I'm Always a Little Skeptical of It: Verification Practices of Blind Users When Working with Generative AI in Spreadsheets

Minoli Perera
Department of Human-Centred Computing
Monash University
Melbourne, Australia
minoli.perera@monash.edu

Swamy Ananthanarayan
Department of Human-Centred Computing
Monash University
Melbourne, Australia
swamy.ananthanarayan@monash.edu

Cagatay Goncu
Department of Human-Centred Computing
Monash University
Melbourne, Australia
cagatay.goncu@monash.edu

Kim Marriott
Department of Human-Centred Computing
Monash University
Melbourne, Australia
kim.marriott@monash.edu

Abstract

Generative AI (GenAI) tools are increasingly used for spreadsheet tasks, yet little is known about how blind users verify their outputs in accuracy-critical contexts. We conducted a study with 12 blind spreadsheet users to explore verification practices across tasks such as information extraction, formula generation, trend analysis, chart creation, and formatting. Participants never fully trusted outputs without verification and employed diverse strategies, including manual checks with screen reader and spreadsheet features, same AI-assisted verification, cross-AI tool validation, leveraging prior knowledge, and human assistance. These approaches were adapted based on task context, perceived risk, and users' expertise. Errors were common, particularly in chart generation and formatting, some detected, others overlooked. While verification improved confidence, it was often effortful, time-consuming, or infeasible for visual tasks. We discuss how blind users utilize GenAI not only as a task performer but also as a verification aid and validator, highlighting design opportunities for more accessible and reliable spreadsheet use.

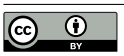
CCS Concepts

• **Human-centered computing → Empirical studies in accessibility**.

Keywords

blind, accessibility, spreadsheets, AI assistants, Generative AI, Copilot, Human-AI verification, assistive technology, screen readers

ACM Reference Format:
Minoli Perera, Swamy Ananthanarayan, Cagatay Goncu, and Kim Marriott. 2026. I'm Always a Little Skeptical of It: Verification Practices of Blind Users When Working with Generative AI in Spreadsheets. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3790988



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790988

1 Introduction

Spreadsheets are among the most widely used tools for tabular data manipulation [62] and remain central to data analysis [17]. Blind users typically access spreadsheets through screen readers (SRs), which read data one cell at a time. This linear access makes it difficult to gain an overview [21], identify patterns and trends [41, 64], and use analysis features such as formulas and conditional formatting [57]. Prior work also shows that SR users struggle with large spreadsheets and often take longer to complete tasks than sighted peers, largely due to accessibility limitations [61].

Generative AI (GenAI) tools powered by Large Language Models (LLMs) have shown potential to address such challenges by introducing an intent-based paradigm, enabling users to specify *what to do* rather than *how to do* it [52]. This shift has transformed workplace practices by offering significant productivity benefits [10, 23]. GenAI tools are now capable of analyzing data, generating summaries, extracting key insights and trends, creating formulas, applying formatting, and generating charts in spreadsheets [51, 53].

GenAI tools are also becoming increasingly integrated into accessibility workflows. They are available via web interfaces (e.g., ChatGPT [54]), screen reader–based tools (e.g., JAWS Picture Smart AI [58]), and mobile or desktop applications (e.g., Be My AI [24], Microsoft Copilot [50], Seeing AI [49]). Prior research has demonstrated their potential to enhance accessibility across a wide range of activities, including image descriptions [2, 4, 28, 29] and visualization [34, 59].

However, GenAI tools explicitly warn users that outputs may be inaccurate and should be checked for correctness. Hallucinations, false or misleading content generated by AI, is a well-known risk, and blind users often approach these tools with a mindset of everyday uncertainty [66]. Prior work has shown that blind people adopt a range of verification strategies when working with GenAI tools, including re-prompting, asking follow up questions, cross-checking with other models, using prior knowledge, consulting search engines, and seeking sighted assistance [2, 4, 13, 29]. Prior research has identified these strategies in the context of visual assistance for image-based tasks (e.g., understanding objects, surroundings [4, 29],

and visualizations [59]), document reading and writing [2, 66], web browsing [66], and productivity applications [56]. However, to our knowledge, there has been no step-by-step exploration of how blind users verify the accuracy of GenAI outputs when working with predominantly numeric datasets, particularly in spreadsheet tasks such as data analysis and manipulation, a domain where accuracy is often critical.

To address this gap, we conducted a remote study with 12 blind spreadsheet users to explore their verification practices and error-response strategies across common spreadsheet tasks, including information extraction, formula generation, chart creation, and formatting. Our primary contributions are:

- An empirical characterization of the verification methods blind users employ to ensure the accuracy of spreadsheet tasks, highlighting practices unique to them (e.g., verification using external AI-based image description tools).
- An analysis of verification and error-response workflows that reflect a skeptical stance toward GenAI, treating outputs as provisional and requiring stepwise validation.
- Design suggestions for GenAI-assisted spreadsheets that foreground multi-model validation, explainability, and prompt-engineering practices.

This study provides timely insights for assistive technology and GenAI tool designers highlighting the verification challenges and unique needs of blind users when working with GenAI. Although these tools continue to evolve rapidly, we believe our findings offer foundational knowledge of blind users' verification practices that remains relevant for designing accessible and reliable GenAI-assisted spreadsheet workflows.

2 Related Work

2.1 Accessibility of Spreadsheets for Blind Users

Spreadsheets achieved global adoption because of their flexible two-dimensional grid structure [12] and extensive built-in features. People who are blind typically access spreadsheets with screen readers (SRs), which read information cell by cell within the multi-dimensional grid. This linear presentation makes it difficult to form a holistic view of a sheet [41], locate specific components [21, 57], and interpret trends [64]. SR users also struggle to interpret information conveyed through visual cues, such as conditional formatting, icons, and data bars. These difficulties stem from cognitive overload, trade-offs between timing and information density, limited awareness of certain features, insufficient system feedback, and delayed or missing SR responses [57].

Research on spreadsheet accessibility has primarily focused on helping blind users interpret tabular (often numeric) data through sonification [64, 65] and understand embedded charts via sonification or haptic presentation [1, 21]. Prior work has also introduced a multi-modal presentation and navigation tool that provided overview information about complex spreadsheets and supported hierarchical speech-based navigation of tables, formulas, and charts [20, 21]. In addition, an answer-verification system was developed to help spreadsheet learners locate and understand errors, which blind and low-vision students found effective for independent study [37]. Despite these advances, a significant gap remains in understanding how GenAI tools could support SR users in data-analysis and spreadsheet-modification tasks.

2.2 Use of Generative AI in Spreadsheet Tasks

Recent research has shown that integrating GenAI in spreadsheets holds promise for enhancing productivity and automating tasks [9, 15, 46]. Modern models can now understand tabular data representations and are accessible through web-based chat interfaces (e.g., OpenAI ChatGPT [54], Google Gemini [30], and Anthropic Claude [16]) as well as embedded within applications like Microsoft Excel [51] (as Copilot [50]) and Google Sheets [31] (as Gemini [31]), enabling users to perform operations through natural language requests. Current systems support a range of data analysis and spreadsheet manipulation tasks, including formula generation, chart creation, information extraction, trend analysis, visual formatting and even code-based automation [40, 46]. Studies highlight persistent reliability problems across these activities, with formula generation often misinterpreting user intent and applying incorrect logic [47, 69].

To overcome these challenges, sighted users employ a range of verification strategies, such as cross-checking results with manual calculations or existing spreadsheets, checking code, comparing outputs across tools, or refining prompts until a satisfactory answer is reached [11, 35, 69]. In some cases, users even turn verification back onto the AI itself, asking it to explain or correct earlier responses [22]. While such practices help identify errors, they introduce significant overhead. Time spent validating results can equal or exceed the time saved through automation, and experienced users often find it faster to construct spreadsheets manually than to repair flawed outputs [11]. More broadly, these findings suggest that GenAI not only changes how tasks are performed but also shifts responsibility for correctness from the system onto the user. Furthermore, prior research has focused almost entirely on sighted users [35], and no studies have explored how blind users engage with GenAI in spreadsheets, even though these tools are central in professional and educational contexts.

2.3 Generative AI Use and Verification Strategies Among Blind Users

GenAI has emerged as a powerful assistive technology for people who are blind [56]. These tools are increasingly adopted across domains, including accessibility, and are available through screen readers (e.g., JAWS Picture Smart AI [58], NVDA AI-content-describer [67]), and mobile or desktop applications (e.g., Be My AI [24], Seeing AI [49], AIRA [3]). Prior studies show that GenAI can enhance accessibility across a wide range of activities, including image description [2, 4, 28, 29], visualization [34, 59], programming [25], content creation [7, 44], visual content generation [7, 42], understanding surroundings [73], and navigation [75]. Recent work has also highlighted the potential of GenAI to improve SR experiences in productivity-oriented tasks [27, 45, 56, 68].

Several studies have explore how blind users engage with GenAI tools [2, 4, 6, 29]. These studies report that blind users often develop flawed or overly simplified mental models of GenAI, shaped by their experiences of blindness, their reliance on assistive technologies,

and the accessibility of the tools [2]. GenAI tools have also been found to present accessibility and usability challenges for SR users [2, 5].

Moreover, GenAI frequently produces false or misleading content, commonly referred to as hallucinations [14, 63], which limits the reliability of these tools for accessibility tasks [2, 4, 39, 66]. Consequently, blind users often approach AI-generated output with skepticism [66] and employ a range of verification strategies, such as asking follow-up questions, cross-referencing across devices and applications, testing different models, comparing with known information, consulting search engines, using sensory cues, and seeking human confirmation [2, 4, 29, 66, 73, 74]. While these practices have been primarily found in object- and scene-based image description tasks [4, 13, 29], visualizations [59], document reading and writing [2, 66], web browsing [66], and productivity applications [56], there has been no exploration of verification strategies involving largely numeric data, particularly in spreadsheet analysis and manipulation. Our work takes an initial step toward understanding blind users' verification practices, error-response strategies, and the challenges they face in such workflows.

3 User Study

We conducted a remote study with 12 blind participants to explore verification and error-response strategies when analyzing data and modifying spreadsheets with GenAI assistance.

3.1 Study Material

We prepared two spreadsheet datasets, each consisting of a single workbook with one table (the original datasets are provided in the supplementary material):

- Inflation Data: This spreadsheet (Figure 1) contained real-world annual inflation data (%) from 2013 to 2023 for 162 countries. The dataset was obtained from the World Bank Group [36]. The original dataset was preprocessed to remove countries with missing values and modified to include a column specifying the region for each country.
- Student Marksheet: This spreadsheet (Figure 2) contained grades for 100 students in a single unit, including two assignments and two exams. The dataset was designed to reflect common real-world uses of spreadsheets, such as student grading, which has also been used in prior spreadsheet-related research [43].

3.2 Study Setup and Procedure

Before the study, all researchers familiarized themselves with how GenAI tools handle spreadsheet tasks such as generating overviews, extracting information, producing formulas and charts, and applying formatting. Prior research has identified these tasks as particularly challenging for blind users to complete using SRs [21, 41, 57]). The first author tested these tasks across a range of tools (ChatGPT [54], Gemini [30], Copilot [50], Claude [16], and DeepSeek [19]) and shared the outputs, which the remaining authors jointly reviewed to develop a shared understanding of their functionality. For chat-based tools, such as ChatGPT and Gemini on Web, spreadsheets (.xlsx files) had to be uploaded or copy-pasted into the interface before running a query. For modification tasks, some tools (e.g., ChatGPT) allowed the edited spreadsheets to be downloaded, while in Microsoft Excel [51] and Google Sheets [31], GenAI features were integrated directly, enabling real-time changes to the sheet.

Participants were free to choose any GenAI model or version they were most comfortable with. For those unfamiliar with the available model versions, we read aloud the model descriptions provided by the GenAI tool (e.g., ChatGPT-4o: *"great for most tasks"*; o4-mini: *"fastest at advanced reasoning"*). We asked them to select the model version they deemed most appropriate for spreadsheet tasks. To preserve privacy of their chat histories, we first asked participants which GenAI tools they typically used, then created dedicated accounts in those tools for the study. Participants logged in during each session, with model learning features disabled and all chat histories were deleted afterwards.

The study was conducted remotely via Zoom, allowing participants to use their personal computers and preferred SRs. Participants were asked to share their screens and enable audio recording (all participants had previously received project information and provided consent via email).

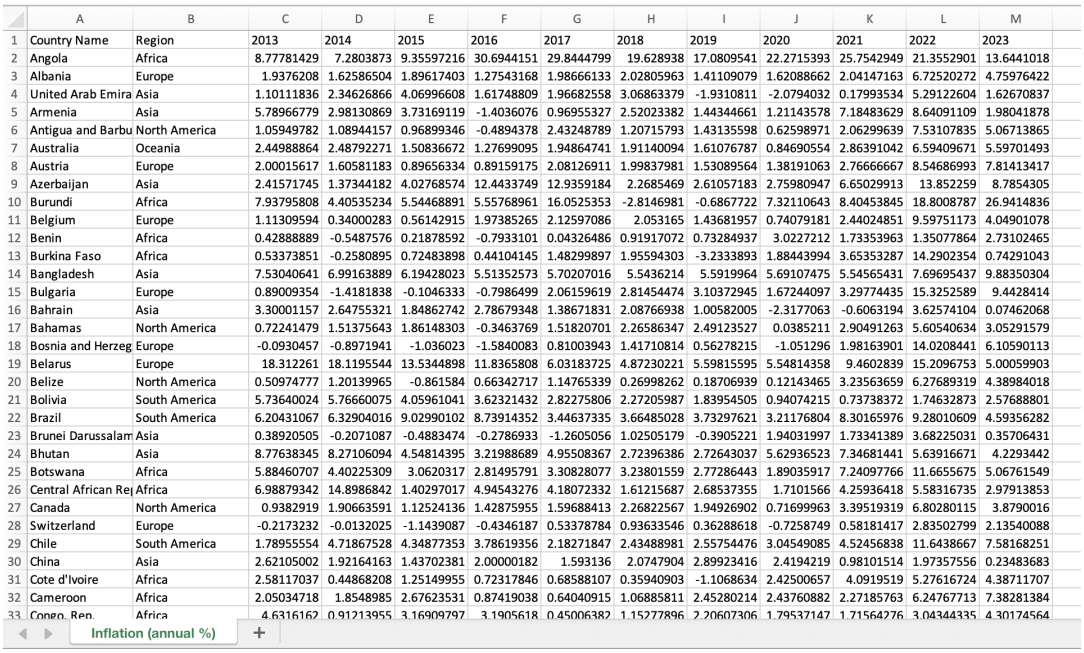
At the start of each session, we collected demographic information and asked participants about their prior experience with spreadsheets and GenAI tools, including how they typically approached trust and verification. Based on their responses, we provided a brief, high-level overview of other verification strategies that could be applied to spreadsheet scenarios. Specifically, we told participants: *"Some ways to verify AI-generated output include asking follow-up questions, cross-checking results with other AI tools, or manually checking the spreadsheet. There may be many other ways to verify the content, and you are free to use any method you prefer."* This brief overview was intended to give participants a general understanding of the types of verification strategies available, without offering detailed guidance on how to apply them. The goal was simply to ensure that all participants were aware of potential approaches they might not have previously encountered.

Participants were then assigned one of two accuracy-critical scenarios reflecting real-world spreadsheet use in workplace and educational contexts and incorporating the tasks identified above. One scenario related to analyzing data in spreadsheets, while the other involved modifying spreadsheet data. Each scenario included three tasks. (Although we initially planned for participants to complete both scenarios, two pilot studies revealed that this was infeasible within a single session). Participants were emailed the relevant spreadsheet and instructed to complete tasks using GenAI assistance, with accuracy emphasized as critical. Before starting the tasks, participants were asked to familiarize themselves with the spreadsheet, either by using GenAI assistance or by navigating the sheet.

The two scenarios and their associated tasks were as follows:

S1: Analyzing Inflation Data: Participants played the role of a finance reporter tasked with analyzing inflation data to write an article for a national newspaper covering global inflation from 2013–2023, with high expectations for accuracy.

- *S1T1*: Identify the country with the highest inflation rate in 2023.
- *S1T2*: Find out how the inflation rate changed over the years in Oceania.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country Name	Region	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2	Angola	Africa	8.77781429	7.2803873	9.35597216	30.6944151	29.8444799	19.628938	17.0809541	22.2715393	25.7542949	21.3552901	13.6441018
3	Albania	Europe	1.9376208	1.62586504	1.89617403	1.27543168	1.98666133	2.02805963	1.41109079	1.62088662	2.04147163	6.72520272	4.75976422
4	United Arab Emira	Asia	1.10111836	2.34626866	4.06996608	1.61748809	1.96682558	3.06863379	-1.9310811	-2.0794032	0.17993534	5.29122604	1.62670837
5	Armenia	Asia	5.78966779	2.98130869	3.73169119	-1.4036076	0.96955327	2.52023382	1.44344661	1.21143578	7.18483629	8.64091109	1.98041878
6	Antigua and Barbu	North America	1.05949782	1.08944157	0.96899346	-0.4894378	2.43248789	1.20715793	1.43135598	0.62598971	2.06299639	7.53107835	5.06713865
7	Australia	Oceania	2.44988864	2.48792271	1.50836672	1.27699095	1.94864741	1.91140094	1.61076787	0.84690554	2.86391042	6.59409671	5.59701493
8	Austria	Europe	2.00015617	1.60581183	0.89656334	0.89159175	2.08126911	1.99837981	1.53089564	1.38191063	2.76666667	8.54686993	7.81413417
9	Azerbaijan	Asia	2.41571745	1.37344182	4.02768574	12.4433749	12.9359184	2.2685469	2.61057183	2.75980947	6.65029913	13.852259	8.7854305
10	Burundi	Africa	7.93795808	4.40535234	5.54468891	5.55768961	16.0525353	-2.8146981	-0.6867722	7.32110643	8.40453845	18.8008787	26.9414836
11	Belgium	Europe	1.11309594	0.34000283	0.56142915	1.97385265	2.12597086	2.053165	1.43681957	0.74079181	2.44024851	9.59751173	4.04901078
12	Benin	Africa	0.42888889	-0.5487576	0.21878592	-0.7933101	0.04326486	0.91917072	0.73284937	3.0227212	1.73353963	1.35077864	2.73102465
13	Burkina Faso	Africa	0.53373851	-0.2580895	0.72483898	0.44104145	1.48299897	1.95594303	-3.2333893	1.88443994	3.65353287	14.2902354	0.74291043
14	Bangladesh	Asia	7.53040641	6.99163889	6.19428023	5.51352573	5.70207016	5.5436214	5.5919964	5.69107475	5.54565431	7.69695437	9.88350304
15	Bulgaria	Europe	0.89009354	-1.4181838	-0.1046333	-0.7986499	2.06159619	2.81454474	3.10372945	1.67244097	3.29774435	15.3252589	9.4428414
16	Bahrain	Asia	3.30001157	2.64755321	1.84862742	2.78679348	1.38671831	2.08766938	1.00582005	-2.3177063	-0.6063194	3.62574104	0.07462068
17	Bahamas	North America	0.72241479	1.51375643	1.86148303	-0.3463769	1.51820701	2.26586347	2.49123527	0.0385211	2.90491263	5.60540634	3.05291579
18	Bosnia and Herzeg	Europe	-0.0930457	-0.8971941	-1.036023	-1.5840083	0.81003943	1.41710814	0.56278215	-1.051296	1.98163901	14.0208441	6.10590113
19	Belarus	Europe	18.312261	18.1195544	13.5344898	11.8365808	6.03183725	4.87230221	5.59815595	5.54814358	9.4602839	15.2096753	5.00059903
20	Belize	North America	0.50974777	1.20139965	-0.861584	0.66342717	1.14765339	0.26998262	0.18706939	0.12143465	3.23563659	6.27689319	4.38984018
21	Bolivia	South America	5.73640024	5.76660075	4.05961041	3.62321432	2.82275806	2.27205987	1.83954505	0.94074215	0.73738372	1.74632873	2.57688801
22	Brazil	South America	6.20431067	6.32904016	9.02990102	8.73914352	3.44637335	3.66485028	3.73297621	3.21176804	8.30165976	9.28010609	4.59356282
23	Brunei Darussalam	Asia	0.38920505	-0.2071087	-0.4883474	-0.2786933	-1.2605056	1.02505179	-0.3905221	1.94031997	1.73341389	3.68225031	0.35706431
24	Bhutan	Asia	8.77638345	8.27106094	4.54814395	3.21988689	4.95508367	2.72396386	2.72643037	5.62936523	7.34681441	5.63916671	4.2293442
25	Botswana	Africa	5.88460707	4.40225309	3.0620317	2.81495791	3.30828077	3.23801559	2.77286443	1.89035917	7.24097766	11.6655675	5.06761549
26	Central African Re	Africa	6.98879342	14.8986842	1.40297017	4.94543276	4.18072332	1.61215687	2.68537355	1.7101566	4.25936418	5.58316735	2.97913853
27	Canada	North America	0.9382919	1.90663591	1.12524136	1.42875955	1.59688413	2.26822567	1.94926902	0.71699963	3.39519319	6.80280115	3.8790016
28	Switzerland	Europe	-0.2173232	-0.0132025	-1.1439087	-0.4346187	0.53378784	0.93633546	0.36288618	-0.7258749	0.58181417	2.83502799	2.13540088
29	Chile	South America	1.78955554	4.71867528	4.34877353	3.78619356	2.18271847	2.43488981	2.55754476	3.04549085	4.52456838	11.6438667	7.58168251
30	China	Asia	2.62105002	1.92164163	1.43702381	2.00000182	1.593136	2.0747904	2.89923416	2.4194219	0.98101514	1.97357556	0.23483683
31	Cote d'Ivoire	Africa	2.58117037	0.44868208	1.25149955	0.72317846	0.68588107	0.35940903	-1.1068634	2.42500657	4.0919519	5.27616724	4.38711707
32	Cameroon	Africa	2.05034718	1.8548985	2.67623531	0.87419038	0.64040915	1.06885811	2.45280214	2.43760882	2.27185763	6.24767713	7.38281384
33	Congo, Rep.	Africa	4.6316162	0.91213955	3.16909797	3.1905618	0.45006382	1.15277896	2.20607306	1.79537147	1.71564276	3.04344335	4.30174564

Inflation (annual %) +

Figure 1: Inflation rate dataset.

	A	B	C	D	E	F
1	Student ID	Student N	Assignment 1	Assignment 2	Exam 1	Exam 2
2	8054	George M	68	53	56	69
3	9625	Min Nair	54	41	31	40
4	7652	Nina Liu	78	71	80	85
5	2320	Omari Abe	77	67	55	55
6	8798	Emma Me	74	71	83	86
7	8109	Zuri Sharm	94	70	95	83
8	1697	Sophie Wa	87	90	93	91
9	1762	Olivia Bak	58	67	55	60
10	1539	George Pa	27	37	41	33
11	7525	Olivia Wo	88	86	76	90
12	7375	Imani Aziz	49	61	66	70
13	2284	Farid Redo	22	14	43	27
14	2245	Isha Wu	88	92	85	97
15	6347	Sarah Mer	54	55	85	66
16	6829	Wei Verm	55	71	68	71
17	7227	Liam Xu	57	78	54	67
18	2344	Hana Adar	66	68	63	47
19	7383	Aarav Des	44	33	38	17
20	8286	Neha Li	87	77	75	67
21	8353	Li Zhang	73	63	60	62
22	6798	Wei Jones	90	96	81	98
23	7311	Emily Choi	59	71	49	70
24	4977	Mei Husse	37	28	21	20
25	5405	Ren Li	62	60	63	71
26	5521	Mateo Wo	81	58	67	64
27	5040	Kwame Ba	90	77	85	74
28	3219	Ren Smith	55	71	64	57
29	6399	Anna Shar	76	76	77	70
30	3230	Ren Singh	75	94	78	87
31	6147	George Sc	44	58	57	55
32	9044	Mei Kim	55	76	62	61
33	7437	Neha Wal	73	85	84	72
34	1440	Leon Lewi	76	60	77	73

Marks +

Figure 2: Student marksheet (We deliberately adjusted the column spacing to truncate the data in the *Student Name* column (col B in the sheet) to create a task focused on formatting the table with GenAI assistance.)

- *S1T3*: Generate a chart showing how inflation changed over the years in Oceania compared to other regions.

S2: Modifying a Student Marksheet: Participants acted as university lecturers finalizing grades for a core subject essential for students' graduation. The spreadsheet included marks for assignments and exams, and the finalized grades would be submitted to the university's education board.

- *S2T1*: Add a column to calculate each student's average mark and assign letter grades (A = 90–100, B = 80–89, C = 70–79, D = 60–69, F = <60).
- *S2T2*: Highlight students who failed in red.
- *S2T3*: Improve presentation of the sheet by formatting the header row (background color and bold text), ensuring all data is visible, and making the sheet look visually appealing.

For each task, we asked participants to describe their experiences, their confidence in the accuracy of outputs, and their verification workflows, including why they chose particular strategies, whether these increased their confidence, and what challenges or suggestions they had.

After participants completed their usual verification process for the initial task, if they had not used the reasoning feature (which displays the model's internal thinking process) [32, 55], we asked whether they were aware of it. Those who were not were then asked to review the reasoning provided alongside the model's response (in tools where it was available) or to prompt the AI to explain its steps. This step was necessary because, at the time of the study, the reasoning feature was relatively new and available only in certain model versions. When accessed through a SR, this feature was often skipped entirely after response generation—either because the SR automatically began reading the response immediately (e.g., ChatGPT) or because the difficulties of finding the *show thinking* button in the interface. We included this step to ensure that all participants were aware and up to date with the tools' existing capabilities, which could support their verification process but might not have been known due to the feature's novelty, limited exposure, or accessibility barriers. Once participants were made aware of the feature, we did not intervene further.

After completing all tasks, participants were asked about their overall experiences with verification, challenges encountered, and suggestions for improving verification workflows. Each study session lasted approximately two hours, and participants received a $100 Amazon gift card as compensation. The study was approved by our institutional ethics committee.

3.3 Participants

Participants were recruited via email from a pool of individuals who had taken part in prior studies, where their proficiency with spreadsheets had already been demonstrated. They were subsequently invited through a call-for-participation email. Inclusion criteria required participants to be (1) blind adults (aged 18 and over) who regularly use Excel with a SR and (2) individuals with prior experience using any GenAI tool, such as ChatGPT or Copilot.

We recruited 12 blind participants (11 totally blind, 1 legally blind; 9 male, 3 female) who relied on a SR for computer access. Seven participants were from the US, three from Australia, and one each from Canada and India. Ages ranged from 18 to over 65. A summary of participant demographics is provided in Table 1.

More than half of the participants (n=8) reported daily use of spreadsheets, while four used them weekly. Nine participants described their proficiency level with spreadsheets as intermediate, two as advanced, and one as beginner. Participants used spreadsheets for a variety of purposes in both work and personal contexts, including recording financial data and budgeting (n=9), maintaining lists (n=7), analyzing data (n=3), and tracking work-related tasks (n=2).

All participants had prior experience with multiple GenAI tools across diverse tasks. Beyond using them as chatbots for information retrieval, participants reported employing GenAI for image descriptions (n=9), coding (n=6), research (n=3), search (n=2), travel planning (n=2), writing (n=2), describing slides (n=2), generating charts (n=2), document generation (n=2), and obtaining real-time descriptions of their surroundings (n=2). With respect to spreadsheet-related GenAI use, four participants (P1, P4, P5, P12) had prior experience generating formulas, P10 had used GenAI to create tables, and P5 had used it to generate pivot tables. Other participants reported no prior experience using GenAI for spreadsheet tasks. During the study, eight participants used ChatGPT, two used Copilot, and two used Gemini.

Participants P1–P6 were assigned the inflation data analysis scenario (S1), while participants P7–P12 worked on the student marksheet modification scenario (S2).

3.4 Data Analysis

We conducted two separate analyses corresponding to the two primary data sources: (1) screen recordings of participants performing the tasks and (2) qualitative feedback provided during and after the tasks.

The objective of the observational analysis was to capture participants' verification strategies, their responses when inconsistencies were encountered, their use of different features, and the challenges faced during verification and while using GenAI tools. Initial codes were generated from reviewing the recordings and were documented for each participant and task. These codes were then organized into categories, including understanding data, verification methods, and verification challenges.

All qualitative feedback was transcribed and analyzed using inductive thematic analysis [8]. Our focus was on identifying verification strategies, participants' rationales for adopting them, the challenges they encountered and their suggestions to improve the verification process. Two authors independently coded one transcript, after which all authors met to discuss the initial codes and resolve discrepancies, resulting in the refinement of codes. The first author then analyzed the remaining transcripts, which were cross-checked by the other authors. The entire team engaged in follow-up discussions to refine and build themes, helping to mitigate individual bias. Low-level codes such as *"verifying using SR features"* and *"verifying using spreadsheet features"* were grouped under the higher-level code *manual verification*, which was then organized under the broader category of *verification methods* (Section 4.2). Finally, these were cross-validated against the observational data from screen recordings, and vice versa.

Table 1: Details of study participants. Proficiency levels are self-reported.

P#	Age	Country	Occupation	Spreadsheet Proficiency	GenAI tools used	Freq. of using GenAI	SR used	GenAI tools used during study
P1	26–35	US	Senior applied scientist	Intermediate	Gemini, Copilot in Teams, PowerPoint, Word, Excel, GitHub Copilot, Meta AI glasses	Daily	JAWS	Gemini Pro
P2	46–55	Canada	UX accessibility consultant	Intermediate	ChatGPT, Copilot (Web, M365), Gemini, Seeing AI, JAWS Picture Smart AI, Meta AI glasses	Daily	JAWS	Gemini Pro, ChatGPT 4o
P3	65+	Australia	Accessibility consultant	Intermediate	ChatGPT, Perplexity, Copilot (Web, M365), Seeing AI, Be My AI, NVDA AI Content Describer	Several times a week	NVDA	ChatGPT o4mini
P4	36–45	US	Sr. director of software engineering	Intermediate	Gemini, Perplexity, Claude, ChatGPT, Be My AI, Seeing AI	Daily	JAWS	ChatGPT 4o
P5	26–35	US	Senior software engineer	Advanced	Copilot in Excel, GitHub Copilot, ChatGPT, Meta AI glasses, Seeing AI, JAWS Picture Smart AI	Daily	JAWS	Copilot (Excel, Web)
P6	46–55	Australia	Academic	Intermediate	Copilot on Edge, ChatGPT (mobile), Be My AI, JAWS Picture Smart AI	Daily	JAWS	Copilot on Edge
P7	26–35	US	Assistive tech trainer	Beginner	ChatGPT, Gemini, Ollama (local), Kagi, Be My AI, AIRA, OurAccessAI, Pixxy Bot, Seeing AI, NVDA AI Content Describer	Daily	NVDA	ChatGPT o4mini
P8	26–35	Australia	Public service planner	Intermediate	Gemini, Be My AI, ChatGPT, JAWS FS Companion	Daily	JAWS	ChatGPT o4mini
P9	56–65	US	Cyber security student	Intermediate	ChatGPT (browser), Be My AI, JAWS Picture Smart AI, Meta AI glasses	Daily	JAWS	ChatGPT 4o
P10	18–25	India	Master's student	Intermediate	ChatGPT, Gemini, Be My AI	Daily	NVDA	ChatGPT 4.1
P11	46–55	US	Software engineer	Intermediate	GitHub Copilot, M365 Copilot, Copilot on Edge, ChatGPT, JAWS Picture Smart AI	Daily	JAWS	Copilot (Edge, Excel)
P12	56–65	US	Consultant	Advanced	ChatGPT, Gemini, Copilot (Web, M365), Ollama (local), GitHub Copilot, JAWS Picture Smart AI, Be My AI, Perplexity	Daily	JAWS	ChatGPT 4o

4 Findings

Participants perceived GenAI as a valuable tool and a useful starting point for performing spreadsheet-related tasks. At the same time, they reported accessibility and usability challenges [2] that influenced how effectively they could integrate these tools into their workflows. Common issues included the absence of progress updates during response generation (n=4), difficulties navigating chat interfaces and locating elements such as charts or tables in responses (n=4), and instances where content was inaccessible to SRs (n=2). While these accessibility considerations shaped participants' interactions, the primary focus of this study was to understand how blind users conduct and verify GenAI-assisted spreadsheet tasks. We organize our findings into three sections: first, we describe participants' overall levels of trust in GenAI tools; second, we explore their verification practices, focusing on how they identified errors in AI outputs; and finally, we analyze their verification workflows, how they responded to inaccuracies and the strategies they employed to address them.

4.1 Trust in GenAI tools

All participants were experienced users of GenAI tools and were fully aware of common limitations such as inaccuracies and hallucinations. This awareness often led them to express skepticism toward the outputs. Participants' level of trust in GenAI was frequently shaped by their prior experiences and the type of task at hand. Tasks involving structured or numeric data tended to elicit higher levels of confidence, whereas tasks requiring more open interpretations, such as image generation or visually descriptive tasks, were met with greater skepticism. Participants (n=3) also reported using different tools for different tasks, drawing on their experiences of how well each tool performed in particular contexts, reflecting tool-switching practices observed in earlier studies [2, 4]. Beyond task type and experience, additional factors influenced trust: for example, some participants (n=3) reported placing greater confidence in outputs based on the reputation of the company behind the model. Interestingly, one participant highlighted how the accessibility and usability of the tool shaped their willingness to accept AI-generated outputs. As P7 explained: *"When the interface is more accessible, it makes it easier to agree with the data, with the analysis, stuff like that. And it's definitely a bias. But when things are more accessible, less stressful, then sometimes I might agree, even when the data is wrong."* However, all participants emphasized the importance of incorporating some level of verification when working with GenAI tools, particularly in accuracy-critical scenarios.

4.2 Verification Methods Used by Blind Users

Although participants described different factors shaping their trust in GenAI, all emphasized the critical importance of accuracy in the study scenarios. As a result, no participant skipped verification entirely; each either employed at least one method or described how they would verify AI outputs.

Participants used a variety of verification methods to assess AI-generated outputs, often combining multiple strategies depending on the task type, accessibility constraints, and the level of perceived risk. Table 2[1] summarizes these primary methods, their associated sub-methods, and support methods (secondary actions that helped participants carry out or strengthen their main verification strategies.) Five main categories of verification approaches were identified:

4.2.1 Manual verification: Most participants relied extensively on manual verification, returning to the dataset (the original dataset in S1 and the GenAI-updated spreadsheets in S2) and using both SR and Excel features to ensure the accuracy of AI-generated outputs. These approaches varied in intensity, ranging from targeted spot-checking to exhaustive reviews, depending on the task.

[1]The table illustrates how many participants used or indicated they would use each method during the study. For instance, although participants did not seek human assistance during the sessions, several noted that they would do so in real-world scenarios.

Table 2: Verification methods, method types, and participant usage across tasks.

Verification Method			Finding max value (S1T1)	Trend analysis (S1T2)	Chart creation (S1T3)	Formula generation (S2T1)	Conditional formatting (S2T2)	Visual appeal (S2T3)
Manual	Sub Method	Check response against data with SR	5	2	–	–	–	–
		Check generated spreadsheet with SR	–	–	–	6	6	6
		Human calculation	–	–	-	1	–	–
		Use Excel functions	3	4	–	5	2	1
	Support Method	Restart SR	–	–	–	–	1	–
		Change SR	–	–	–	–	1	1
Same AI	Sub Method	Ask for a subset of data	1	–	–	–	–	–
		Ask further questions	2	–	1	1	2	2
		Ask to check the response against the data in a new chat	–	1	–	–	–	–
		Ask to check the generated image/sheet in a new chat	–	–	2	–	1	–
		Check reasoning/code	–	4	2	–	1	1
	Support Method	Ask for instructions	–	2	–	–	3	1
		Ask for the response to be accessible	1	–	2	1	–	–
		Ask to regenerate the response	–	–	1	–	–	–
		Ask to retain formulas in the sheet	–	–	–	3	–	–
		Ask to use Excel functions	1	–	–	–	–	–
Different AI	Sub Method	Go through image descriptions	–	–	5	–	2	4
		Ask further questions	–	–	3	–	2	4
		Ask to check response against data	–	1	–	–	–	–
		Ask to check generated sheet	–	–	–	–	1	–
Human			–	–	6	–	2	5
Prior knowledge about data			1	2	–	–	–	–

For example, in the task, where participants were asked to highlight students who had failed in red (S2T2), all participants employed a spot-checking strategy. They returned to the sheet and used their SR features to examine both students with grades of F and those without. By interrogating formatting details such as font and background color (using Insert+F keys), they verified whether the cells had been accurately color coded. As P11 explained: *"I checked one person with an F, and I check one person without an F, and so it seems to have gotten that logic right. My confidence level at that point is enough to where I feel comfortable not checking everything else."* However, for tasks such as writing an article, most participants emphasized that they would manually fact-check all AI-generated content before publishing it. In one case, P8, while verifying average marks (S2T1), mentally estimated the values to check whether the AI's calculated average appeared reasonable.

While these manual approaches often increased participants' confidence in the output and were regarded as the most trustworthy verification method, they noted that manual verification required considerable effort and was frequently described as time-consuming and cognitively demanding, particularly when it involved writing more complex formulas for tasks such as analyzing trends (S1T2). In some cases, manual verification was impossible–for instance, when verifying AI-generated charts (S1T3) or assessing the overall visual appearance of spreadsheets (S2T3).

4.2.2 Same AI-assisted Verification: In addition to manual checks, participants leveraged the same AI tool as a resource for verification. This AI-assisted verification method involved testing the model's internal consistency and checking its intermediate steps and reasoning.

Some participants (n=3) verified outputs by examining whether the AI remained consistent under follow-up questioning. They asked clarifying questions to confirm its output [2, 66]. As P5 explained, *"I ask clarifying questions just to see if it's going to stick to what it said or change its answer. When it contradicts itself, it's kind of a red flag."* Similarly, after asking the AI to apply visual formatting to the spreadsheet (S2T3), P8 followed up with prompts such as *'Do you think this looks okay? Is it easy to read visually?'* to further verify the quality of the output.

Participants adopted prompting strategies to obtain customized responses that better supported their initial verification workflows. They employed techniques such as explicitly revealing their disability [66] (e.g., *"I am a blind user"*), requesting descriptions *"for a blind person/screen reader user"* or mentioning the type of the SR to make outputs more accessible. These strategies were applied in tasks such as obtaining spreadsheet overviews (n=4), generating image descriptions (n=2), and requesting instructions to manually perform tasks (n=2). By tailoring prompts in this way, participants were able to access more usable information and begin verification within the model's output.

One participant also prompted the AI to generate a subset of the underlying data alongside the main response, enabling them to cross-check results within the same output. For example, P1 asked, *'Tell me inflation rates for all countries in 2023 and give me the highest and lowest from that'* which allowed them to verify the named highest country against the full list within the response itself, reducing the burden of navigating the original dataset. P1 reflected, *"I'd say maybe 70% confident because I made it generate the actual list and not just an answer, because sometimes it doesn't have all the data in context and gives a wrong answer."* P1 further re-asked the same question in different ways, such as *'So, in conclusion, what was the highest inflation rate in 2023?'* to see whether the model converged on the same result. Although this iterative questioning

was a lightweight way to surface discrepancies, they also noted it could reinforce errors if the model repeated the same mistake.

To avoid anchoring on a single conversation, some participants opened a new chat session in the same model [2] and questioned the prior output there. For instance, P10 in S2T2 uploaded the AI-modified spreadsheet (after requesting color-coding by condition) to a new session, believing that the new session did not retain memory of previous interactions, and asked whether the file actually contained any highlighting. P1 also opened a new session in the same model to critique the output from a previous run of S1T2, which analyzed trend fluctuations for Oceania. They uploaded the relevant spreadsheet and pasted the previously generated paragraph, instructing the model to act as a senior editor and fact-check every claim against the attached sheet—prompting it to *'list each fact with a judgment of accurate/inaccurate.'* While no inaccuracies were identified, the participant noted that they would still manually spot-check some claims, particularly the first and last, before publishing, explaining that the model *'sometimes goes on a tangent midway'*.

Participants had varying opinions about including a confidence rating alongside the AI output (similar to the findings from [4]). Some (n=3) found it useful for deciding their next steps in the verification process, as it provided a rough sense of accuracy. Others (n=6), however, felt it would not be useful and would instead become another datapoint to navigate. They argued that since such ratings were generated by the same model, they could be biased and equally prone to hallucinations.

Participants verified results by examining how the AI arrived at them. Some requested intermediate calculations, step-by-step explanations, or the code used to produce an answer. For example, in S2T1, P9 explicitly asked for the formula the AI had used to calculate average marks. Three participants also requested that formulas be retained in the sheet so they could access them with their SRs. For P11, who used Copilot in Excel, the formulas used for calculations were included directly within the response along with explanations, which helped build confidence in the accuracy of the output. During the study, one participant (P1) on their own initiative examined the model's step-by-step reasoning. By reviewing the chain-of-thought process, P1 was able to better understand the model's inner workings and confirm whether the steps had been carried out accurately.

Most participants (n=9), however, were initially unfamiliar with chain-of-thought reasoning. Once exposed to this method, two participants began incorporating it into their workflow for subsequent tasks. Overall, ten participants reported that access to reasoning information increased their confidence, particularly when they had doubts about an output. As P8 reflected after trying this approach for the first time: *"I really like that[thought process] because I've been kind of questioning 'Well, what if? what if it hasn't understood me?' Or 'what if it's not taking into account all the data?', all these sort of questions that have been hanging over my head, I can kind of see Oh, yeah, it's got this. It knows what I'm asking and it's thought about how to present the answers in an accessible way."*

However, some of the reasoning provided was not to our participants expectations. For example, in S2T3 (applying visual formatting to the table), P7 reviewed the reasoning and noted that it was not useful, as it failed to capture specific information about the actions they had performed in the sheet—such as which colors were applied or how the text was aligned. Similarly, participants emphasized the limitations of this approach. They noted that while reasoning served as a useful starting point for verification, it was insufficient to ensure full confidence.

Reviewing the underlying code was perceived as a more trustworthy verification method. Two participants (P1, P5) with prior coding knowledge found that accessing the code behind responses significantly strengthened their trust. P1 remarked, *"I think my confidence went up to almost 100% reading the code."* Both P1 and P5 asked the AI to include code in its responses so they could verify outputs directly. However, two participants who encountered code embedded within reasoning found it confusing, as they were unfamiliar with the syntax.

The same AI tool also acted as a support mechanism for manual verification. In cases where participants lacked the necessary knowledge, several (n=5) used the AI as a verification guide, prompting it for step-by-step instructions on how to independently check the accuracy of outputs. For example, participants asked questions such as *"As a blind person, how can I confirm the red color of the rows?"* and *"What is the formula for calculating the mean of a column of numbers in Excel?"*. Participants also turned to the same AI when they wanted the content in a more accessible format (e.g., requesting chart data points in a table), when they were unable to locate certain elements such as charts in the response (e.g., P1 in S1T3, asking to regenerate the response), or when they needed help performing spreadsheet functions as part of verification (e.g., P5 in S1T1, asking the AI to sort a column within the spreadsheet).

4.2.3 Different AI-assisted Verification: Beyond manual and same AI verification, several participants (n=9) turned to other GenAI tools in their workflows, viewing them as a more independent form of verification. This approach involved prompting other AI models to re-check information or comparing outputs across different GenAI tools—a method also reported in prior work [2, 13, 66]. When different models converged on the same answer, participants reported greater confidence in the accuracy of the output. For example, P2 verified trend information for Oceania in task S1T2, including average percentage values and a summary of fluctuations, by consulting ChatGPT after first working with Gemini. ChatGPT confirmed the content as accurate and reported no discrepancies, which gave P2 sufficient confidence to include the information in the article.

The most common use of additional AI tools was for verifying visual tasks such as chart generation (S1T3) and formatting (S2T2, S2T3). Participants drew on a range of visual assistance tools for this purpose, including JAWS Picture Smart AI (n=4), Be My AI (n=2), NVDA AI Content Describer (n=2), Seeing AI (n=1), and Copilot Vision (n=1). Most participants (n=8) used these tools to obtain descriptions of screen content (e.g., to generate descriptions of AI created charts), while some (n=5) went further by asking follow-up questions to confirm whether tasks had been completed accurately. For example, two participants (P7, P11) asked whether tables were formatted correctly with appropriate column spacing and alignment; P11, for instance, queried Picture Smart AI, *'Is the content from all 8 columns visible?'*

Participants also used other GenAI tools to clarify inconsistencies identified during manual verification. For example, when Copilot stated "Here is the chart..." without actually producing one, P5 and P6 navigated the response and then asked Seeing AI and Be My AI, respectively, whether a chart was present on the screen. Similarly, after detecting inconsistencies in conditionally formatted files generated by ChatGPT, P10 turned to Be My AI to confirm whether students who failed were highlighted. Some participants utilized multiple-model (Claude and ChatGPT) description features integrated into visual assistance tools such as Picture Smart AI for cross-checking. For example, P6 compared chart descriptions, while P12 compared conditional formatting outputs. P6 suggested that cross-model verification could be far more efficient if automated: *"It would be lovely if it... already did the verification, using other models or to say we found discrepancies, checking this with another model would have automated my verification process."*

While most participants recognized the benefits of different AI tools for verifying outputs, some (P6, P8, P10, P11, P12) reported confusion about which tool to trust when discrepancies arose. For example, when P11 asked Picture Smart AI whether all content in a table was visible, the tool reported that some cells in a column were cut off. P11 described feeling uncertain about how to act on this information: *"I would not make any changes based on that. Like if it had told me 'Okay, it's all visible'... then I would have been like, 'All right, that's fine.'–that's as good as I'm going to be able to do. Now that it tells me that it's not quite visible, like I don't know, I don't know what to do."* Although in this case the tool provided an accurate response, there were also instances where these tools hallucinated and further confused participants. For example, (P10, P12) received incorrect information about color-coded cells in the student mark-sheet spreadsheet (S2T2), and P10 was also given inaccurate descriptions of visual formatting, such as text alignment and cell borders (S2T3). At the same time, some participants (P4, P8, P11) expressed skepticism about relying on other GenAI tools, emphasizing that errors could propagate across multiple models. P4 explained, *"I mean, they could all be wrong. You would have to use at a bare minimum at least three and get two of them to agree. But I don't know that I would trust AI if I tried them on three models, and only in two of them got it correct. They could have just both been wrong in the same way."* These uncertainties prompted participants to seek human verification for tasks that relied heavily on visual presentation, such as chart generation and formatting.

Apart from inaccuracies, participants also encountered other issues with different GenAI tools that made them less useful in certain scenarios, leading them to seek alternative verification methods. For example, when P2 (in S1T3) attempted to take a screenshot of the generated chart to be read by Picture Smart AI, the ChatGPT prompt input box obstructed part of the chart, preventing a complete view. As a result, P2 reverted to the same AI tool and asked follow-up questions for verification. In S2T3, P8 raised similar concerns, noting that such tools have inherent limitations because screenshots may not capture the entire screen for description. As P8 explained, *"I think, because there are a hundred rows, I don't actually know whether all 100 rows fit on the screen at any given time. So I feel like if I were to ask Be My AI, I don't know if it would be able to see everything anyway."*

4.2.4 Sighted Human Verification: When verification methods or strategies were ineffective and they were unsure how to verify results independently, participants turned to sighted assistance as a reliable way to validate AI-generated outputs (similar to findings from [2, 4, 29, 66, 73]). This was particularly common for tasks that relied heavily on visual presentation, such as graph generation (S1T3) and formatting (S2T2, S2T3). All participants except P9 reported that they would seek sighted assistance in such cases; P9, however, expressed sufficient confidence in working with AI alone and did not feel the need for sighted support. Participants (n=3) emphasized that GenAI tools were often unable to provide accurate subjective feedback on visual qualities, noting that *"AI is not there yet"*, which made sighted verification necessary.

Participants (n=5) mentioned turning to trusted sources, such as colleagues, friends, or remote agents such as Be my Eyes or AIRA (n=3). They sought feedback on whether a chart had been generated correctly with appropriate sizing and colors, whether overlaps with other elements occurred, whether formatting such as column spacing and alignment was adequate, whether colors had sufficient contrast, and whether the output *'looked good'* overall. P5 explained that they would first ask and confirm with the AI before turning to a sighted person for verification: *"I would ask the AI first, just to not waste someone's time. But if the AI feels confident, and I asked a few questions about the chart, then I just check with someone."*

While sighted verification provided reassurance, some participants reflected on how it affected their sense of independence. For example, P6, after identifying inconsistencies in chart descriptions using Picture Smart AI's multi-model description feature, explained that relying on human assistance felt like a last resort—necessary, but more time-consuming and costly: *"I want to be independent, and I probably would have accepted it if it made sense, if they[Claude and ChatGPT] were correlating to what I expected. But they are not, and therefore now I'm going to my human intervention, which is more time consuming, and there's a cost involved, because I can't independently do it."*

4.2.5 Verification through Contextual Knowledge. In addition to manual, AI-assisted, and sighted verification, participants also drew on their prior knowledge and responses from previous tasks to verify outputs to some extent. For example, when answering S1T1 (identifying the country with the highest inflation in 2023), P3 recalled seeing the same country flagged earlier as an outlier when requesting an overview of the dataset. The repetition of this result across tasks gave them some confidence in its accuracy. Similarly, two participants partially verified trend fluctuations (S1T2) by noting that a spike in inflation during specific years aligned with their understanding of the pandemic's impact. However, no participant relied solely on prior knowledge for verification; all combined this strategy with other methods to confirm the information.

Participants employed a diverse set of strategies to verify accuracy across tasks. These strategies varied with task type, context, perceived importance, and participants' familiarity with different methods. In some cases, verification relied on a single approach, while in others participants combined multiple strategies, especially when they detected inconsistencies.

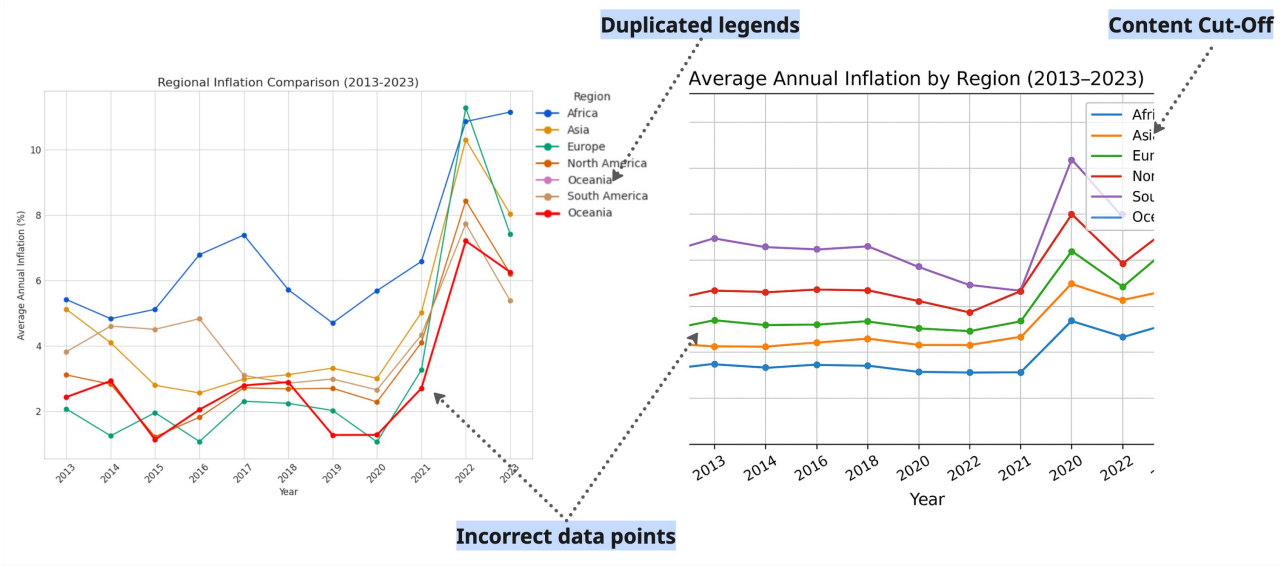


Figure 3: Chart generation errors.

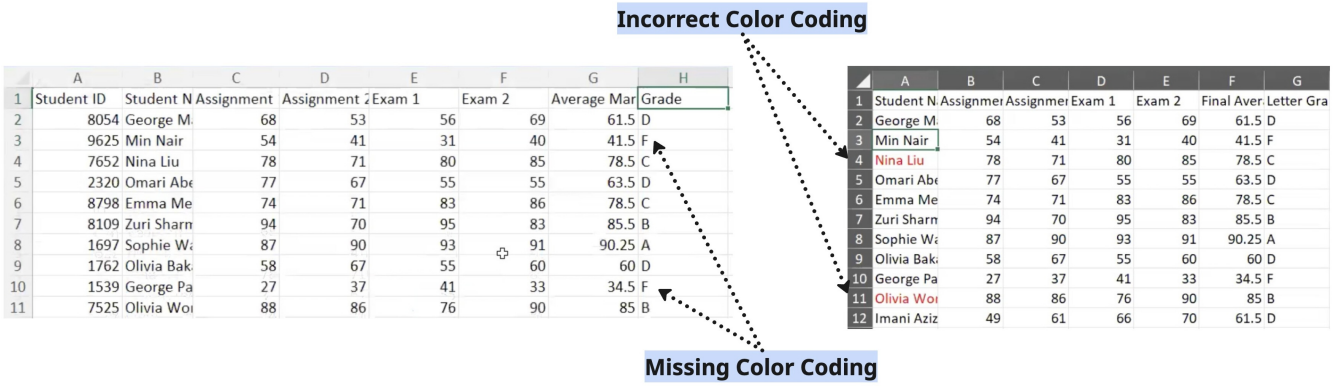


Figure 4: Conditional formatting errors.

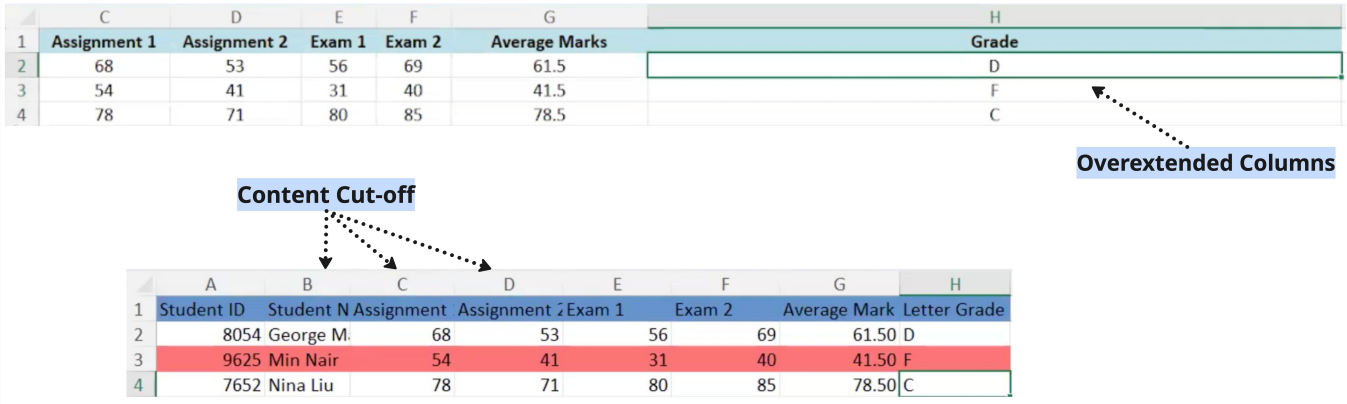


Figure 5: Visual formatting errors.

Table 3: Errors observed across tasks during the study and the participants' corresponding identification outcomes.

Task	Error	P#	Error identified?
S1: Overview	Missing information: regions listed as Africa, Europe, Asia, North America; missing Oceania, South America	P2	Unnoticed
S1T1: Finding the maximum value	Inaccurate information: country with highest inflation listed as Venezuela	P4	Identified
S1T3: Chart generation (Figure 3)	Missing chart	P5	Identified
		P6	Identified
	Incorrect datapoints in chart	P1	Noticed inconsistency; error not identified
		P5	Unnoticed
		P6	Unnoticed
	Duplicated legends	P1	Identified
		P6	Identified
	Chart content cut-off	P5	Unnoticed
	Inaccurate information: incorrect region-wise average inflation rate (%) values	P3	Unnoticed
S2: Overview	Inaccurate information: incorrect minimum values for assignments and exams	P9	Unnoticed
S2T2: Applying conditional formatting (Figure 4)	Missing color coding	P10	Noticed inconsistency; unsure whether there is an error
	Incorrect color coding	P12	Identified
S2T3: Applying Visual formatting (Figure 5)	Overextended columns	P7	Identified
		P10	Identified
		P12	Unnoticed
	Incorrect column content visibility	P11	Noticed inconsistency; unsure whether there is an error

4.3 Verification Workflows and Error Response Strategies

Participants' verification practices were often multi-step rather than linear. The user flows for each task are presented in Figures 6–11. These diagrams illustrate participants' verification workflows and possible transitions between different methods. Multicolored nodes represent distinct verification methods, while dotted-line nodes indicate supporting verification approaches. Nodes outlined in gray denote strategies employed after an error was identified. The weighted arrows indicate the frequency of participant transitions along each path. An asterisk (*) shown next to some participants indicates verification methods they intended to use in such scenarios (e.g., human verification) but did not pursue during the study due to practical or time constraints. Participants shown in red indicate that their AI output contained an error.

In our study, the most common workflow involved verifying outputs manually whenever possible and turning to alternative methods, such as querying the same AI model, using a different GenAI tool for cross-verification, or seeking sighted assistance when independent verification as insufficient (e.g., visual tasks; Figures 8, 9, and 11). While some relied on a single verification method for certain tasks (e.g., using SR features or inspecting code), others combined multiple strategies, especially when they encountered conflicting results—similar to observations in prior work [66].

Encountering hallucinations and inaccuracies during the verification process (Figure 6, 8, 9 and 11), whether within the same AI response or across different tools used for verification, further complicated participants' workflows. Across the study, participants encountered a range of errors while verifying AI-generated output. Table 3 summarizes the tasks in which errors occurred.

Error identification and response workflows were typically multi-step and mixed-method, with seeking human assistance serving as the last resort for visual tasks. During the study five participants encountered multiple errors, and all but P8 experienced at least one. Some errors were successfully identified (N = 8), while others went completely unnoticed (N = 7). In one case (S1T3: Figure 11, P1), although the participant noticed an inconsistency and attempted verification, the underlying error was not identified. Similarly, in two other instances (S2T2: Figure 8, P10; S2T3: Figure 9, P11), participants noticed inconsistencies but remained unsure whether an actual error existed, even after verification.

Our findings suggest that participants' prior experience with using GenAI in spreadsheets and coding may have influenced the verification methods they used. For example, P1 and P5, who both had prior experience with GenAI in spreadsheet tasks, relied less on manual verification and used more AI-based methods, either through the same AI or across different models. Given their coding backgrounds, they frequently verified outputs by examining the model's reasoning steps or reviewing the generated code.

Verification workflows also varied depending on the task. Tasks grounded in structured numerical reasoning, such as identifying the country with the highest inflation (S1T1; Figure 6), understanding inflation rate fluctuations (S1T2; Figure 7) or computing grades (S2T1; Figure 10), tended to yield relatively simpler, shorter verification flows. In cases S1T1 and S2T1, participants often confirmed the AI's answer using a few targeted checks, such as sorting a column with their SR or verifying that embedded formulas matched their expectations. For the trend analysis task (S1T2: Figure 7), three participants initially relied on same-model verification methods (e.g., checking the AI's reasoning or comparing its response in a new chat against spreadsheet data) but then diverged to use Excel formulas to confirm the computed values. These patterns indicate that most participants typically reverted to manual verification for numerical reasoning tasks.

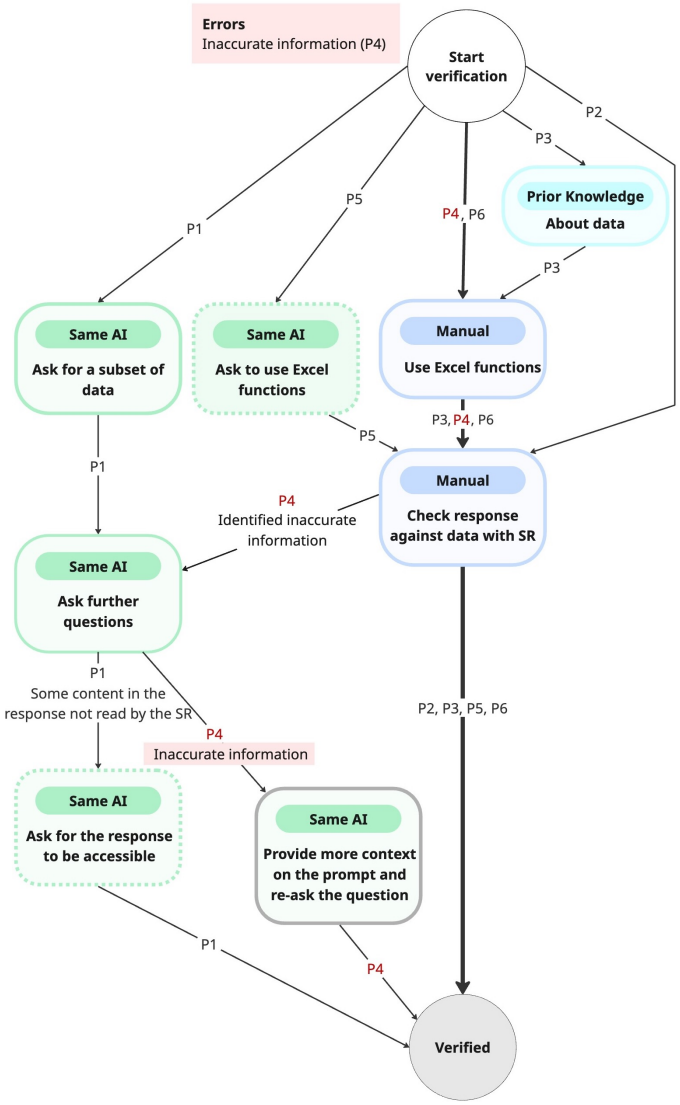


Figure 6: Finding the maximum value (S1T1). Verification methods for this task included manual and same-AI approaches, as well as leveraging prior knowledge. The most common workflow involved manual verification using Excel functions (e.g., sorting) and checking sheet data with the screen reader (n=4).

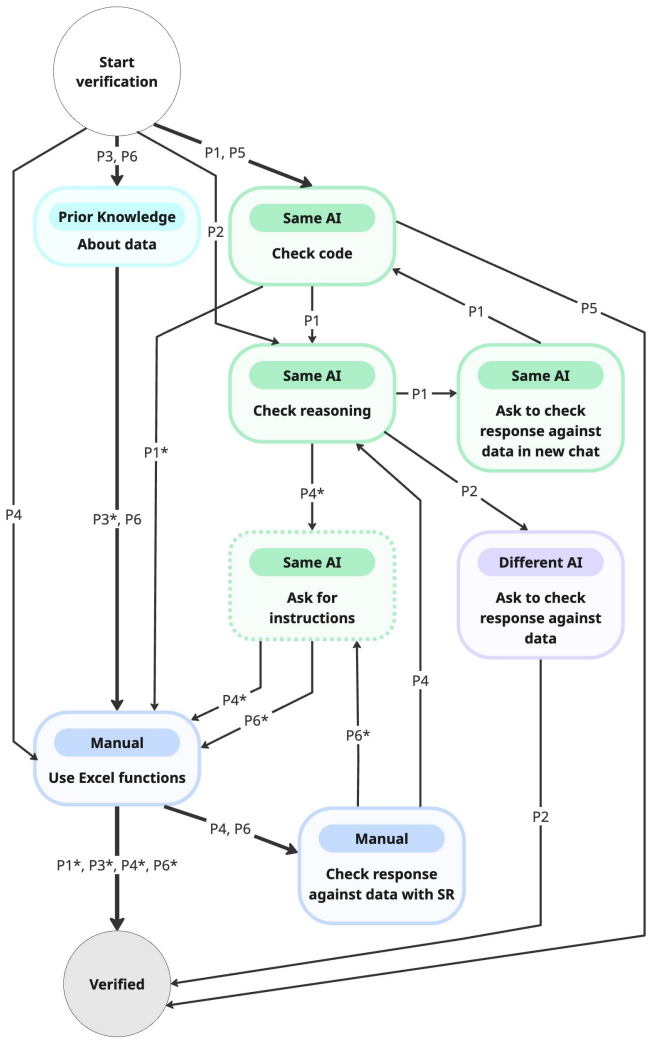


Figure 7: Trend analysis (S1T2). Participants used a variety of verification methods for this task, including manual, same-AI, different-AI, and prior-knowledge based approaches. The most common workflow involved using Excel formulas to verify fluctuation data points (n=4). While some participants began directly with manual verification, others first used same-AI reasoning or code-explanation features (n=3). Two participants relied solely on AI-based verification.

Verifying visual tasks (Figures 8, 9, and 11) proved to be the most difficult for participants, and these tasks also generated the highest number of errors. Most errors occurred in chart generation (N = 9), conditional formatting (N = 2), and applying visual styles to spreadsheet tables (N = 4) (Figures 3, 4, and 5).

When participants detected an inconsistency in a visual task using their SRs, they often turned to other AI tools for verification. However, these tools sometimes introduced additional inaccuracies, further confusing participants and leading them through multiple verification steps. As a result, these tasks often evolved into complex, recursive loops using several verification strategies. Figure 12 illustrates a representative case: P10 attempted to verify conditional formatting (red-colored cells) after ChatGPT claimed it had been applied, even though the formatting was absent from the sheet. In addition to ChatGPT's incorrect claim, the other tool P10 used for verification (Be My AI) also provided misleading feedback. This prompted P10 to apply multiple layers of verification methods, yet they remained uncertain about the output's accuracy and ultimately stated that they would seek sighted assistance (Figure 8).

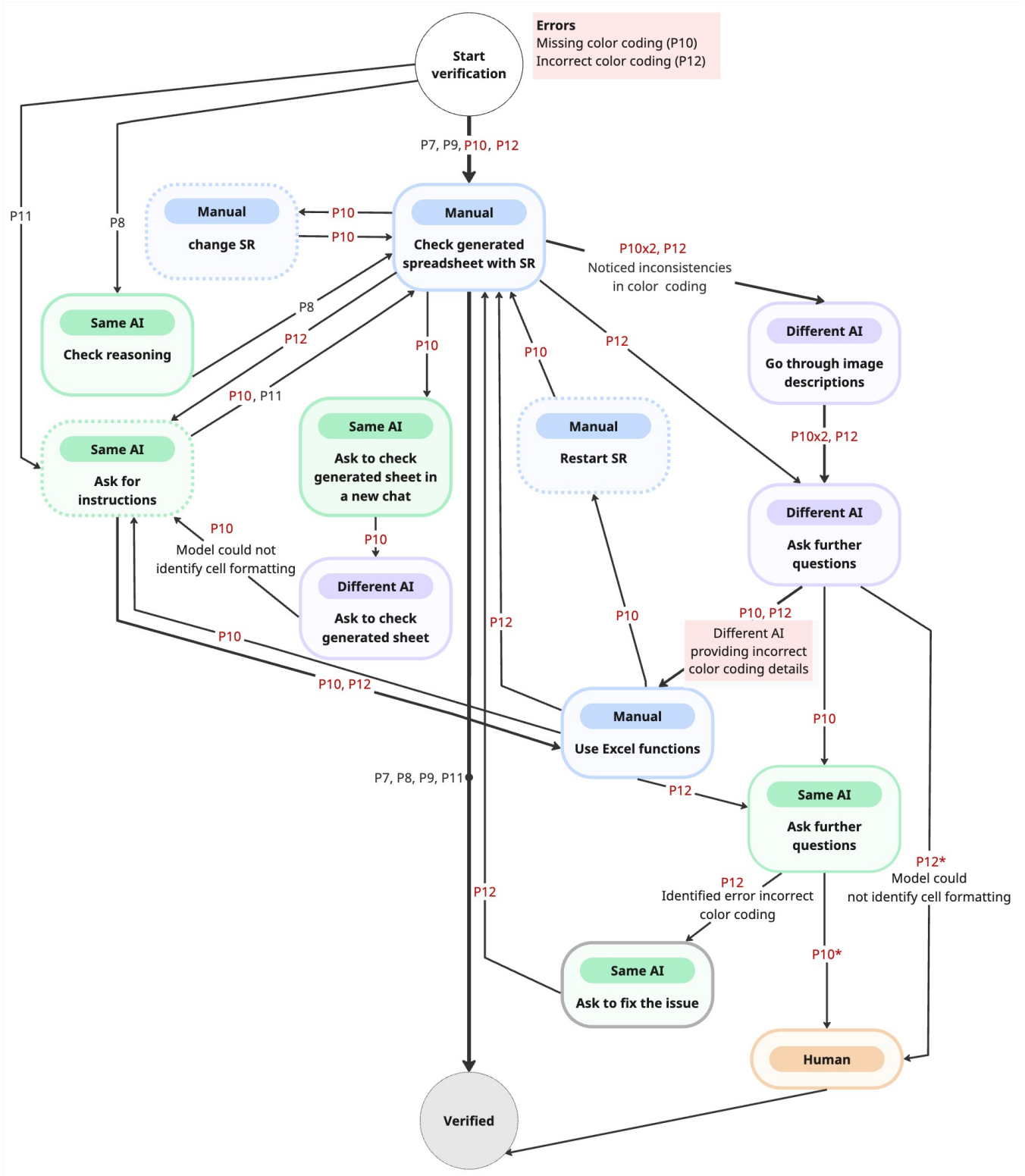


Figure 8: Applying conditional formatting (S2T2). All participants used manual methods to check formatting using SR commands. This approach allowed four participants to successfully verify the applied formatting. The remaining participants encountered more complex workflows in response to errors introduced by the AI tools.

A similar situation occurred for P12, where ChatGPT applied conditional formatting using an incorrect formula, causing non-F students to be highlighted in red (Figure 4). The secondary AI tool (e.g., Picture Smart AI) compounded the confusion by incorrectly describing the formatted cells, leaving P12 uncertain about their SR's feedback. However, when P12 explicitly prompted ChatGPT with *"check this sheet for the name Min Nair; when I view this, the name is not formatted as red and it seems like it should be, why?"*, the model acknowledged the error it had introduced and suggested a potential fix.

Both P10 and P12, along with three other participants, noted that they would ultimately seek sighted assistance for spreadsheet tasks requiring visual judgment (Figure 9). P12 further emphasized the need for complete accuracy in such cases, explaining, *"I believe as a person who's blind, making visually compelling spreadsheets on a regular basis is not a task that most people who are blind can do effectively, independently. I don't know if there's data to prove that, but my own anecdotal data says so. This is an area where, if the technology could do that, it'd be great—but it can't be halfway. It can't even be 95% of the way; if it's not 100%, it might as well be Zero."*

Models also generated consecutive errors that undermined trust (Figure 6). For example, when P4 asked ChatGPT to identify the country with the highest inflation rate in 2023 (S1T1), the model first returned Venezuela, then another incorrect country, before arriving at the correct answer (Lebanon) only after P4 manually verified the dataset and supplied clarifying details. These consecutive inaccuracies reduced P4's confidence in the model's reliability. Similarly, three participants reported that errors lowered their confidence in subsequent tasks and six said accurate performance increased their trust moving forward.

After identifying an error, participants employed several strategies beyond seeking human assistance to correct it. For instance, three participants (P4, P5, P6) refined their prompts by adding contextual information before re-asking the question, while others (P1, P7, P12) used the same AI model to attempt error correction. Although the model successfully provided accurate fixes in some cases (e.g., for P7 in S2T3 and P12 in S2T2), participants never concluded their verification workflows without turning to alternative methods for confirmation.

Participants emphasized the trade-off between speed, reliability, and effort when working with GenAI tools. While they valued AI for accelerating tasks that typically required significant time, effort, or expertise, such as generating charts, creating formulas, applying formatting, or analyzing trends, some participants (n=4) noted that errors often negated these benefits by requiring extensive verification. For straightforward tasks (e.g., sorting a column to find the highest or lowest value), participants reported it was faster and more trustworthy to perform them manually in Excel using their SRs. Although verification was sometimes challenging for certain tasks (e.g., visual tasks), all participants acknowledged its value in spreadsheet work. They mentioned that AI was particularly useful for gaining an overview of the sheet (n=11), understanding trends (n=6), completing formula-related tasks (n=5), answering data-related questions (n=4), and summarizing data (n=3). For more complex or visually intensive tasks, participants (n=8) described GenAI as a useful starting point before seeking sighted verification,

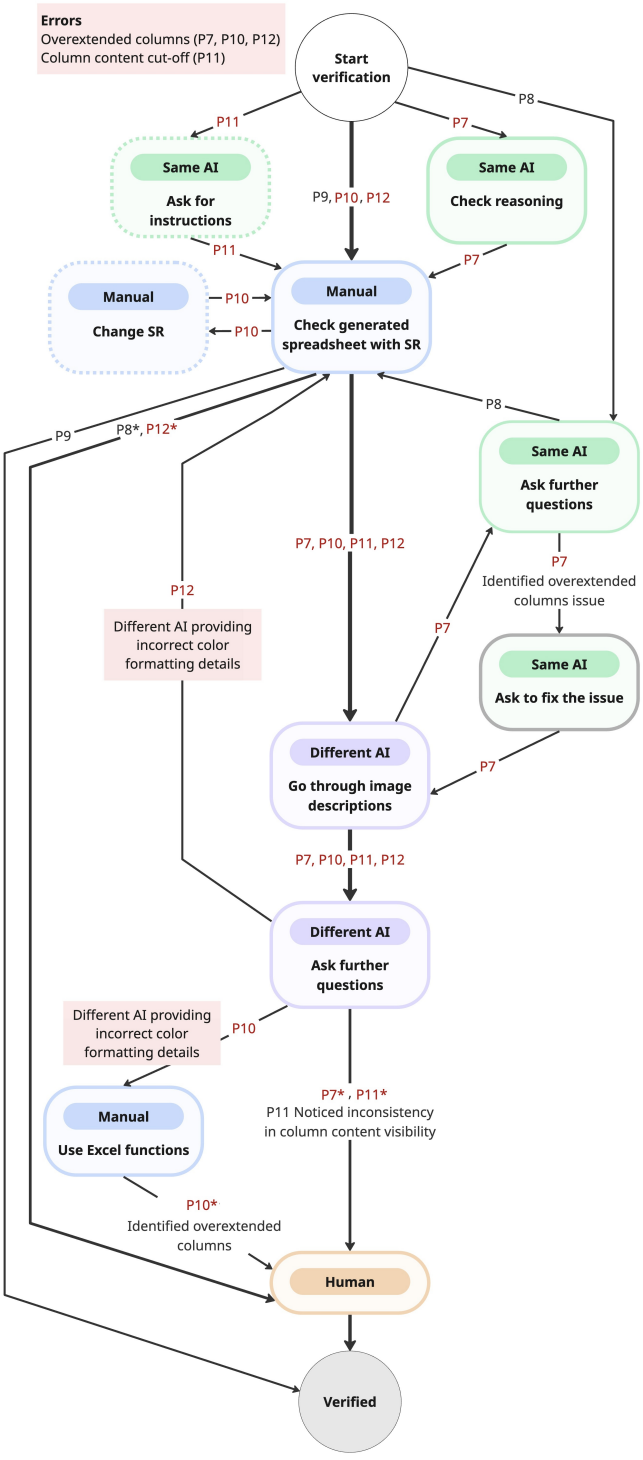


Figure 9: Applying visual formatting (S2T3). Participants used a variety of methods for this task, but the most common workflow involved manually checking the applied formatting with their SR and then turning to different AI tools for further verification. Most participants (n=5) said that they would use human confirmation as the final step to ensure accuracy, regardless of whether errors were detected.

as they believed AI outperformed their own ability to handle visually dependent tasks. As P11 explained: *"I would definitely use AI for it, for sure. As low as my confidence was in the AI, I have less confidence in myself—so especially for something visual, I probably would have just had AI do it."*

Several participants also highlighted that AI tools increased their efficiency and fostered a sense of equity, enabling them to work on par with sighted peers despite persistent accessibility barriers. As P1 shared: *"I think it's been pretty great. Being blind, I was really struggling sometimes in a very highly competitive kind of environment. I always had to work many, many long hours to keep up with everyone. But now with AI, I'm so much faster at things that are accessible that I don't mind struggling on a few that are not very accessible. I still get a lot more done within the same number of hours because I'm just so much better with AI than a lot of other people."*

Overall, participants weighed potential time savings against the verification workload introduced by errors, making speed, effort, and reliability central factors in deciding whether to rely on AI or revert to manual or alternative verification methods.

5 Discussion

5.1 Comparison of Verification Strategies Between Blind and Sighted Spreadsheet Users

Our findings show that blind spreadsheet users adopt verification practices similar to those reported in studies of sighted analysts working with GenAI [35], but they do so under unique constraints imposed by SR access. Like sighted users, participants relied on quick sanity checks rather than exhaustive validation. For example, several participants described manually creating formulas or spot-checking portions of the values provided in trend analyses. This approach extended to data modification tasks: in formula generation and conditional formatting scenarios, participants typically checked one or two rows before assuming the remainder had been applied consistently. In some cases, participants asked the same AI model to critique its own output. While prior work cautions that relying on language models to audit themselves is an anti-pattern [33], some participants used this strategy to surface potential inconsistencies. However, they never relied solely on self-critiques, instead combining them with manual spot checks to build confidence in the output. Participants with prior programming experience also drew on validation practices such as reviewing model's reasoning or code [35].

Unlike sighted analysts, who can rely on quick *"eyeballing"* strategies [35], our participants encountered significant difficulties. To access data, they had to navigate tables cell by cell using their SRs, which greatly increased the manual effort required for verification. The primary challenge involved visually intensive tasks, such as interpreting charts, which were impossible to accomplish with a SR. Participants could not detect issues with incorrect data points, chart formatting, or rendering, even when reviewing the model's reasoning or underlying code. Verification of visual content was further complicated by inaccuracies and inconsistencies in secondary verification tools, creating multi-step workflows that demanded additional time and labor. Together, these findings show that verification for blind users is not simply a slower version of sighted workflows but a qualitatively different, more complex, and layered process.

5.2 Cross-Domain Comparison of Blind Users' Verification Strategies and Trust in GenAI

In spreadsheet verification scenarios, our participants employed methods similar to those used in image verification tasks (object and scene descriptions, visualizations), such as asking follow-up questions, cross-referencing across devices and applications, testing with different models, comparing against known information, and seeking human confirmation [2, 4, 29, 59, 66, 73, 74].

Our findings also aligned with Tang et al.'s [66] concept of everyday uncertainty, in which blind people consistently approached GenAI outputs with skepticism. Across both data analysis and spreadsheet modification scenarios, participants remained cautious and did not fully trust GenAI outputs without verification. This contrasts with prior work showing that blind users have sometimes placed undue trust in AI-generated descriptions [48]. Nevertheless, undue trust occasionally surfaced when a different AI tool validated and agreed with a GenAI output without exposing inconsistencies. Agreement between multiple GenAI tools sometimes led users to conclude their verification process and accept inaccurate results when the models produced matching outputs. At the same time, participants expressed concerns that multiple tools could fail in the same way—a concern borne out when different models jointly reported formatting changes in a spreadsheet that had not actually occurred. Although such agreement did not result in complete reliance on the output, it created confusion about which verification strategies were dependable and even led participants to question the accuracy of trusted methods such as manual verification with SRs. Ultimately, this uncertainty motivated participants to seek human verification as a final safeguard–a common practice, reported in prior work, that is used by blind users in accuracy-critical scenarios [4].

Our study extends prior research showing that conflicting results from different GenAI-based verification tools, particularly for image descriptions, create uncertainty for blind users [4, 73]. We further observed inconsistencies among multiple models within secondary AI tools (e.g., discrepancies between ChatGPT and Claude within Picture Smart AI), which added to participants' confusion. To reduce uncertainty, participants employed dialog-based verification strategies, engaging in back-and-forth questioning with other AI tools to cross-check and validate accuracy, a workflow similar to that identified in the use of external tools for understanding surroundings [73].

Similar to prior work showing that blind users engage in multi-step verification when interacting with image descriptions, documents or web content [2, 29, 59, 66], our participants constructed multi-method verification ecosystems that spanned SRs, spreadsheet features, the same AI model, different AI tools, and, when needed, sighted assistance. These workflows reveal how blind users mobilized strategies identified in earlier research—such as asking further questions (e.g., re-prompting, clarifying questions), cross-checking, and seeking human help [2, 4, 29]—while integrating newer capabilities, including reasoning features and code-based

support [32, 55], into their verification processes. In doing so, participants extended these practices into the domain of accuracy-critical spreadsheet manipulation and data analysis, producing verification workflows that were often more complex and multi-step than those observed in prior visual-description tasks.

5.3 Improving Verification to Accommodate Blind User Needs

Although blind users often feel the need to verify each GenAI-produced output, indicating limited trust in these systems, our findings show that this verification requirement does not preclude meaningful adoption. Participants noted that for straightforward tasks (e.g., using the sort feature to identify the highest value), they typically retained their established workflows because these methods remained more efficient. However, GenAI became especially valuable as task complexity increased. Importantly, GenAI helped address long-standing accessibility challenges faced by blind spreadsheet users, including difficulties with obtaining content overviews, accessing trend information, generating formulas, and interpreting visual formatting [41, 64, 65]—tasks that were otherwise extremely difficult, due to cell-by-cell navigation and excessive SR maneuvering, or entirely inaccessible using SRs alone. Even in cases where participants anticipated needing human assistance for visually intricate tasks, they perceived GenAI as a valuable entry point with greater capability in interpreting visual aesthetics than their own knowledge. However, the GenAI verification process introduced new challenges for blind users (Section 4.3), which could be mitigated through the following improvements:

5.3.1 Improving Feedback on Visual Tasks. Prior work has extensively documented inaccuracies in visual content generated by multimodal LLMs, particularly for charts and tabular structures, highlighting their susceptibility to hallucination and misrepresentation [26, 71, 72]. Consistent with these findings, our study showed that GenAI tools produced errors more frequently in visually oriented tasks (15 of 18 error instances), such as chart generation and visual formatting, than in data extraction and manipulation tasks. Notably, about half of these visual errors went unnoticed by participants.

Although participants supplemented their workflows with external GenAI-based visual assistance tools (e.g., Picture Smart AI, Be My AI, Seeing AI), these tools did not reliably flag visual inaccuracies in their generated descriptions. As a result, users remained unaware of errors such as incorrect chart formatting, truncated table cells, or overextended columns. While some issues could be uncovered through targeted prompting [18], participants often lacked the confidence or domain knowledge to formulate such visual-specific queries. These findings highlight the need to train external AI tools with visual error detection capabilities that proactively surface issues without requiring expert prompting from blind users.

Participants also emphasized the value of embedding multiple GenAI model image verification directly within their verification workflows. Rather than relying on one model for image descriptions [2, 4], they proposed that GenAI systems could automatically cross-check visual descriptions across multiple models and highlight discrepancies. Existing tools like Picture Smart AI already offer multi-model descriptions but present each output in isolation, forcing users to manually compare alternative descriptions—an effort that is cognitively demanding and prone to oversight, especially in tasks that require identifying trends or higher-level patterns [59]. Chen et al. [13] demonstrated that surfacing variations across multi-model image descriptions, by highlighting agreement and disagreement among model responses, can help blind users identify inaccurate or incomplete image descriptions. This automated comparison could be used to identify inconsistencies in spreadsheet-based visual tasks observed in our study, such as duplicated legends, truncated chart regions, or misaligned columns.

However, our analysis also suggests that surfacing description-level variations alone is insufficient for spreadsheet-related tasks. Unlike general image description scenarios, users must verify not only whether different tools' textual descriptions match the visual output, but also whether the visual output itself is accurate relative to the underlying dataset. Several participants encountered charts that mis-plotted values or omitted data entirely—errors they were unable to detect. Here, multi-model verification could include extracting underlying data points from the generated chart, comparing them to the source dataset, and reporting inconsistencies. Such cross-checking would benefit both blind and sighted users, as chart hallucinations have been widely observed in regular GenAI use [38, 60].

Additionally, GenAI-based visual assistance tools often operate on screenshots, meaning that only a subset of the spreadsheet is captured. As with known challenges in photo-based image descriptions [2, 13], this limitation can obscure formatting problems that exist outside the visible frame. Multi-model systems that accept the entire generated spreadsheet (rather than a screenshot) could detect issues such as off-screen misalignments, or incomplete formatting.

5.3.2 Improving Step-by-step Reasoning. Explainability features, such as step-by-step explanations of how a model produces a response, are known to improve transparency in spreadsheet tasks [70]. In our study, most participants (n=10) found this functionality valuable. When blind participants could access the reasoning behind a result, they reported greater confidence in its accuracy because it helped them confirm that the model had correctly interpreted their intent and followed an appropriate process to complete the task. However, unlike examining the underlying code, step-by-step reasoning alone never provided a single, definitive means of verification capable of ensuring full confidence.

Detailed reasoning in visually intensive tasks was considered useful. Although prior research has shown that SR users often find it difficult to detect and interpret visual formatting in spreadsheets [57], our participants indicated that explanations describing specific formatting details, such as styles, colors, or formatting applied to particular cells, in the reasoning itself could help them verify tasks that sighted users could confirm at a glance.

Our findings further reveal accessibility barriers to discovering reasoning features, extending prior reports of usability challenges for SR users of GenAI interfaces [2, 4]. Many participants were initially unaware that chain-of-thought explanations were available until this was pointed out to them. Although the model interfaces included this option, it was not easily discoverable with SRs. Because tools like ChatGPT begin reading the generated output immediately,

any reasoning related controls positioned between the query and the response were skipped entirely. To ensure that SR users can fully benefit from these features, explainability options need to be clearly integrated into the SR navigation flow.

5.3.3 Prompt Engineering to Support Verification. Our findings show that prompting was not a simple one-shot interaction but a dynamic, iterative process that became central to blind participants' verification workflows [33]. Consistent with prior work describing GenAI prompting as *messy* and contingent on experimentation [22, 66], our participants, like sighted users, engaged in extensive back-and-forth dialogue to clarify outputs, surface inconsistencies, and refine requests until they were confident in the results. During this process, they employed several distinct strategies to support verification:

- *Revealing disability for customized output*: Participants frequently disclosed their blindness to obtain responses tailored for non-visual verification [66], particularly when requesting spreadsheet overviews, detailed visual descriptions, or step-by-step SR instructions for performing tasks manually.
- *Prompting for selective views*: Participants also crafted prompts to generate *selective views* of underlying data, asking the model to output targeted subsets of a spreadsheet. This approach reduced the need to switch between multiple applications or navigate large grids to locate relevant data.
- *Asking for self-correction*: When they detected discrepancies through manual SR checks, participants often informed the model of the inconsistency and requested a correction. In some cases, they explicitly challenged the AI's claims to test its reasoning.

Crafting effective verification prompts required substantial metacognitive effort to express goals at the right level of abstraction [47], and participants often did not know which questions to ask, particularly when seeking visual details they could not directly perceive. Future GenAI interfaces could better support this process by scaffolding follow-up questions or offering prompt suggestions to aid blind users.

6 Limitations

To observe how blind participants used spreadsheets, GenAI tools, and SRs while performing tasks, we asked them to share their screens and allowed session recordings. This arrangement may have influenced their behavior, prompting them to avoid strategies they might typically use when encountering difficulties. Future work could employ more creative and unobtrusive methods to capture blind users' natural interactions with GenAI-assisted spreadsheet tasks.

Although some participants (n=5) had limited prior experience using GenAI for spreadsheet tasks, such as formula generation, none had extensive experience with GenAI-assisted spreadsheet workflows. Consequently, our findings may not fully capture the verification strategies and challenges that emerge with sustained use. Longitudinal studies are needed to explore how SR users adapt to GenAI-assisted spreadsheets over time and to identify any enduring accessibility barriers.

Our participants reported an intermediate level of spreadsheet proficiency on average: only two described their Excel skills as advanced and one as a beginner. Expert users may employ different verification strategies or workarounds, and future research including highly proficient spreadsheet users could reveal additional practices. Moreover, participants were highly educated, early adopters of technology, and primarily from English-speaking countries (e.g., the U.S. and Australia), which may limit the generalizability of our findings to the broader blind population. Finally, we acknowledge our positionality as sighted accessibility researchers, which may have influenced our interpretation of blind users' experiences and perspectives when interacting with these technologies.

7 Conclusion

We conducted a study with 12 blind screen reader users to explore their verification strategies and the challenges they face when performing AI-assisted spreadsheet tasks. We found that verifying GenAI outputs required additional time and effort, particularly for visually intensive tasks such as chart generation and formatting. Participants employed a range of approaches, including manual checks, cross AI-assisted verification, external AI tools, prior knowledge, and sighted human assistance, often combining these methods in a multi-step workflow to detect and correct errors.

Our study shows that blind users' verification workflows are not merely slower versions of those of sighted users but are more complex and qualitatively different. These findings offer valuable insights for designers of GenAI tools and assistive technologies, highlighting the need to enhance verification experiences in AI-assisted spreadsheet tasks. Improving error-identification features will better support blind users in independently verifying and working with information, with benefits for education and employment.

Acknowledgments

We thank our participants for their valuable contributions in sharing their experiences and insights. We also thank Ramona Mandy for helping with the evaluation of the screen reader accessibility of GenAI tools, as well as the other members of the Monash Assistive Technology and Society (MATS) Centre for their support and valuable discussions.

References

[1] Iyad Abu Doush, Enrico Pontelli, Dominic Simon, Tran Cao Son, and Ou Ma. 2009. Making Microsoft Excel™: multimodal presentation of charts. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 147–154.

[2] Rudaiba Adnin and Maitraye Das. 2024. "I look at it as the king of knowledge": How Blind People Use and Understand Generative AI Tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 64, 14 pages. doi:10.1145/3663548.3675631

[3] Aira. 2026. Aira. Retrieved January 26, 2026 from https://aira.io/

[4] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N. Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 61, 17 pages. doi:10.1145/3663548.3675659

[5] Bedour Alshaigy and Virginia Grande. 2024. Forgotten Again: Addressing Accessibility Challenges of Generative AI Tools for People with Disabilities. In *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction*. Association for Computing Machinery, New York, NY, USA, 1–6.

[6] Alex Atcheson, Omar Khan, Brian Siemann, Anika Jain, and Karrie Karahalios. 2025. " I'd Never Actually Realized How Big An Impact It Had Until Now": Perspectives of University Students with Disabilities on Generative Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[7] Cynthia L Bennett, Renee Shelby, Negar Rostamzadeh, and Shaun K Kane. 2024. Painting with Cameras and Drawing with Text: AI Use in Accessible Creativity. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–19.

[8] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[9] Martin Broadhurst. 2024. Leveraging ChatGPT for Excel: How large language models are changing spreadsheet practices. *Journal of AI, Robotics & Workplace Automation* 3, 3 (2024), 26–40.

[10] Alexia Cambon, Brent Hecht, Ben Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan. 2023. *Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity.* Technical Report. Microsoft. https://www.microsoft.com/en-us/research/publication/early-llm-based-tools-for-enterprise-information-workers-likely-provide-meaningful-boosts-to-productivity/

[11] Alexander Campbell. 2024. Using ai chatbots to produce engineering spreadsheets in an advanced structural steel design course. In *2024 ASEE Annual Conference & Exposition*.

[12] George Chalhoub and Advait Sarkar. 2022. "It's Freedom to Put Things Where My Mind Wants": Understanding and Improving the User Experience of Structuring Data in Spreadsheets. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[13] Meng Chen, Akhil Iyer, and Amy Pavel. 2025. Surfacing Variations to Calibrate Perceived Reliability of MLLM-generated Image Descriptions. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–17.

[14] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 245–255.

[15] Yibin Chen, Yifu Yuan, Zeyu Zhang, Yan Zheng, Jinyi Liu, Fei Ni, Jianye Hao, Hangyu Mao, and Fuzheng Zhang. 2025. SheetAgent: towards a generalist agent for spreadsheet reasoning and manipulation via large language models. In *Proceedings of the ACM on Web Conference 2025*. 158–177.

[16] claude. 2026. Claude. Retrieved January 26, 2026 from https://claude.ai/

[17] Grenville J Croll. 2007. The importance and criticality of spreadsheets in the city of London. *arXiv preprint arXiv:0709.4063* (2007).

[18] Amit Kumar Das, Mohammad Tarun, and Klaus Mueller. 2025. Charts-of-Thought: Enhancing LLM Visualization Literacy Through Structured Data Extraction. *arXiv preprint arXiv:2508.04842* (2025).

[19] DeepSeek. 2025. DeepSeek. Retrieved January 26, 2026 from https://www.deepseek.com/en/

[20] Iyad Abu Doush and Enrico Pontelli. 2010. Non-visual navigation of tables in spreadsheets. In *Proc. of ICCHP*. 108–115.

[21] Iyad Abu Doush and Enrico Pontelli. 2013. Non-visual navigation of spreadsheets: Enhancing accessibility of Microsoft Excel™. *Universal access in the information society* 12 (2013), 143–159.

[22] Ian Drosos, Advait Sarkar, Xiaotong Xu, Carina Negreanu, Sean Rintel, and Lev Tankelevitch. 2024. "It's like a rubber duck that talks back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Newcastle upon Tyne, United Kingdom) *(CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, Article 16, 21 pages. doi:10.1145/3663384.3663389

[23] Benjamin G Edelman, Donald Ngwe, and Sida Peng. 2023. Measuring the Impact of AI on Information Worker Productivity. *Available at SSRN 4648686* (2023). https://ssrn.com/abstract=4648686

[24] Be My Eyes. 2026. Introducing: Be my AI. Retrieved January 26, 2026 from https://www.bemyeyes.com/blog/introducing-be-my-ai

[25] Claudia Flores-Saviaga, Benjamin V Hanrahan, Kashif Imteyaz, Steven Clarke, and Saiph Savage*. 2025. The Impact of Generative AI Coding Assistants on Developers Who Are Visually Impaired. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[26] James Ford, Xingmeng Zhao, Dan Schumacher, and Anthony Rios. 2025. Charting the future: Using chart question-answering for scalable evaluation of LLM-driven data visualizations. In *Proceedings of the 31st International Conference on Computational Linguistics*. 7497–7510.

[27] Anujay Ghosh, Monalika Padma Reddy, Satwik Ram Kodandaram, Utku Uckun, Vikas Ashok, Xiaojun Bi, and IV Ramakrishnan. 2024. Screen Reading Enabled by Large Language Models. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–5.

[28] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 1–8.

[29] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–21.

[30] Google. 2026. Gemini. Retrieved January 26, 2026 from https://gemini.google.com/

[31] Google. 2026. Gemini in Google Sheets. Retrieved January 26, 2026 from https://workspace.google.com/intl/en_au/resources/spreadsheet-ai/

[32] Google. 2026. Gemini thinking. Retrieved January 26, 2026 from https://ai.google.dev/gemini-api/docs/thinking

[33] Andrew D Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, et al. 2023. Co-audit: tools to help humans double-check AI-generated content. *arXiv preprint arXiv:2310.01297* (2023).

[34] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. 2024. Vizability: Enhancing chart accessibility with llm-based conversational interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[35] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M Drucker. 2024. How do analysts understand and verify ai-assisted data analyses?. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.

[36] Jongrim Ha, M. Ayhan Kose, and Franziska Ohnsorge. 2023. One-Stop Source: A Global Database of Inflation. *Journal of International Money and Finance* 137 (October 2023), 102896. doi:10.1016/j.jimonfin.2023.102896

[37] Isao Hiramatsu, Kousuke Mouri, Bipin Indurkhya, and Keiichi Kaneko. 2019. Development of an Answer Verification System for Helping Visually Impaired People to Learn a Spreadsheet Software. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 190–195.

[38] Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[39] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. In *Workshop on Demographic Diversity in Computer Vision@ CVPR 2025*.

[40] Jeevana Priya Inala, Chenglong Wang, Steven Drucker, Gonzalo Ramos, Victor Dibia, Nathalie Riche, Dave Brown, Dan Marshall, and Jianfeng Gao. 2024. Data analysis in the era of generative AI. *arXiv preprint arXiv:2409.18475* (2024).

[41] Chutian Jiang, Wentao Lei, Emily Kuang, Teng Han, and Mingming Fan. 2023. Understanding Strategies and Challenges of Conducting Daily Data Analysis (DDA) Among Blind and Low-vision People. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 1–15.

[42] Rie Kamikubo, Farnaz Zamiri Zeraati, Kyungjun Lee, and Hernisa Kacorri. 2024. AccessShare: Co-designing Data Access and Sharing with Blind People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.

[43] Bennett Kankuzi, Bassey Isong, and Lucia Letlonkane. 2017. Using the spreadsheet paradigm to introduce fundamental concepts of programming to novices. *Proceedings of SACLA'17* (2017), 3–5.

[44] Gyeongdeok Kim, Chungman Lim, and Gunhyuk Park. 2025. I-Scratch: Independent Slide Creation With Auditory Comment and Haptic Interface for the Blind and Visually Impaired. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[45] Satwik Ram Kodandaram, Utku Uckun, Xiaojun Bi, IV Ramakrishnan, and Vikas Ashok. 2024. Enabling Uniform Computer Interaction Experience for Blind Users through Large Language Models. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 73, 14 pages. doi:10.1145/3663548.3675605

[46] Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhao-Xiang Zhang. 2023. Sheetcopilot: Bringing software productivity to the next level through large language models. *Advances in Neural Information Processing Systems* 36 (2023), 4952–4984.

[47] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 598, 31 pages. doi:10.1145/3544548.3580817

[48] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *proceedings of the 2017 CHI conference on human*

factors in computing systems. Association for Computing Machinery, New York, NY, USA, 5988–5999.
[49] Microsoft. 2024. Seeing AI. Retrieved September 9, 2025 from https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/
[50] Microsoft. 2026. Microsoft 365 Copilot. Retrieved January 26, 2026 from https://www.microsoft.com/en-us/microsoft-365/enterprise/microsoft-365-copilot
[51] Microsoft. 2026. Microsoft Excel. Retrieved January 26, 2026 from https://www.microsoft.com/en-au/microsoft-365/excel
[52] Jakob Nielsen. 2023. AI: First New UI Paradigm in 60 Years. Retrieved January 26, 2026 from https://www.nngroup.com/articles/ai-paradigm/
[53] OpenAI. 2024. Improvements to data analysis in ChatGPT. Retrieved January 26, 2026 from https://openai.com/index/improvements-to-data-analysis-in-chatgpt/
[54] OpenAI. 2026. ChatGPT. Retrieved January 26, 2026 from https://openai.com/chatgpt/
[55] OpenAI. 2026. Reasoning Models. Retrieved January 26, 2026 from https://platform.openai.com/docs/guides/reasoning
[56] Minoli Perera, Swamy Ananthanarayan, Cagatay Goncu, and Kim Marriott. 2025. The Sky is the Limit: Understanding How Generative AI can Enhance Screen Reader Users' Experience with Productivity Applications. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[57] Minoli Perera, Bongshin Lee, Eun Kyoung Choe, and Kim Marriott. 2024. Visual Cues for Data Analysis Features Amplify Challenges for Blind Spreadsheet Users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16.
[58] Freedom Scientific. 2026. New and Improved Features in JAWS: Picture Smart AI. Retrieved January 26, 2026 from https://www.freedomscientific.com/training/jaws/new-and-improved-features/
[59] JooYoung Seo, Sanchita S Kamath, Aziz Zeidieh, Saairam Venkatesh, and Sean McCurry. 2024. MAIDR Meets AI: Exploring Multimodal LLM-Based Data Visualization Interpretation by and with Blind and Low-Vision Users. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–31.
[60] Philip Wootaek Shin, Jack Sampson, Vijaykrishnan Narayanan, Andres Marquez, and Mahantesh Halappanavar. 2025. Losing the Plot: How VLM responses degrade on imperfect charts. *arXiv preprint arXiv:2509.18425* (2025).
[61] AM Silverman, LP Rosenblum, EC Bolander, CR Rhoads, and K Bleach. 2022. Technology and accommodations: Employment experiences of US adults who are blind, have low vision, or are deafblind. American Foundation for the Blind. Retrieved January 26, 2026 from https://www.afb.org/sites/default/files/2022-01/AFB_Workplace_Technology_Report_Accessible_FINAL.pdf
[62] Mukul Singh, José Cambronero Sánchez, Sumit Gulwani, Vu Le, Carina Negreanu, Mohammad Raza, and Gust Verbruggen. 2023. Cornet: Learning Table Formatting Rules By Example. *Proceedings of the VLDB Endowment* 16, 10 (2023), 2632–2644.
[63] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of ChatGPT: findings from a questionnaire study of early users. In *Proceedings of the 5th international conference on conversational user interfaces*. 1–10.
[64] Tony Stockman. 2004. The design and evaluation of auditory access to spreadsheets. In *ICAD 04-Tenth Meeting of the International Conference on Auditory Display*.
[65] Tony Stockman, Christopher Frauenberger, and Greg Hind. 2005. Interactive sonification of spreadsheets. In *ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*.
[66] Xinru Tang, Ali Abdolrahmani, Darren Gergle, and Anne Marie Piper. 2025. Everyday Uncertainty: How Blind People Use GenAI Tools for Information Access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[67] Carter Temm. 2025. AI-Content-Describer: NVDA add-on that provides descriptions for controls and images. https://github.com/cartertemm/AI-content-describer
[68] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 43–52.
[69] Simon Thorne. 2023. Experimenting with chatgpt for spreadsheet formula generation: Evidence of risk in ai generated spreadsheets. *arXiv preprint arXiv:2309.00095* (2023).
[70] Simon Thorne. 2024. Understanding and Evaluating Trust in Generative AI and Large Language Models for Spreadsheets. *arXiv preprint arXiv:2412.14062* (2024).
[71] Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. 2025. Is this chart lying to me? Automating the detection of misleading visualizations. *arXiv preprint arXiv:2508.21675* (2025).
[72] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems* 37 (2024), 113569–113697.
[73] Jingyi Xie, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M Carroll. 2025. Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[74] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. Emerging practices for large multimodal model (lmm) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv:2407.08882* (2024).
[75] He Zhang, Nicholas J Falletta, Jingyi Xie, Rui Yu, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2025. Enhancing the Travel Experience for People with Visual Impairments through Multimodal Interaction: NaviGPT, A Real-Time AI-Driven Mobile Navigation System. In *Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work*. 29–35.

A Verification Flows for Tasks

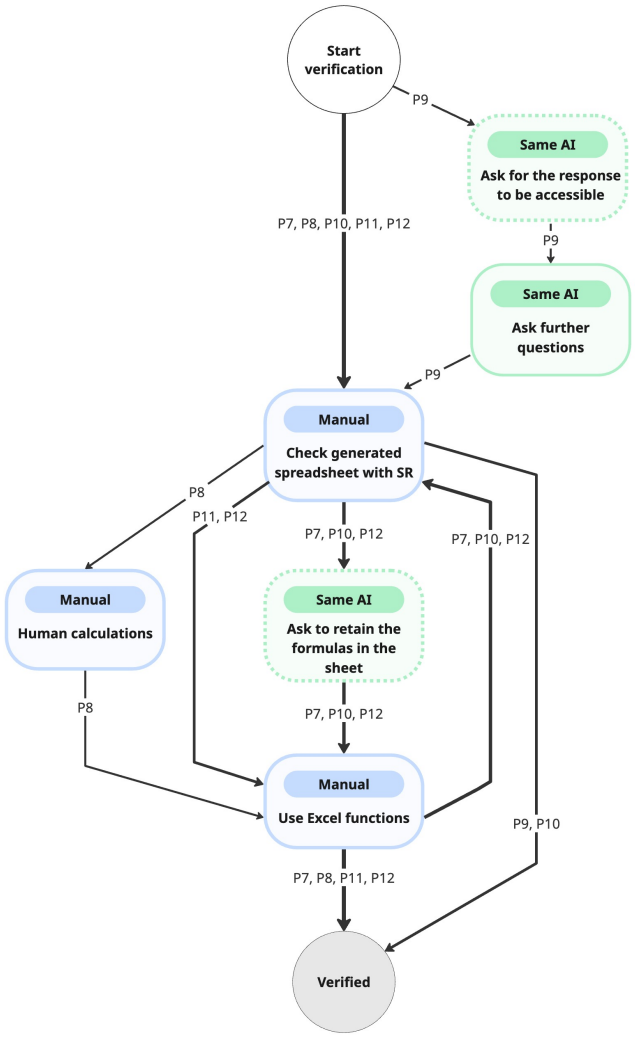


Figure 10: Formula Generation (S2T1). For this task, participants relied solely on manual verification or a combination of manual and same-AI verification methods. The most common workflow involved first checking the generated spreadsheet for existing formulas; if formulas were missing, participants asked the AI to retain them, then manually verified their accuracy using Excel features to reveal and inspect the formulas (n=3).

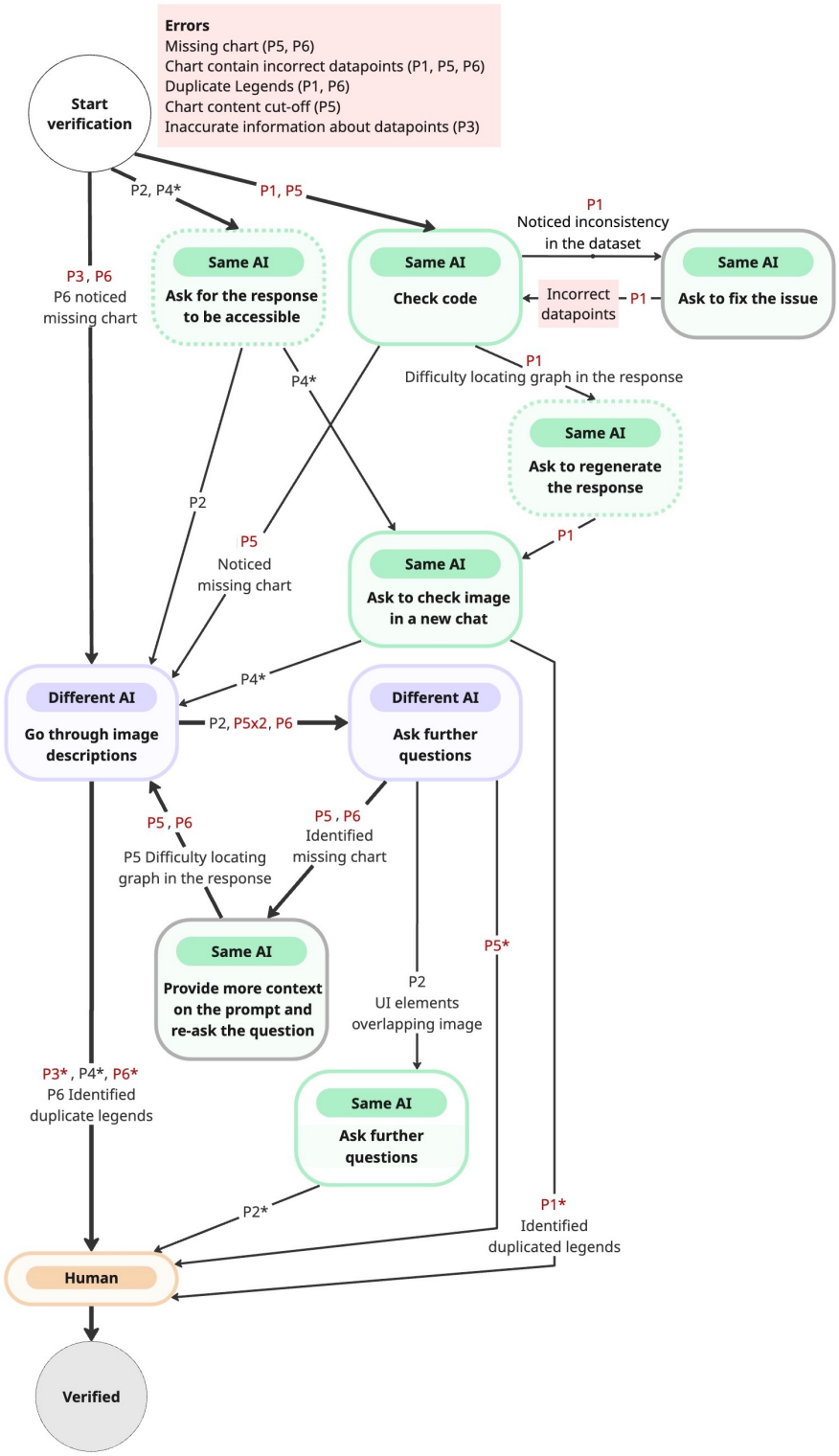


Figure 11: Chart creation (S1T3). All participants used a combination of AI-based verification methods, involving either the same AI model or different models. Most participants (n=5) used alternative AI tools rather than continuing verification within the same tool. Encountering errors and accessibility issues often complicated their workflows, resulting in multi-step verification processes. Regardless of whether they encountered an error, all participants reported that they would seek human assistance to confirm the accuracy of the chart.

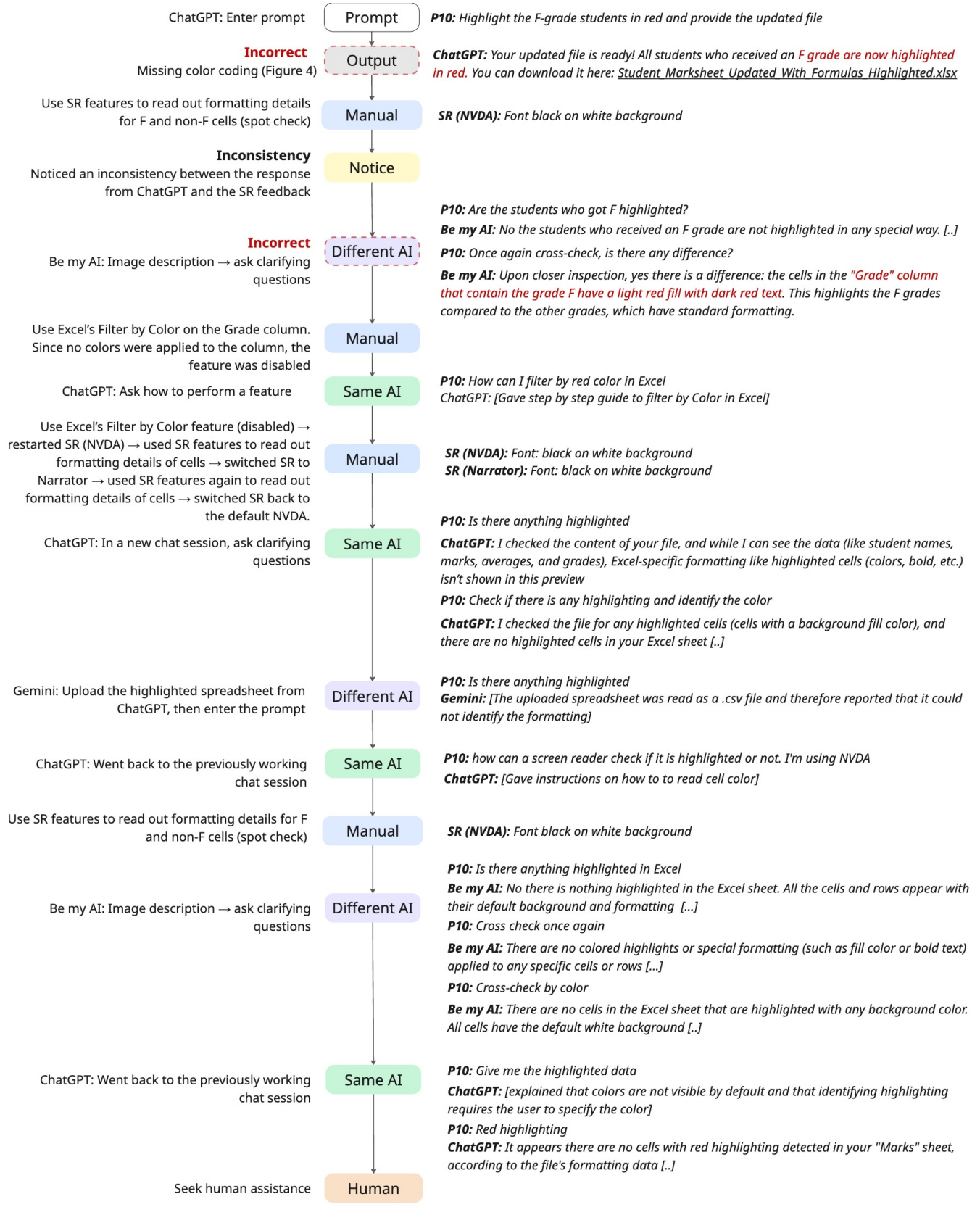


Figure 12: Detailed verification workflow of P10 during task S2T2 (applying conditional formatting)