisards/FYI.

Supplemental Materials

FYI is available as open-source software at https://github.com/DataVisards/FYI. Supplemental material is available in PCS (the IEEE Xplore digital repository) and includes (1) a video demonstration of FYI outlining its claim detection, multi-tool verification, and verdict submission workflows, (2) the study article and movie dataset used in the evaluation, (3) all LLM prompts for claim detection, Auto Check, and AI Chat, (4) the interview protocol and questionnaire, and (5) the interaction log analysis scripts and figure generation code.

References

[1] 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, 2023. doi: 10.1038/s41586-023-06883-y

[2] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu, Q. Massonnat, A. Deiana et al. Statistical claim checking: StatCheck in action. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4798–4802, 2022. doi: 10.1145/3511808.3557198

[3] 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 Transactions on Visualization and Computer Graphics*, 29(1):1113–1123, 2022. doi: 10.1109/TVCG.2022.3209495

[4] I. Boldova and K. Božič Dimovski. Cognitive overload, anxiety, cognitive fatigue, avoidance behavior and data literacy in big data environments. *Information Processing & Management*, 61(1):103284, 2024. doi: 10.1016/j.ipm.2023.103284

[5] B. Botnevik, E. Sakariassen, and V. Setty. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 2117–2120, 2020. doi: 10.1145/3397271.3401396

[6] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa

[7] A. Buholayka and et al. Reference hallucination score for medical artificial intelligence chatbots. *JMIR Medical Informatics*, 12:e54345, 2024. doi: 10.2196/54345

[8] J. H. Chae and D. Tewksbury. Perceiving ai intervention does not compromise the persuasive effect of fact-checking. *New Media & Society*, 2024. doi: 10.1177/14614448241286881

[9] A. M. Chegini, K. Rezaei, H. Eghbalzadeh, and S. Feizi. RePanda: Pandas-powered tabular verification and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2503.11921.

[10] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[11] Z. Chen and H. Xia. CrossData: Leveraging text-data connections for authoring data documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI)*, art. no. 95, 2022. doi: 10.1145/3491102.3517485

[12] Z. Cheng, T. Xie, P. Shi, C. Li, R. Nadkarni, Y. Hu et al. Binding language models in symbolic languages. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[13] A. K. Das and K. Mueller. MisVisFix: An interactive dashboard for detecting, explaining, and correcting misleading visualizations using large language models. *IEEE Transactions on Visualization and Computer Graphics*, 32(1), 2025. doi: 10.1109/TVCG.2025.3633884

[14] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50):e2409662121, 2024. doi: 10.1073/pnas.2322823121

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

[16] 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 *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–20, 2024. doi: 10.1145/3654777.3676359

[17] K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012. doi: 10.1136/amiajnl-2011-000089

[18] L. Graves. Understanding the promise and limits of automated fact-checking. Factsheet, Reuters Institute for the Study of Journalism, University of Oxford, Feb. 2018. doi: 10.60625/risj-nqnx-bg89

[19] 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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4971–4983, 2022. doi: 10.18653/v1/2022.emnlp-main.331

[20] Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454

[21] A. Hanselowski and et al. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 6859–6866, 2018. doi: 10.1609/aaai.v32i1.11660

[22] N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1803–1812, 2017. doi: 10.1145/3097983.3098131

[23] T. L. Horstmann, B. Geisenberger, and M. Alam. T-REX: Table – refute or entail eXplainer. In *Machine Learning and Knowledge Discovery in Databases: Demo Track (ECML-PKDD)*, Lecture Notes in Computer Science, 2025.

[24] Y. Huang and A. Narechania. Facilitating proactive and reactive guidance for decision making on the web: A design probe with webseek. *arXiv preprint arXiv:2601.15100*, 2026.

[25] K. Hyland. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192, 2005. doi: 10.1177/1461445605050340

[26] F. Jahanbakhsh and D. R. Karger. A browser extension for in-place signaling and assessment of misinformation. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pp. 1–21, 2024. doi: 10.1145/3613904.3642473

[27] S. M. Jones-Jang and et al. How do people react to ai failure? automation bias, algorithmic aversion, and perceived agency controllability. *Journal of Computer-Mediated Communication*, 28(1):zmac029, 2023. doi: 10.1093/jcmc/zmac029

[28] P. Juneja and T. Mitra. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), art. no. 418, 2022. doi: 10.1145/3555143

[29] D. Karagiannis and et al. Identification and verification of simple claims about statistical reports. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13175–13182, 2020. doi: 10.1609/aaai.v34i05.6838

[30] G. Karagiannis, M. Saeed, P. Papotti, and I. Trummer. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *Proceedings of the VLDB Endowment*, 13(11):2085–2098, 2020. doi: 10.14778/3407790.3407841

[31] D. H. Kim, S. Choi, J. Kim, V. Setlur, and M. Agrawala. EmphasisChecker: A tool for guiding chart and caption emphasis. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):919–929, 2024. doi: 10.1109/TVCG.2023.3327150

[32] D. H. Kim, E. Hoque, J. Kim, and M. Agrawala. Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 423–434, 2018. doi: 10.1145/3242587.3242617

[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 *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 623–639, 2024. doi: 10.1145/3640543.3645155

[34] D. M. Lane and et al. The effect of cognitive load on decision making with graphically presented numerical information. *Human Factors*, 56(1):20–34, 2014. doi: 10.1177/0018720813491846

[35] 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 *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–22, 2025. doi: 10.1145/3706598.3713778

[36] G. Lim and S. T. Perrault. XAI in automated fact-checking? the benefits are modest and there's no one-explanation-fits-all. In *Proceedings of the 35th Australian Computer-Human Interaction Conference (OzCHI)*, pp. 323–336, 2023. doi: 10.1145/3638380.3638388

[37] C. Liu, Q. Zhou, X. Shen, X. B. Liu, T. Wu, and X. Chen. Behavioral indicators of overreliance during interaction with conversational language models. *arXiv preprint arXiv:2602.11567*, 2026.

[38] T. Lloyd, T. Nguyen, K. Levy, and M. Naaman. Beyond community notes: A framework for understanding and building crowdsourced context systems for social media. *arXiv preprint arXiv:2509.15434*, 2025.

[39] D. Metropolitansky and J. Larson. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. doi: 10.18653/v1/2025.acl-long.348

[40] 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 Transactions on Visualization and Computer Graphics*, 31(1):1235–1245, 2024. doi: 10.1109/TVCG.2024.3456144

[41] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, 2021. doi: 10.1109/TVCG.2020.3030378

[42] 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 *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 189–199, 2018. doi: 10.1145/3242587.3242666

[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. Accessed: 2026-03-31.

[44] OpenAI. GPT-4.1. Model card, 2025.

[45] A. Pesaranghader and E. Li. Hallucination detection and mitigation in large language models. *arXiv preprint arXiv:2601.09929*, 2026.

[46] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as inspired by a sensemaking model. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, pp. 376–380. SAGE Publications, 2005. doi: 10.1177/154193120504900357

[47] P. Pirolli and D. M. Russell. Introduction to this special issue on sensemaking. *Human–Computer Interaction*, 26(1-2):1–8, 2011. doi: 10.1080/07370024.2011.556557

[48] J. S. Rubinstein, D. E. Meyer, and J. E. Evans. Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance*, 27(4):763–797, 2001. doi: 10.1037/0096-1523.27.4.763

[49] 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. *arXiv preprint arXiv:2602.15865*, 2026.

[50] D. E. P. Schultz. *Truth goggles: automatic incorporation of context and primary source for a critical media experience*. PhD thesis, Massachusetts Institute of Technology, 2012.

[51] N. Sultanum and A. Srinivasan. DATATALES: Investigating the use of large language models for authoring data-driven articles. In *Proceedings of IEEE Visualization and Visual Analytics (VIS)*, pp. 231–235, 2023. doi: 10.1109/VIS54172.2023.00055

[52] M. Theologitis and D. Suciu. Thucy: An LLM-based multi-agent system for claim verification across relational databases. *arXiv preprint arXiv:2512.03278*, 2025.

[53] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074

[54] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan et al. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609

[55] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 422–426, 2017. doi: 10.18653/v1/P17-2067

[56] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[57] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. In *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 649–658, 2016. doi: 10.1109/TVCG.2015.2467191

[58] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand et al. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pp. 2648–2659, 2017. doi: 10.1145/3025453.3025768

[59] 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 *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 174–184, 2023. doi: 10.1145/3539618.3591708

[60] M. Yin and et al. Understanding and calibrating human reliance on ai for human-ai collaborative decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):194, 2019. doi: 10.1145/3359320

[61] Z. Yuan, H. Yuan, C. Tan, W. Wang, and S. Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

[62] R. Zamith. Transparency, interactivity, diversity, and information provenance in everyday data journalism. *Digital Journalism*, 2019. doi: 10.1080/21670811.2018.1554409

[63] Y. Zhang and et al. Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, 15(2):e0229132, 2020. doi: 10.1371/journal.pone.0229132