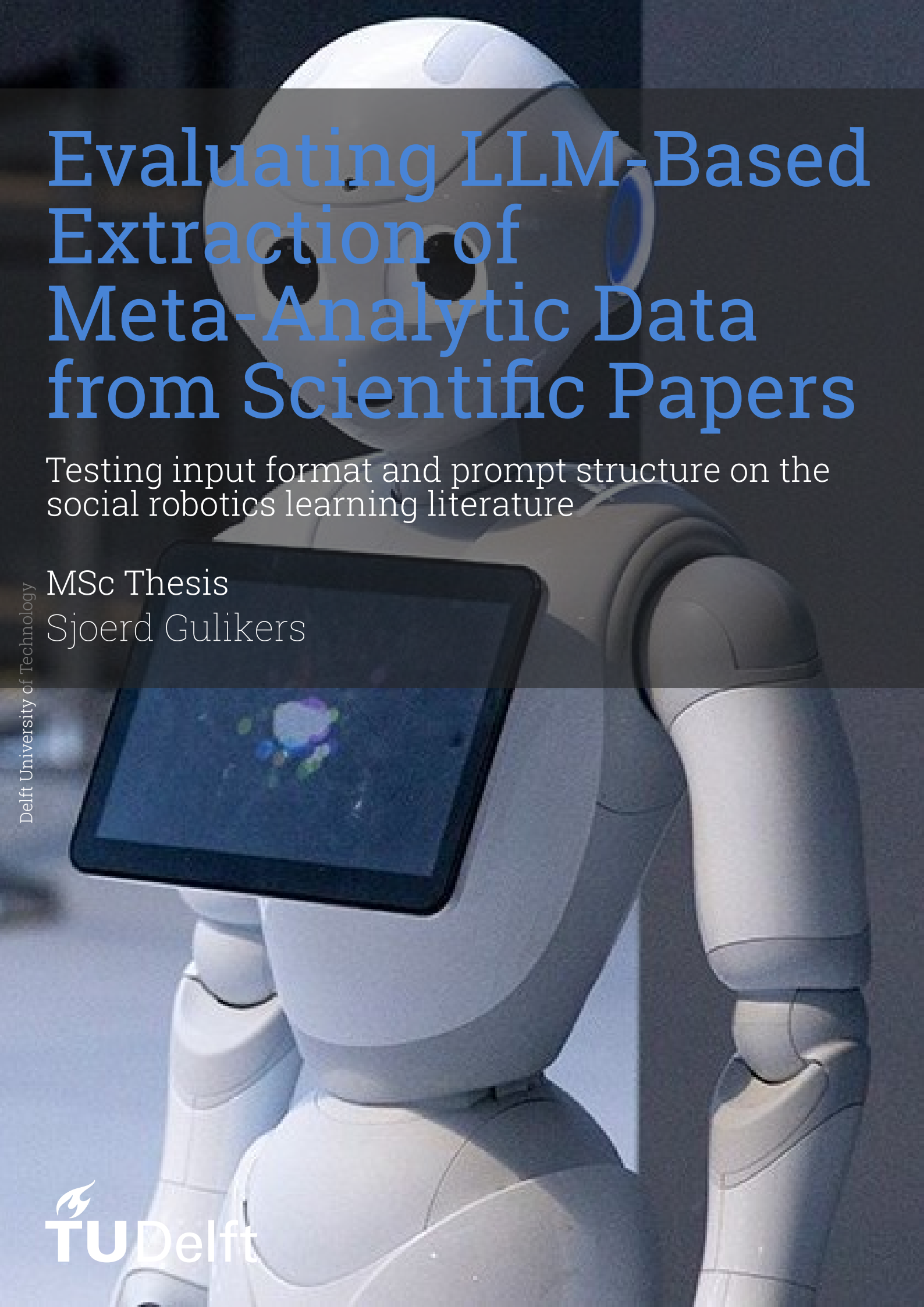

Evaluating LLM-Based Extraction of Meta-Analytic Data from Scientific Papers

Testing input format and prompt structure on the social robotics learning literature

MSc Thesis

Sjoerd Gulikers

Delft University of Technology

TU Delft

Evaluating LLM-Based Extraction of Meta-Analytic Data from Scientific Papers

Testing input format and prompt structure on the social robotics learning literature

by

Sjoerd Gulikers

Student number:	5400090
Project:	MSc thesis
Supervisor:	J. de Winter
Programme:	MSc Robotics
Department:	Cognitive Robotics
Faculty:	Mechanical Engineering, Delft University of Technology
Project duration:	June 16, 2025 – May 22, 2026

Cover:	Pepper robot image by Högskolan i Skövde, via Science Park Skövde.
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Abstract

This study evaluates recent large language models for extracting meta-analytic data from scientific articles in which relevant results are reported in text, tables, and figures. The study compares GPT-5.2, Claude Opus 4.6, Gemini 3.1 Pro Preview, and Gemini 3.1 Flash Lite Preview across three input representations: the original PDF, structured Markdown derived from the PDF, and a combined input consisting of the PDF together with cropped tables and figures. The final analyzed corpus contains 56 papers from a social robotics meta-analysis and includes 193 statistical data rows for which pre- and post-intervention values had to be extracted. Each row was scored on five numerical fields relevant for effect size calculation (Pre-Mean, Pre-SD, Post-Mean, Post-SD, and n), yielding 965 scored cells in total. Model outputs were compared against manually corrected ground-truth values from the original meta-analysis, which were verified against the source papers. Predicted rows were first matched to the corresponding ground-truth rows, after which the individual numerical values were scored using predefined numerical tolerances. This design allowed extraction errors to be interpreted not only as aggregate model failures, but also in relation to source format, input representation, target-row construction, and numerical reading. Across the full evaluation corpus, Markdown was often the strongest or near-strongest input representation overall, although the strongest representation differed across text-, table-, and figure-dominant papers. The highest overall performance was achieved by Claude Opus 4.6 with Markdown input. Papers in which the target values had to be extracted mainly from figures were the most difficult. A diagnostic follow-up on 17 papers that repeatedly showed errors in identifying the correct paper-condition combinations found that performance improved when the relevant condition was specified in advance and the model only had to extract the corresponding values. This suggests that many remaining errors were associated with identifying the correct extraction target row rather than with numerical reading alone.

Contents

1

Introduction

The concept of 'Meta-analysis', introduced by Gene Glass in 1976, provided a formal framework for the statistical integration of results from independent studies [1]. By aggregating data from multiple papers, researchers can estimate effects across a broader evidence base than is possible from a single study. The method functions as a structured interrogation of the data, requiring the researcher to rigorously extract and synthesize information scattered across text, tables, and figures to uncover patterns not visible in single studies. This was intended to move beyond purely narrative comparison and focus on observable statistics as a more systematic basis for synthesis. The Cochrane Collaboration, established in 1993, became one of the major organizations dedicated to producing and maintaining rigorous systematic reviews [2]. Meta-analysis eventually became a central method in the pursuit of cumulative knowledge in the ensuing decades.

1.1. The Extraction Bottleneck

In the years since the adoption of meta-analysis, the method has faced substantial criticism regarding its feasibility and utility. Ioannidis has highlighted the harms of the 'mass production' of systematic reviews and meta-analysis, arguing that many are redundant, misleading, and conflicted [3]. Regarding the mechanics of producing these reviews, O'Connor et al. argue that manual data extraction acts as a "rate-limiting step" in the efficient production of synthesized evidence [4]. This is because the manual process is slow and human-resource intensive, typically taking months or years to complete [5]. This challenge becomes even greater as the scientific literature continues to grow and as review methods increasingly move toward 'living systematic reviews', a newer mode of evidence synthesis that relies on continuous updates [6]. In such settings, sustaining manual extraction becomes particularly difficult without advanced automation technologies.

To address these human-resource constraints, researchers explored more constrained and task-specific software approaches for automated data extraction. However, automated extraction remained technically challenging, particularly when relevant information was not available as straightforward text. Early tools focused on simple text parsing, and they often failed when data was embedded in complex visuals. Because many crucial statistical findings are exclusively presented in plots and charts, failing to parse non-textual elements leads to substantial data loss. As Tsafnat et al. highlight, "primary information is only published in graphical form and primary data needs to be extracted from these plots as accurately as resolution permits" [7]. This limitation reveals a broader challenge for review automation: extraction

systems must handle information that is not always available as straightforward narrative text [8].

Building upon the limitations of early text-parsing tools, researchers have increasingly turned to Large Language Models (LLMs) as a potential solution for multimodal extraction. Unlike earlier software that struggled with varied linguistic structures, modern LLMs are trained on large text corpora and can often parse varied reporting styles. Recent studies have begun to evaluate LLMs for tasks such as abstract screening [9] and the extraction of textual study characteristics [10], suggesting that these models may accelerate parts of the review pipeline. The feasibility of deploying these models at a massive scale has already been demonstrated in live conference peer-review systems [11] and human-in-the-loop meta-reviewer assistants [12], emphasizing the urgency of rigorously auditing their extraction reliability [13]. Recent work has reported 93.1% weighted accuracy for data extraction when reproducing full Cochrane systematic reviews; however, the authors also noted that some data points could not be extracted when relevant information was inaccessible in supplementary files, meaning this result should not be interpreted as evidence that fully multimodal extraction from all tables, figures, and supplements has been solved [10]. Despite this progress, the application of general-purpose LLMs to scientific articles remains challenging when relevant information is distributed across data-rich visual components. While frontier models like GPT-5.2 [14], Claude Opus 4.6 [15], and Gemini 3.1 Pro [16] are positioned as advanced general-purpose systems, recent benchmarking work suggests that even frontier systems still exhibit factuality and calibration weaknesses [17]. These limitations are especially consequential in scientific extraction settings, where models must link text with visual components without hallucinating relationships [18], and where performance can degrade when they are required to jointly recover study structure and extract numerical values.

1.2. The Multimodal Challenge

Even when models can ingest PDFs directly, reconstructing hierarchical table structures can remain difficult, a problem often described as the “serialization bottleneck” [19]. Because standard transformer models process data as one-dimensional token sequences, they struggle to preserve the two-dimensional spatial logic of a grid. In tables with nested headers, multi-row spans, or unconventional layouts, one important risk is failure at “cell-to-header alignment” [20]. This can lead to structural hallucinations, where the model may accurately read a numerical value but assign it to the wrong variable or study arm [21].

Similarly, extracting data from charts and plots demands “visual reasoning” capabilities that extend beyond the Optical Character Recognition (OCR) found in standard pipelines. Benchmarks on scientific figures reveal that even advanced Multimodal LLMs (MLLMs) struggle with precise coordinate mapping. Turan and Sparks (2025) report that models such as GPT-4V and Gemini can interpret charts at a high level, but may still struggle to translate visually presented data into accurate numerical tables without external aid [22, 23].

Consequently, the optimal methodology for prompting frontier models remains unclear. It is not yet established which input representations best mitigate these bottlenecks, or whether performance improves when models are given more explicit guidance about which study conditions and outcomes should be extracted. Without understanding these boundary conditions, fully automated meta-analysis remains brittle and incomplete. Establishing these boundary conditions is highly timely, as the field is currently investigating what level of automation is practically ‘good enough’ for evidence synthesis [24], confirming that targeted LLM extraction requires careful evaluation against human expert baselines to ensure scientific validity [25].

1.3. Research Question

To address this gap, this study evaluates GPT-5.2, Claude Opus 4.6, Gemini 3.1 Pro Preview, and Gemini 3.1 Flash Lite Preview on the extraction of statistical parameters from scientific articles containing text, tables, and figures. Fundamentally, this research investigates the performance of frontier models under varying conditions of input representation, with particular attention to structurally complex cases, meaning the model must correctly identify the relevant study arms and experimental conditions before the specific numerical values can be extracted. Addressing this structural complexity is critical; recent diagnostic evaluations demonstrate that while LLMs excel at recognizing isolated entities, they suffer from severe relational binding failures when required to reconstruct complete combinations of condition, outcome, comparison, and assessment interval without explicit guidance [26]. This study employs a custom-built dataset of ground truth (GT) values derived from an existing meta-analysis [27] to test these limitations. The empirical evaluation in this study is based specifically on scientific articles about the learning effectiveness of social robotics, allowing the evaluation to be conducted within a single, thematically coherent application domain. Comparing multiple frontier model families is relevant not only because they are among the strongest currently available general-purpose systems, but also because differences in how they handle document inputs and integrate textual and visual evidence may lead to different strengths across source formats.

This study adds to prior work by making the extraction problem more diagnostic. Existing evaluations of LLM-assisted evidence synthesis often report aggregate extraction performance or evaluate broader review workflows. Such evaluations are useful, but they make it difficult to determine whether errors originate from the document representation, the reporting source, the construction of the target row, or the numerical reading step itself. The present study therefore contributes in four ways. First, it evaluates extraction at the source-paper level rather than only at the level of already summarized review outputs. Second, it separates papers by the dominant reporting source of the target values, namely text, tables, and figures. Third, it compares three document representations within the same corpus: native PDF, structured Markdown, and PDF plus automatically generated visual crops. Fourth, it introduces a target-condition extraction variant to distinguish errors caused by numerical reading from errors caused by row construction and condition linking. The contribution is therefore not only to compare model performance, but also to provide a source-level diagnostic evaluation of where meta-analytic extraction fails.

Therefore, the primary research question of this study is:

> *"How do variations in input representation affect the extraction performance of state-of-the-art Large Language Models when extracting statistical parameters from multimodal scientific literature, particularly in cases where the experimental conditions are structurally complex to identify?"*

Hypothesis: We hypothesize that specific input representations, such as structured Markdown for tables and targeted image crops for figures, will improve extraction performance across frontier models. We further expect that isolating the extraction step—by providing the model with pre-specified study conditions—will improve performance in cases where models struggle to reconstruct the study structure in one unguided attempt.

We further investigate three specific components of this problem through the following sub-questions.

First, we examine performance across different representations:

"Which input representation, Native PDF, Structured Markdown, or an explicitly multimodal

approach (PDF + image crops), yields the highest extraction performance across different data structures (Text, Tables, and Figures)?"

Hypothesis: We expect Structured Markdown to yield the highest performance for text and tables by removing layout noise, while the explicit multimodal input (incorporating object-detection crops [28]) may perform best for figures by preserving relevant visual information.

Second, we test whether providing the target condition in advance improves extraction in cases where unguided extraction fails:

"To what extent does a guided extraction prompt (where the required study conditions are pre-specified) improve performance in papers where the model fails to correctly identify the study arms in a single, unguided attempt?"

Hypothesis: A guided extraction prompt will yield higher performance in structurally complex papers, indicating that many errors in automated meta-analysis stem from misinterpreting the study's underlying experimental design rather than an inability to read the numerical data.

Finally, we evaluate model capabilities:

"How do different frontier models differ in extraction performance across representations and structurally challenging source formats?"

Hypothesis: Architectural differences will result in varying sensitivities to input representations and structurally difficult source materials, leading to differences in extraction performance across models and source formats.

2

Methods

To address the research questions on input representation, model performance, and the effect of providing pre-specified target conditions in structurally difficult cases, we designed a comparative extraction experiment. This section details the dataset construction, input-file generation, model setup, and evaluation metrics used to quantify performance.

2.1. Dataset Construction and Reporting Categories

We initially constructed a specialized evaluation corpus comprising $N = 57$ scientific publications selected from the meta-analysis by de Winter et al. [27], which originally contained 147 papers. This meta-analysis on social robotics was selected as the source domain for both practical and methodological reasons. Practically, it provided access to the original extraction table, the source papers, and expert consultation for resolving ambiguities in the GT values. Methodologically, it offered a coherent domain in which comparable statistical targets were reported across narrative text, tables, and figures. This made it possible to evaluate source-format effects while limiting variation caused by differences between scientific disciplines. Not all papers from the source meta-analysis were suitable for the present analysis. For the category-based comparison, each included paper had to be assignable to one dominant reporting source for the target numerical values, and the three source categories were initially balanced by both paper count and statistical data rows. During post-selection verification, one paper originally assigned to the Figure category was found to violate this criterion because the target values were also reported in a table. This paper was therefore excluded from all reported quantitative analyses. Because model outputs had already been generated and inspected, no replacement paper was added post hoc, to avoid outcome-informed reselection. The final analyzed corpus therefore comprised $N = 56$ papers, 193 statistical data rows, and 965 scored cells.

To construct the evaluation corpus, we screened all papers included in the meta-analysis by de Winter et al. [27] for suitability in our extraction setting. Papers were selected to populate three distinct source categories, based on whether the target numerical values were mainly reported in text, tables, or figures.

The three resulting categories are defined as follows:

2.1.1. The Text Category ($n_{papers} = 19$, $n_{rows} = 67$)

In this category, the target numerical values are reported in the narrative text of the paper. This includes both explicit cases, where values are directly stated (e.g., "The mean score was 0.5 (SD = 1.2)"), and implicit cases, where one or more required values had to be derived from other reported summary statistics such as confidence intervals (CI), standard errors (SE), or interquartile ranges (IQR).

2.1.2. The Table Category ($n_{papers} = 19$, $n_{rows} = 67$)

In this category, the target numerical values used for evaluation are located within structured tables. Correct extraction therefore depends on resolving row-column relations, headers, subheaders, grouped cells, and any accompanying table-specific context needed to interpret the reported values. In contrast to narrative text, the interpretation of a numerical value in this category depends on its position within the table structure and, in some cases, on how dispersion statistics are reported.

2.1.3. The Figure Category ($n_{papers} = 18$, $n_{rows} = 59$)

In this category, the target numerical values are presented in figures, such as bar charts or line plots. Extraction in this setting requires visual interpretation of the figure as a whole, including the plotted bars or points, axis scales, labels, legends, and value estimation. Similar to the text and table categories, some figure-based cases also required interpretation of reported summary statistics, for example when figures reported SE or CI rather than SD directly.

The unit of analysis in this experiment is the *statistical data row*, defined here as a single row containing the five target numerical fields Pre-Mean, Pre-SD, Post-Mean, Post-SD, and n, with `null` values where applicable. The original selection contained 19 papers and 67 rows per category, resulting in 57 papers, 201 statistical data rows, and 1005 scored cells before post-selection verification. After excluding one Figure-category paper that violated the single-source category rule, the final analyzed corpus contained 56 papers, 193 statistical data rows, and 965 scored cells. The Text and Table categories each contained 19 papers, 67 rows, and 335 scored cells, whereas the Figure category contained 18 papers, 59 rows, and 295 scored cells. The final category-level comparison should therefore be interpreted as slightly unbalanced, although the original selection procedure was designed to balance both paper count and row count across categories.

Papers reporting the target numerical values across multiple formats were excluded from the category-based comparison, because they could not be assigned cleanly to one dominant source type. We additionally recorded whether each included case was explicit or implicit. Explicit cases were those in which the target values were directly reported, whereas implicit cases required conversion or derivation from other reported statistics, such as SE, CI, or IQR. Because implicit reporting was relatively rare, especially in the Text and Table categories, we considered this label during selection but did not analyze implicit versus explicit reporting as a separate experimental factor in the main results.

We first identified an initial set of eligible papers in the Figure category, which was the most constrained category. This initial Figure set contained 19 papers and 67 extractable rows and served as the balancing reference for the other categories. For both the Text and Table categories, selection was performed in two steps. First, all eligible implicit papers were retained where possible, because implicit cases were rare and would otherwise have been underrepresented. Second, eligible explicit papers were added from the screened list until a set was obtained that matched the initial Figure category in both paper count and row count, namely 19 papers and 67 statistical data rows. This selection was completed before model evalua-

tion, and model performance, expected extraction difficulty, and extraction outcomes were not used as selection criteria. During later verification, one paper from the Figure category was excluded because the target values were also available in a table, which violated the category-purity rule used for this evaluation. No replacement paper was added after model outputs had been inspected, to avoid outcome-informed re-selection. Papers labelled as "could include" in the screening overview were therefore not necessarily unsuitable; they were eligible papers that were not selected once the balanced target for that category had been reached. A complete overview of all screened papers from the source meta-analysis, including their assigned category, inclusion status, and implicit/explicit labels, is provided in Appendix A.

The final dataset composition ($N_{total} = 193$ rows) is distributed as follows:

Table 2.1: Overview of Analyzed Studies by Category and Reporting Style

Category	Reporting Style	Included Studies
Text	Explicit	[29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42]
	Implicit	[43, 44, 45, 46, 47]
Tables	Explicit	[48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63]
	Implicit	[64, 65, 66]
Figures	Explicit	[67, 68, 69, 70, 71, 72, 73, 74, 75, 76]
	Implicit	[77, 78, 79, 80, 81, 82, 83, 84]

The GT values were based on the extracted values recorded in the original meta-analysis dataset [27], but these values were not treated as immutable. All included papers were checked against the source articles to verify the numerical entries and the interpretation of the extracted rows. This verification was carried out by first using the known study descriptors from the original extraction, such as condition labels, measurement names, time intervals, and robot identities, to locate the reported values in the source papers. The reported values and their locations were then checked manually against the source article, and whenever discrepancies were found, the paper was inspected in detail and the evaluation set was corrected accordingly. After the model runs, additional potential discrepancies were flagged when multiple systems produced the same answer that did not match the current GT. These cases were manually re-examined against the source papers, and values were changed only if the published article supported the correction. In total, 14 scored cells were corrected. All corrected cases are documented in Table C.4. The paper excluded after post-selection verification was not included in the quantitative scoring, category-specific analyses, statistical tests, or qualitative error counts reported below. Any discrepancies identified in that paper were therefore treated as part of the exclusion rationale rather than as GT corrections in the final analyzed benchmark.

The verified GT values in this study were used as an evaluation reference, not as information available to the models during single-pass extraction. This is different from a newly conducted meta-analysis, where no GT table exists at the start of the review. In such a setting, an LLM would more plausibly be used to draft candidate rows and values, identify possible source locations, or flag uncertain cases for human review. The present benchmark does not evaluate that full co-extraction workflow directly, but it provides a controlled way to measure the reliability of the underlying extraction step.

2.2. Extraction Design

The extraction design was structured to separate three potential sources of failure. First, models may fail because the document representation does not preserve the relevant evidence in an accessible form. This motivated the comparison between native PDF, structured Markdown, and PDF plus visual crops. Second, models may fail because the relevant values are reported in different source formats, such as narrative text, tables, or figures. This motivated the category-based corpus construction. Third, models may fail because they identify the wrong study arm, outcome, comparison, or assessment interval before numerical extraction begins. This motivated the target-condition extraction variant. The main experiment therefore evaluated single-pass LLM-based information extraction across multiple input representations, while the target-condition extraction variant was added for the subset of papers in which models repeatedly failed to select the intended rows. These design choices were intended to move beyond an aggregate performance comparison and instead diagnose where in the extraction pipeline failures occur.

2.2.1. Extraction Task and Prompt Logic

The extraction task required mapping diverse study designs into the standardized pre-post row format used in the source meta-analysis [27]. In this format, each extracted row represents a specific condition, outcome measure, and assessment interval for which the five numerical fields are recorded:

- **Unit of Extraction:** Every extracted row must represent a single experimental condition containing paired Pre-test and Post-test statistics (M, SD, n), with `null` values where a pre-test or post-test value is not applicable or not reported.
- **Handling Between-Subjects Designs:** All experimental arms were extracted as distinct rows. Distinct control groups (e.g., Human Teacher vs. No Training) were kept separate to preserve control-type granularity. When multiple robot interventions represented different experimental arms, these were also kept separate.
- **Handling Within-Subjects Designs:** Single-group pre-post comparisons were extracted by treating the specific *outcome measures* as the row-defining variables (e.g., ”Vocabulary Score” vs. ”Grammar Score”). This ensures that multiple cognitive tests administered to the same group are captured as discrete data rows.
- **Handling Longitudinal Data:** Longitudinal assessments (e.g., immediate post-test vs. delayed retention) were extracted as discrete data rows. This ensured that comparisons were made within the same assessment interval.

Single-Pass Extraction Setup

In the main single-pass setup, each model performed the extraction task in one generation step. The model had to identify which condition, outcome measure, and assessment interval should form each row, and then populate the target JSON output with the corresponding numerical values. This means that document interpretation, row selection, and numerical extraction were performed together rather than as separate sub-tasks. The same prompt instructions were used for the PDF, Markdown, and Multimodal conditions; only the provided input representation differed between conditions. The JSON schema also requested descriptive fields, such as Presence of human tutor and Source means and SDs, to guide the extraction. However, the quantitative alignment and scoring procedure used only the five numerical fields Pre-Mean, Pre-SD, Post-Mean, Post-SD, and n. The prompt template is provided in section B.2.

Exploratory prompt variants were also tested during development and yielded small per-

formance improvements in some cases. However, because these gains were limited relative to the additional methodological complexity, extensive prompt optimization was considered outside the scope of the present study.

2.2.2. Data Preprocessing & Input File Generation

To evaluate how document representation affected extraction performance, we generated three distinct input representations for every document: (1) *Raw PDF* (native PDF input), (2) *Structured Markdown* (LlamaParse-derived text and layout representation), and (3) *Multimodal* (Raw PDF plus cropped visual appendix). Prior to the model runs, all collected papers were processed in the same way to generate these inputs. Because the raw PDF already contained the original figures and tables, the Multimodal condition should be understood as the PDF plus an additional cropped visual appendix, not as an image-only condition.

1. **PDF Acquisition:** The full corpus of 56 papers was manually retrieved and stored in their native format. These files served as the input for all *Raw PDF* conditions.
2. **Extracted Markdown text (LlamaParse):** All PDF documents were processed using LlamaParse V2 [85]. We utilized the “Agentic Plus” mode with the “Agentic Specialized Chart Parsing” add-on to improve the quality of the textual representation and figure descriptions. This representation was chosen because the Markdown condition was intended to test whether an externally structured textual representation could reduce layout and serialization problems relative to native PDF input, especially for tables and figure descriptions. A simpler OCR-only or plain-text extraction pipeline would have tested whether the text was accessible, but would not have addressed the table-structure and figure-description bottlenecks that motivated the study. The Markdown condition should therefore be interpreted as a structured document-conversion condition rather than as plain text extraction. The text in these files served as the input for all *Structured Markdown* conditions.
3. **Visual Crop Generation (YOLO + SigLIP):** We employed a two-step computer vision pipeline to isolate potentially data-rich visual regions. The crop-generation pipeline was designed to create a scalable and reproducible approximation of how a multimodal extraction system might focus attention on visually relevant document regions. Manual crop selection was avoided because it would introduce researcher judgement after document inspection and would be difficult to reproduce at scale. The two-stage design was chosen because the task required two different operations: document-layout localization and semantic relevance filtering. YOLOv11 was used as the first stage because it could localize bounded document regions such as figures and tables. However, YOLO alone would retain many visual regions that were not relevant for effect-size extraction. SigLIP was therefore used as a second-stage filter because not every detected picture or table-like region contained statistical information relevant to the target fields. Conversely, using SigLIP without a layout detector would not by itself define where on the page the relevant document regions should be cropped. The combination of YOLO and SigLIP therefore separated document-layout detection from semantic relevance filtering:
    (a) **Step 1: Extraction (YOLOv11):** We used a YOLOv11 model fine-tuned on the DocLayNet dataset [86] to locate and crop tables and figures from the documents. This model targets document layout regions such as “Table” and “Picture”. Pages were rendered at 200 DPI, and a 50-pixel padding was applied to each crop to reduce the risk of cutting off relevant content at the crop boundaries. To document the expected detection performance for the relevant layout classes, we used the normalized confusion matrix reported in the model repository [86]. This reported

matrix showed a recall of 0.93 for *Picture* and 0.89 for *Table*, as shown in Figure 2.1.

(b) **Step 2: Zero-Shot Classification (SigLIP):** To filter non-informative crops, we used the SigLIP vision-language model [87], specifically `google/siglip-so400m-patch14-384`. Each crop was classified in a forced-choice setting by selecting the highest-probability label (*argmax*) from a predefined set of candidate class descriptions. Images assigned to the two data-relevant classes, "a statistical chart, graph, or boxplot with axes" and "a data table containing numerical values", were initially retained, while all remaining labels were mapped to a discard category. To evaluate this filtering step, all 863 candidate crops were subsequently reviewed and manually labeled by the researcher, and the SigLIP predictions were compared against these reference labels. This yielded an overall accuracy of 91.43%, with 298 true positives, 491 true negatives, 69 false positives, and 5 false negatives. The false-negative rate among manually relevant crops was therefore 1.65%. Crops predicted as discard but manually labeled as relevant were reinserted into the Multimodal input set, so that relevant visual evidence was not lost before downstream extraction. Crops predicted as relevant by SigLIP but manually labeled as non-relevant were retained, because no manual pruning of false positives was performed. As a result, the Multimodal condition reflects not only the original PDF representation, but also an additional visual context generated through crop detection and filtering. After filtering and reinserting false negatives, the Multimodal input contained on average 6.30 crops per paper (range: 0 to 36).

Context Construction: For the *Multimodal* conditions, the retained crops were added to the model input as a sequential visual appendix. The images were base64-encoded and appended to the API payload as discrete `input_image` blocks, positioned immediately after the textual prompt and the raw PDF file. This design also means that the Multimodal condition evaluates a complete preprocessing strategy rather than only the intrinsic visual reasoning ability of the LLM. Any improvement or lack of improvement in this condition must therefore be interpreted as the combined effect of native PDF input, crop detection, crop filtering, crop ordering, and the model's ability to use the additional visual appendix.

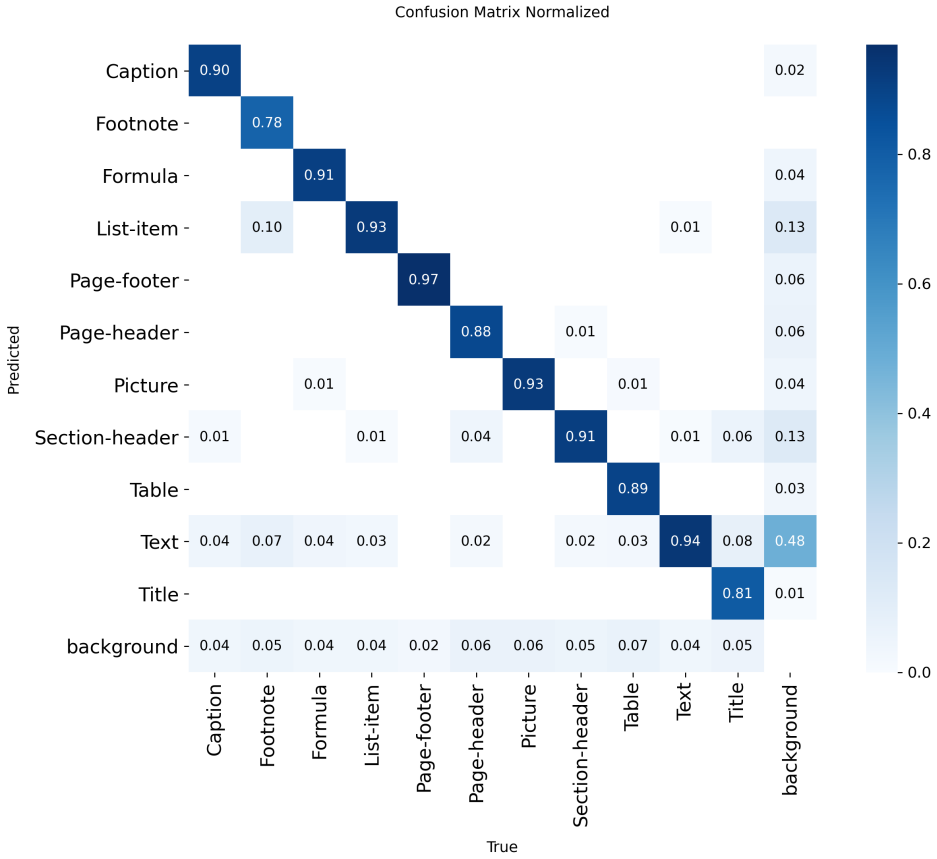

Figure 2.1: Normalized confusion matrix of the DocLayNet-trained YOLOv11 layout detector, reproduced from the model repository [86]. Columns denote true classes and rows predicted classes; diagonal entries therefore represent class-wise recall. In particular, recall is 0.93 for *Picture* and 0.89 for *Table*.

2.2.3. Model Specifications

To evaluate the performance of frontier models on the extraction task, we included four recent LLMs: *GPT-5.2* (gpt-5.2), *Claude Opus 4.6* (claude-opus-4-6), *Gemini 3.1 Pro Preview* (gemini-3.1-pro-preview), and *Gemini 3.1 Flash Lite Preview* (gemini-3.1-flash-lite-preview). Each model was run once per paper and per input representation to approximate a realistic single-run extraction scenario. This design did not assess run-to-run reliability, which is treated as a limitation of the study. In a small number of cases, a model returned no usable output, for example due to an empty response or failure to follow the required output schema. When this occurred, the identical request was repeated without any modification to the prompt, input files, crops, or model settings until a valid output was produced. These repeated calls were therefore treated as recovery from output-format failures rather than as substantive reruns under altered experimental conditions. Reasoning was enabled for all models, although the corresponding control parameter differed across providers. For GPT-5.2, we used OpenAI's `reasoning_effort` parameter with a `medium` setting. For Claude Opus 4.6, reasoning was configured through Anthropic's adaptive thinking mode using `thinking={type: "adaptive"}`, with the `effort` parameter set to `medium`. For Gemini 3.1 Pro Preview and Gemini 3.1 Flash Lite Preview, reasoning was controlled through Google's `thinking_level` parameter, using the `medium` setting. Temperature was not explicitly set for any model and was therefore kept at the provider default setting. Thus, while provider-specific interfaces differed, reasoning was enabled at a medium level across all models to reduce unnecessary variance attributable to

inference configuration rather than extraction difficulty.

2.3. Evaluation

Model outputs were evaluated in three steps: first, numerical tolerances were defined; second, predicted rows were aligned with GT rows; third, the individual cells in each aligned row were classified.

Step 1: Matching Criteria and Tolerances

Before alignment, we established numerical equivalence thresholds (δ) to determine whether a predicted value matched a GT value. The tolerance rules differed between the continuous fields (Pre-Mean, Pre-SD, Post-Mean, and Post-SD) and the sample-size field (n):

- **Text & Tables (δ_{text}):** A $\pm 1\%$ tolerance was applied to the mean and SD fields to account for minor rounding differences and small discrepancies introduced by reported precision or derived statistics.
- **Figures (δ_{vis}):** A $\pm 4\%$ tolerance was applied to the mean and SD fields to accommodate the additional uncertainty introduced when values had to be visually estimated from figures or converted from visually reported statistics.
- **Sample size (δ_n):** A stricter tolerance of $\pm 0.2\%$ was applied to n. This threshold was chosen because the evaluation target was field-level extraction correctness rather than downstream effect-size equivalence. In this setting, n was treated as a discrete participant count: two different sample sizes are different extracted facts, even if a one-participant difference may sometimes have only a small effect on the final standardized effect size or its weight. The 0.2% threshold was therefore used as a conservative near-exact rule to avoid rewarding incorrect participant counts while still keeping n within the same percentage-based matching framework as the other numerical fields. A separate threshold was necessary because n is normally reported as an integer count, whereas means and SDs can be affected by decimal rounding, reported precision, conversion from other statistics, or visual estimation. Treating n with the broader continuous-value thresholds would risk accepting different participant counts as correct, which would be inappropriate for an extraction-reliability benchmark.

To assess whether the continuous-value thresholds were reasonable, we conducted a sensitivity analysis in which the tolerance was varied from 0.1% to 20% and the resulting F1-scores were compared across source categories. As shown in Figure 2.2, performance for text and tables improved only marginally beyond 1%, indicating that larger thresholds would provide little additional benefit while increasing the risk of rewarding substantively incorrect values. For figures, performance improved more substantially at lower thresholds and began to level off around 4% to 5%. We therefore retained $\pm 1\%$ for text and tables and selected $\pm 4\%$ for figures as a conservative compromise between robustness to visual estimation error and discriminative validity. The stricter sample-size threshold was retained separately to avoid treating different participant counts as correct under the broader continuous-value thresholds. Because the tolerance was percentage-based and applied separately to each numerical field, smaller values such as SDs had narrower absolute tolerance bands than larger mean values. This should be considered when interpreting errors in figure-based SD extraction.

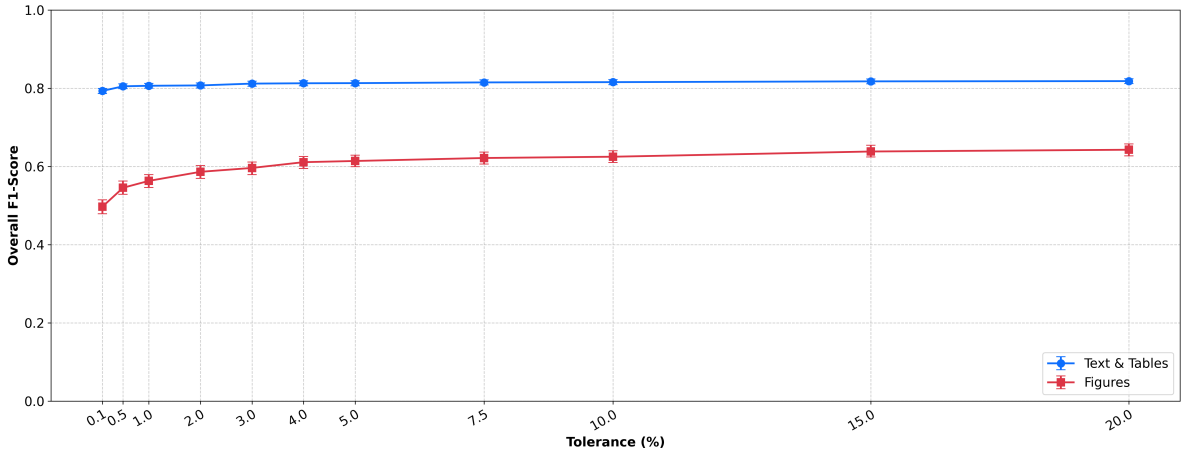

Figure 2.2: Sensitivity analysis of the numeric matching tolerance, computed across all single-pass outputs from the four models and three input representations. The tolerance was varied from 0.1% to 20%, and the resulting overall F1-scores were computed separately for text & tables and for figures. Performance for text & tables stabilizes quickly, whereas figure-based extraction shows larger gains at lower thresholds before leveling off.

2.3.1. Step 2: Alignment Algorithm

Because the model generated an unordered list of data rows, we implemented a *matching algorithm* to align predicted rows with GT rows before cell-level scoring. This was necessary because model outputs were not guaranteed to preserve the same row order or row count as the GT. A strict row-by-row comparison would therefore incorrectly penalize correct numerical extractions that appeared in a different order.

Several alternative alignment strategies were considered but were less suitable for the present evaluation. A strict row-order comparison was inappropriate because the models were not instructed to preserve the GT row order and often produced valid rows in a different sequence. Matching primarily on descriptive labels was also problematic, because equivalent conditions, outcomes, and assessment intervals could be described using different wording across model outputs. Fully manual alignment would have reduced this problem, but would also have introduced researcher judgement into the quantitative scoring procedure. The staged numerical matching procedure was therefore chosen as a transparent and reproducible compromise.

The alignment used only the five numerical fields required for effect-size calculation: Pre-Mean, Pre-SD, Post-Mean, Post-SD, and n. Descriptive fields such as condition labels, outcome names, and other study descriptors were not used for quantitative matching. This focused the scoring on the numerical extraction task, but it also means that a row could be aligned based on numerical overlap even if the descriptive condition label was imperfect. The alignment procedure should therefore be interpreted as an automated scoring rule rather than as a manually adjudicated best-case alignment.

The algorithm used five matching levels. At each level, it scanned the pool of remaining unmatched predictions against the remaining GT rows. A pair was provisionally aligned if it satisfied the level criterion, and both rows were removed from the pool before the next level began:

- **Level 1 (5-field match):** All 5 attributes (Pre-Mean, Pre-SD, Post-Mean, Post-SD, n) match within the defined tolerance δ.
- **Level 2 (4-field match):** Exactly 4 of the 5 attributes match.

- **Level 3 (3-field match):** Exactly 3 of the 5 attributes match.
- **Level 4 (2-field match):** Exactly 2 of the 5 attributes match.
- **Level 5 (1-field match):** Exactly 1 of the 5 attributes matches.

This five-stage procedure was used only for row alignment. Final scoring validity was determined in the subsequent classification step. In particular, provisional one-field alignments were not treated as valid scored row matches, because a single shared value, such as a repeated sample size, could occur by coincidence. This limited the risk that weak numerical overlap would be rewarded as a correct extraction, while preserving an automated and reproducible evaluation procedure.

2.3.2. Step 3: Classification Logic (Cell-Level)

Once the alignment phase was complete, we performed a direct cell-by-cell comparison for each provisionally aligned row pair. To prevent weak numerical overlaps from being treated as valid row matches, we imposed a minimum validity threshold: only aligned row pairs with at least 2 matching attributes were treated as valid matches for scoring purposes. This threshold was chosen because a single matching value, especially a common sample size or a repeated mean, could occur by coincidence and would provide insufficient evidence that the predicted row and GT row referred to the same extraction target. Row pairs aligned only at the 1-of-5 level were therefore rejected and treated as unmatched residuals.

Accordingly, the data were classified as follows:

- **Matched Rows:** Provisionally aligned row pairs with at least 2 matching attributes were scored through direct cell-by-cell comparison.
- **Unmatched Predictions:** Predicted rows with no valid aligned GT counterpart were treated as spurious predicted rows, including rows rejected after a 1-of-5 alignment.
- **Unmatched GT Rows:** GT rows with no valid aligned prediction were treated as missed target rows.

The specific classification definitions were as follows:

- **True Positive (TP):** The model predicted a numerical value within an *aligned* row that falls within the allowed tolerance (δ) of the corresponding GT value.
- **True Negative (TN):** The model correctly identified a value as missing (outputting `null`) within an *aligned* row where the GT also contains no value.
- **False Positive (FP):** A predicted value is penalized as a False Positive in two scenarios:
  1. **Inaccurate Value:** In an aligned row, the model predicts a value that falls outside the tolerance δ or predicts a value where the GT is missing.
  2. **Spurious Prediction:** All numerical values contained within an unaligned predicted row are classified as FPs, regardless of coincidental numerical overlaps.
- **False Negative (FN):** A target value is penalized as a False Negative in two scenarios:
  1. **Missing Value:** In an aligned row, the model fails to extract a value (outputting `null`) where a value exists in the GT.
  2. **Missed Row:** All numerical values contained within an unmatched GT row are classified as FNs.

Metrics: We calculated the F1-score (harmonic mean of Precision and Recall) based on these cell-level counts. We also computed Accuracy, Precision, and Recall, which are reported in the Results section alongside F1. F1 was treated as the primary metric because it penalizes the model both for missing target values and for generating incorrect ones. By contrast, Accuracy should be interpreted more cautiously, because null-null matches, such as pre-test values in post-only designs, can increase Accuracy even though they do not contribute to Precision, Recall, or F1. In addition to reporting these quantitative metrics, we manually inspected papers with low or unexpected scores to identify underlying failure modes, such as row overgeneration or misinterpretation of derived statistics. Here, row overgeneration refers to cases in which the model produced redundant or extra rows beyond the target structure. These were penalized quantitatively through the false-positive counts assigned to unmatched predicted rows. This qualitative error analysis informed the interpretation of the final results. For figure-based cases classified as genuine model extraction failures, we performed an additional diagnostic review of SD-related errors. This review checked whether the source figure or accompanying text actually reported SD values, or whether the visible error bars instead represented another dispersion statistic such as SE or 95% CI. We also inspected whether models left SD fields empty, attempted to infer SD values from the figure, or appeared to extract a value without verifying what the error bars represented. This additional review was used only to interpret the qualitative error patterns and was not used to change the quantitative scoring procedure.

2.3.3. Target-Condition Extraction Variant

During evaluation, we observed that a subset of papers produced persistent errors across models because models selected rows outside the intended meta-analytic target structure before numerical extraction could be meaningfully assessed. In these cases, models frequently generated too many rows relative to the target rows used in the GT. To diagnose whether these errors were mainly caused by row selection rather than by reading the numerical values, we created an additional GT representation for these papers that contained an explicit `target_condition` field. This field specified the intended condition, outcome, comparison, and, where relevant, assessment interval for each target row. The papers were then rerun with an extraction-only prompt in which the model was asked to extract values only for these specified target conditions. This variant therefore removed part of the original row-selection task and should be interpreted as a diagnostic comparison, not as the same task under a shorter or merely simplified prompt. The purpose of this variant was not to simulate a fully automated meta-analysis workflow, because the target rows have already been specified in advance. Instead, it was used to isolate whether failures in the single-pass setup were caused mainly by selecting the wrong target condition or by failing to read the numerical values once the target condition was known. This distinction is important because the two failure modes imply different practical solutions: row-selection errors require better study-structure modelling or human guidance, whereas numerical-reading errors require better table parsing, figure interpretation, or value conversion. The papers included in this targeted rerun were: [29, 30, 35, 38, 47, 41, 52, 56, 57, 58, 78, 73, 74, 75, 76, 46, 62]. The dedicated prompt used for this variant is provided in section B.3.

2.3.4. Statistical Analysis

To assess whether observed performance differences between input conditions were statistically significant, we accounted for the hierarchical structure of the data. Individual scored cells, namely Pre-Mean, Pre-SD, Post-Mean, Post-SD, and n, are nested within extracted rows, which are further nested within papers. Across the final analyzed dataset, this corresponds to 193 rows and 965 scored cells. Standard tests of independence are therefore inappropriate

because they would ignore within-paper correlation.

Accordingly, we applied two complementary statistical tests at different levels of analysis:

1. **Paper-Level Analysis (Wilcoxon Signed-Rank Test):** Because the primary performance metric of this study is the F1-score, the main inferential analysis was conducted at the paper level. For each paper and input condition, we computed a paper-level F1-score based on the underlying cell-level classification results. Paired differences in F1-scores between conditions were then evaluated using the Wilcoxon signed-rank test.
2. **Cell-Level Analysis (Obuchowski's Adjusted McNemar's Test):** As a secondary robustness check, we evaluated binary extraction success versus failure at the cell level using the adjusted McNemar's test for clustered matched-pair data [88]. Unlike the standard McNemar test, this method accounts for intra-cluster correlation among scored cells within the same paper through a variance inflation adjustment. This prevents underestimation of uncertainty and inflation of Type I error rates when comparing paired extraction outcomes at a finer level of granularity.

For the single-pass experiment, these analyses were used to compare paired model-input representation conditions across the final set of 56 analyzed papers. For the target-condition extraction variant, results were analyzed separately on the rerun subset after applying the same paper-exclusion rule.

3

Results

This section reports the results in the order of the research questions. Each subsection begins with a brief answer to the corresponding research question, followed by the supporting evidence. Supplementary visualizations, category-level result tables, full numerical input representation comparisons, and detailed qualitative case inventories are reported in Appendix C.

3.1. RQ1: Effects of Input Representation

Across the complete dataset, input representation influenced extraction performance, but the effect was not uniform across models or source categories. Overall, Markdown was the strongest or near-strongest input representation for most model groups, while figure-based papers remained the most difficult source category across all settings.

To examine how input representation affected extraction performance, all 12 single-pass model-input representation conditions were evaluated on the final analyzed dataset of 56 papers, 193 rows, and 965 scored cells. This dataset reflects the exclusion of one initially selected Figure-category paper that violated the single-source category rule because the target values were also available in a table. Table 3.1 reports the aggregate Accuracy, Precision, Recall, and F1-score with 95% confidence interval for each condition.

Table 3.1: Overall performance across all single-pass model-input representation conditions on the complete dataset, with 95% confidence intervals for the F1-scores.

Condition	Accuracy	Precision	Recall	F1 (95% CI)
GPT-5.2 PDF	0.5766	0.6073	0.8350	0.7032 [0.6791, 0.7270]
GPT-5.2 Markdown	0.6390	0.6575	0.8905	0.7565 [0.7352, 0.7769]
GPT-5.2 Multimodal	0.5894	0.6100	0.8625	0.7146 [0.6907, 0.7370]
Claude Opus 4.6 PDF	0.6557	0.7075	0.8414	0.7687 [0.7457, 0.7919]
Claude Opus 4.6 Markdown	0.7134	0.7607	0.8694	0.8114 [0.7903, 0.8310]
Claude Opus 4.6 Multimodal	0.7037	0.7392	0.8856	0.8058 [0.7852, 0.8260]
Gemini 3.1 Pro PDF	0.6488	0.6742	0.8768	0.7623 [0.7397, 0.7841]
Gemini 3.1 Pro Markdown	0.6645	0.6873	0.8853	0.7738 [0.7541, 0.7937]
Gemini 3.1 Pro Multimodal	0.6533	0.6832	0.8773	0.7682 [0.7463, 0.7889]
Gemini 3.1 Flash Lite Preview PDF	0.6229	0.6783	0.8094	0.7381 [0.7153, 0.7631]
Gemini 3.1 Flash Lite Preview Markdown	0.6373	0.6641	0.8794	0.7567 [0.7337, 0.7782]
Gemini 3.1 Flash Lite Preview Multimodal	0.6641	0.7003	0.8650	0.7740 [0.7512, 0.7964]

The highest overall F1-score was obtained by Claude Opus 4.6 with Markdown input ($F1 = 0.8114$, 95% CI: [0.7903, 0.8310]), followed by Claude Opus 4.6 with Multimodal input ($F1 = 0.8058$, 95% CI: [0.7852, 0.8260]) and Gemini 3.1 Flash Lite Preview with Multimodal input ($F1 = 0.7740$, 95% CI: [0.7512, 0.7964]). A full ranked visualization of these overall F1-scores is provided in Appendix C (Figure C.1).

To determine whether input representation effects differed across source types, performance was also evaluated separately for the text, table, and figure categories. The complete category-specific F1-scores for all 12 conditions are reported in Appendix C (Table C.1), while Figure 3.1 visualizes the category-level patterns.

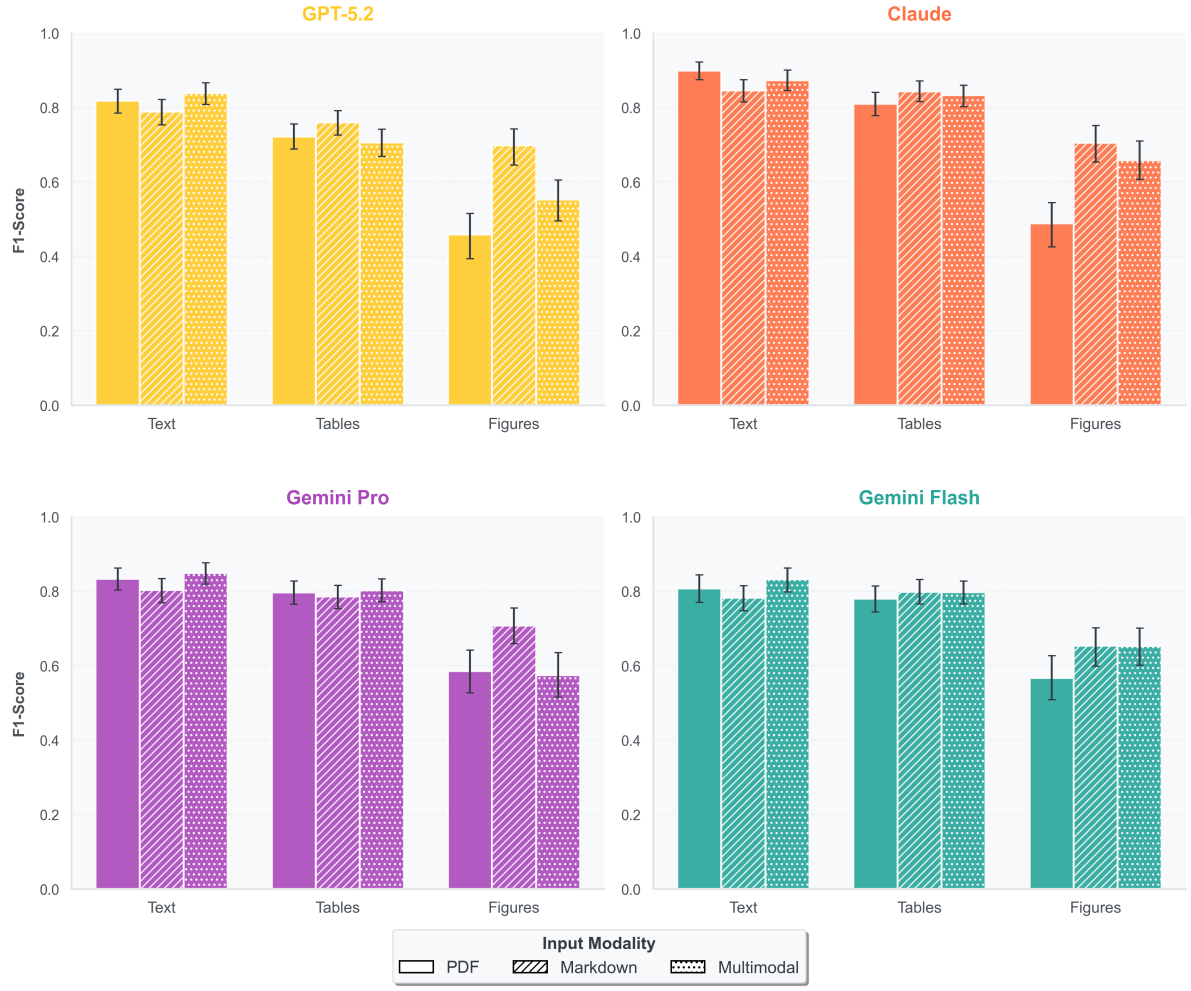

Figure 3.1: Category-specific F1-scores for each model across the PDF, Markdown, and Multimodal input representations. Error bars denote 95% confidence intervals. The Text and Table categories each contain 19 papers, 67 statistical data rows, and 335 scored cells; after exclusion of one source-category-ineligible paper, the Figure category contains 18 papers, 59 statistical data rows, and 295 scored cells.

Three patterns stand out. First, text and table papers were generally extracted more accurately than figure papers. Second, figure-category performance showed the largest spread across conditions, indicating that source representations containing charts, plots, or visually dense material remained the most challenging. Third, the strongest model-input representation varied by category: the best text-category score was obtained by Claude Opus 4.6 PDF ($F1 = 0.8992$, 95% CI: [0.8748, 0.9224]), the best table-category score by Claude Opus 4.6

Markdown ($F1 = 0.8431$, 95% CI: [0.8160, 0.8712]), and the best figure-category score by Gemini 3.1 Pro Markdown ($F1 = 0.7075$, 95% CI: [0.6589, 0.7547]).

3.2. RQ2: Effects of Providing Target Conditions in Structurally Difficult Cases

For RQ2, the relevant result is the comparison between the original single-pass setting and the target-condition extraction variant on the 17-paper rerun subset. In this subset, the single-pass F1-scores were low across all model-input representation combinations, ranging from 0.3969 to 0.4841 before the target conditions were provided.

Table 3.2 reports the comparison between single-pass and target-condition extraction F1-scores for each model-input representation combination on this subset.

Table 3.2: Comparison of single-pass and target-condition extraction performance on the 17-paper rerun subset, with 95% confidence intervals for the F1-scores.

Model	input representation	Single-pass F1 (95% CI)	Target-condition F1 (95% CI)	△F1
GPT-5.2	PDF	0.4140 [0.3623, 0.4645]	0.8698 [0.8307, 0.9085]	+0.4558
GPT-5.2	Markdown	0.4302 [0.3718, 0.4828]	0.8203 [0.7703, 0.8640]	+0.3902
GPT-5.2	Multimodal	0.4370 [0.3824, 0.4910]	0.8662 [0.8235, 0.9028]	+0.4292
Claude Opus 4.6	PDF	0.4776 [0.4206, 0.5353]	0.8875 [0.8479, 0.9231]	+0.4099
Claude Opus 4.6	Markdown	0.4841 [0.4257, 0.5374]	0.8396 [0.7926, 0.8824]	+0.3555
Claude Opus 4.6	Multimodal	0.4771 [0.4202, 0.5298]	0.8910 [0.8535, 0.9250]	+0.4138
Gemini 3.1 Pro	PDF	0.4030 [0.3471, 0.4587]	0.8875 [0.8503, 0.9212]	+0.4845
Gemini 3.1 Pro	Markdown	0.3969 [0.3476, 0.4473]	0.7986 [0.7396, 0.8446]	+0.4016
Gemini 3.1 Pro	Multimodal	0.4330 [0.3779, 0.4824]	0.8328 [0.7859, 0.8746]	+0.3998
Gemini 3.1 Flash Lite Preview	PDF	0.3981 [0.3342, 0.4550]	0.8133 [0.7665, 0.8599]	+0.4152
Gemini 3.1 Flash Lite Preview	Markdown	0.4384 [0.3836, 0.4950]	0.7726 [0.7143, 0.8251]	+0.3341
Gemini 3.1 Flash Lite Preview	Multimodal	0.4706 [0.4147, 0.5235]	0.8173 [0.7682, 0.8626]	+0.3467

After the target conditions were provided, F1 increased in all 12 model-input representation combinations. Improvements ranged from $+0.3341$ to $+0.4845$ F1. The largest absolute gain was observed for Gemini 3.1 Pro PDF, while the smallest improvement was observed for Gemini 3.1 Flash Lite Preview Markdown. A supplementary visualization of this comparison is provided in Appendix C (Figure C.2).

A second analysis restricted the data to papers where the model predicted exactly the same number of rows as the GT. The exact row-count subset showed a similar pattern. Depending on the model-input representation condition, exact row-count agreement with the GT was achieved for 31 to 41 of the 56 papers. Table 3.3 reports the conditional F1-score for these papers, together with the number of papers retained in each condition. The full-dataset F1 point estimate from Table 3.1 is repeated only as a reference value.

Table 3.3: Conditional F1-score restricted to papers with exact row-count agreement, with 95% confidence intervals for the F1-scores. The $\triangle$F1 column reports the difference between the conditional F1-score and the corresponding full-dataset F1-score from Table 3.1.

Model	input representation	Conditional F1 (95% CI)	Matched Papers	$\triangle$F1
GPT-5.2	PDF	0.8997 [0.8781, 0.9183]	32/56	+0.1965
GPT-5.2	Markdown	0.9437 [0.9294, 0.9568]	32/56	+0.1872
GPT-5.2	Multimodal	0.9319 [0.9152, 0.9477]	31/56	+0.2174
Claude Opus 4.6	PDF	0.9024 [0.8853, 0.9191]	40/56	+0.1337
Claude Opus 4.6	Markdown	0.9364 [0.9232, 0.9480]	41/56	+0.1250
Claude Opus 4.6	Multimodal	0.9381 [0.9245, 0.9517]	41/56	+0.1322
Gemini 3.1 Pro	PDF	0.9210 [0.9048, 0.9358]	40/56	+0.1587
Gemini 3.1 Pro	Markdown	0.9444 [0.9312, 0.9572]	34/56	+0.1706
Gemini 3.1 Pro	Multimodal	0.9174 [0.9009, 0.9325]	40/56	+0.1492
Gemini 3.1 Flash Lite Preview	PDF	0.8660 [0.8447, 0.8876]	38/56	+0.1280
Gemini 3.1 Flash Lite Preview	Markdown	0.9089 [0.8925, 0.9258]	37/56	+0.1522
Gemini 3.1 Flash Lite Preview	Multimodal	0.9041 [0.8865, 0.9208]	40/56	+0.1301

For all 12 conditions, the conditional F1-score was higher than the corresponding full-dataset F1-score. Conditional F1-scores ranged from 0.8660 to 0.9444, compared with full-dataset F1-scores ranging from 0.7032 to 0.8114. This comparison should not be interpreted as an unbiased estimate of the gain from correct row construction, because the conditional scores are based only on papers for which exact row-count agreement was already achieved. Rather, the result shows that extraction performance was substantially higher in the subset of cases where the model produced the same number of rows as the GT. A supplementary visualization of this comparison is provided in Appendix C (Figure C.3).

3.3. RQ3: Differences Between Frontier Models

The results for RQ3 show that the frontier models were not equivalent in extraction performance. Claude Opus 4.6 was the most consistent top-performing model across the dataset, especially with Markdown and Multimodal inputs, while Gemini 3.1 Pro was competitive overall and achieved the strongest figure-category result when paired with Markdown. Gemini 3.1 Flash Lite Preview performed competitively in several settings, particularly with Multimodal input, but was less consistent across input representations. GPT-5.2 improved clearly under Markdown input relative to PDF, but did not outperform Claude Opus 4.6 on the complete dataset.

3.4. Statistical Comparisons Between Input Representations

The within-model input representation comparisons provided limited inferential support for input representation differences. The full numerical results are reported in Appendix C, and a supplementary heatmap is provided there as well (Figure C.4).

At the paper level, none of the pairwise comparisons reached the conventional threshold of $p < 0.05$. At the cell level, nominally significant differences were observed for Claude Opus 4.6 PDF versus Markdown ($p = 0.03284$), Claude Opus 4.6 PDF versus Multimodal ($p = 0.01439$), and Gemini 3.1 Flash Lite Preview PDF versus Multimodal ($p = 0.02214$). Because three pairwise input representation comparisons were conducted within each model group, we also considered a Bonferroni-adjusted significance threshold of $\alpha = 0.0167$. Under this correction, only the Claude Opus 4.6 PDF versus Multimodal comparison remained statistically significant at the cell level, whereas the other two nominally significant comparisons no longer met the

adjusted threshold.

3.5. Qualitative Error Analysis

The qualitative review showed that low-scoring cases were not attributable to a single failure source. Instead, four recurring issue sources were identified: GT issues, paper inconsistencies, task-definition or paper-structure ambiguities, and genuine model extraction failures.

During this review, papers were first checked for whether some model outputs had already extracted them fully correctly. A paper was treated as extractable if at least three model-input representation conditions recovered every scored value correctly. In total, this applied to $n = 20$ papers. These papers were therefore not counted as genuine model extraction failures. The issue categories were not always mutually exclusive: in some papers, genuine extraction failures co-occurred with task-definition ambiguities, paper inconsistencies, or GT issues. As a result, individual cases could reflect a combination of difficulty sources rather than a single isolated cause. The excluded Figure-category paper was not counted in this qualitative overview, because its issue concerned source-category eligibility rather than model extraction performance within the final analyzed benchmark.

Table 3.4: Summary of qualitative issue sources identified during manual review.

Issue source	Studies	Description
GT issues	8	The GT did not contain the correct value, for example because a reported value such as SD_{post}, n, or a row assignment was incorrect.
Paper inconsistencies	2	The source paper itself contained inconsistencies, for example when the same sample size was reported differently at two points in the paper.
Task-definition / paper-structure ambiguities	17	The extracted output contained too many rows, reflecting ambiguity in how the paper structure should be mapped onto the GT row structure.
Genuine model extraction failures	18	The model had difficulty extracting the target values, resulting in missing values or incorrect extracted values.

The qualitative review suggests that not all extraction errors should be interpreted as model failures. Some cases reflected GT corrections, while others reflected ambiguity in the mapping from the source paper to the target rows or inconsistencies in the source paper itself. Among the papers classified as genuine model extraction failures, 15 occurred in the Figure category (83%), showing that these failures were strongly concentrated in cases where the target values had to be recovered from visual material. Missing SD values were the most common reviewed problem within these cases, especially in visually noisy or figure-heavy papers.

Because error bars do not always represent SDs, the figure-based SD mismatches were inspected in more detail. This audit focused on the 10 figure-category papers for which at least one model-input representation condition attempted to extract or infer one or more SD values and where SD-related mismatches remained. The remaining figure-category papers were not included in this SD-mismatch audit for three reasons: one paper was labelled as a paper inconsistency [67], three papers were already treated as extractable because at least three model-input representation conditions recovered all scored values correctly [69, 71, 82],

and four papers were not informative for SD-mismatch analysis because no model-input representation condition attempted to extract SD values [68, 80, 74, 75]. The audit therefore specifically examined cases in which at least one model-input representation condition made an SD-related extraction attempt, so that the remaining mismatches could be interpreted in terms of visual estimation, error-bar interpretation, conversion to SD, or target-condition linking.

This audit showed that the mismatches were not caused by a single mechanism. In some cases, the relevant statistic was clearly communicated but required an additional conversion step, for example from 95% CI, SE, or IQR to SD. In these cases, a small visual interpolation error could be amplified after conversion to SD, producing a larger absolute difference in the derived SD even when the originally read value was only slightly inaccurate. Because the scoring used a percentage-based tolerance, this conversion did not by itself change whether a value was classified as correct or incorrect when the same proportional error was preserved. However, these cases show why a fixed absolute tolerance would be inappropriate. Other cases reflected different issues, such as extracting values from the wrong figure, relying on an easier but incorrect learning-gain value, cautious non-extraction of very thin error bars, or incorrect target-condition linking. The full paper-level audit is reported in Appendix C (Table C.7).

To make the category of genuine model extraction failures more concrete, four illustrative papers are discussed below. These cases were selected because the required values were in principle extractable for the target rows used in this study, yet extraction still failed in one or more conditions.

3.5.1. Illustrative cases of genuine model extraction failure

Together, the examples show why the genuine model extraction failure category should not be reduced to one explanation. Some failures involved visual estimation and error-bar interpretation, whereas others involved statistical transformation, subgroup aggregation, or maintaining the correct relation between a value and the intended target row.

Paper 43 (Zhexenova et al., 2020) [65]

Paper 43 illustrates a non-visual genuine extraction failure, namely failure in combining subgroup statistics into the target condition. In this case, the boys and girls subgroups reported in the table had to be merged to match the GT row structure. At least three model-input representation conditions performed this merge correctly, showing that the target row was recoverable in the evaluated setup. However, other conditions had particular difficulty merging the SD values correctly, even when the relevant subgroup rows had been identified. This suggests that the main problem was not locating the relevant table entries, but correctly aggregating subgroup dispersion information into a single value matching the GT row. The case is important because it shows that extraction errors do not always stem from missing or misread values. They can also arise when the model identifies the correct source rows, but fails to transform them properly into the required target-row format.

Paper 44 (Kanda et al., 2012) [78]

Paper 44 represents a genuine extraction failure because the target values were reported in a form that matched the target-row definition, and at least three model-input representation conditions recovered the target row correctly. The main extraction problem in this case was figure-based recovery of the SD values. While the extracted mean was close to the target, it still fell outside the tolerance, and the SD values were not recovered correctly. A likely reason is that the relevant figure was visually small and reported values only at limited precision,

which made the central values somewhat difficult to estimate and the error bars even harder to interpret, as illustrated in Figure 3.2. This paper is illustrative because it shows an important asymmetry in the figure results: even when the bar heights could still be approximated reasonably well, the error bars were much more frequently interpreted incorrectly. The case therefore points to a specific weakness in extracting dispersion information from figures rather than a general inability to identify the correct outcome.

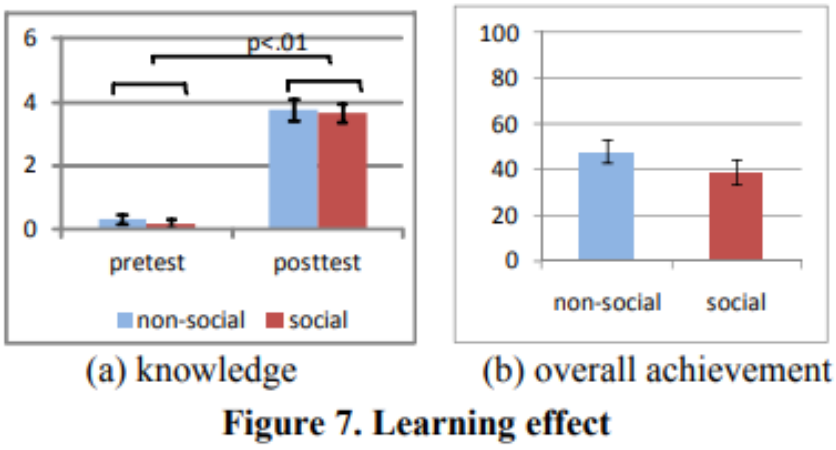

Figure 3.2: Relevant figure from Paper 44 illustrating the extraction difficulty. The central values must be approximated visually from the bar heights, but the error bars are especially small and imprecise, which likely contributed to incorrect extraction of the SD values.

Paper 46 (Levinson et al., 2021) [36]
A further example is provided by Paper 46, in which the relevant values were available but extraction remained unreliable in part of the evaluated setup. The main problem in this case was row assignment rather than number reading: a pre-test value should have been applied to two target rows, but models often assigned it to only one of them. Since at least three conditions recovered the target values correctly, the case was not inherently unextractable. Rather, it reflects a failure to preserve the correct row structure once the relevant values had been located. This example is important because it shows that extraction errors can arise even when the source statistics themselves are identifiable. In such cases, the bottleneck is not the recognition of the number, but the maintenance of the correct relationship between condition, assessment interval, and target row.

Paper 75 (Suzuki and Kanoh, 2017) [73]
Paper 75 illustrates that the qualitative issue categories could overlap. The paper was initially labelled as a task-definition or paper-structure ambiguity, because the single-pass outputs often selected values that did not correspond to the target structure used in the GT. However, even after the target conditions were provided in the target-condition extraction variant, extraction remained poor. This indicates that the case was not only a row-selection problem, but also a genuine model extraction failure.

In this case, the extraction problem was not limited to the SD values. Both the mean values and the SD values had to be inferred visually from the figure, and both were extracted unreliably, as illustrated in Figure 3.3. This paper therefore represents a more difficult figure-reading case in which the visual interpretation problem extended beyond the error bars alone. Even so, the failure on the SD values remains consistent with the broader pattern observed across the figure-category model failures, namely that error bars were the least reliably interpreted

part of the figure.

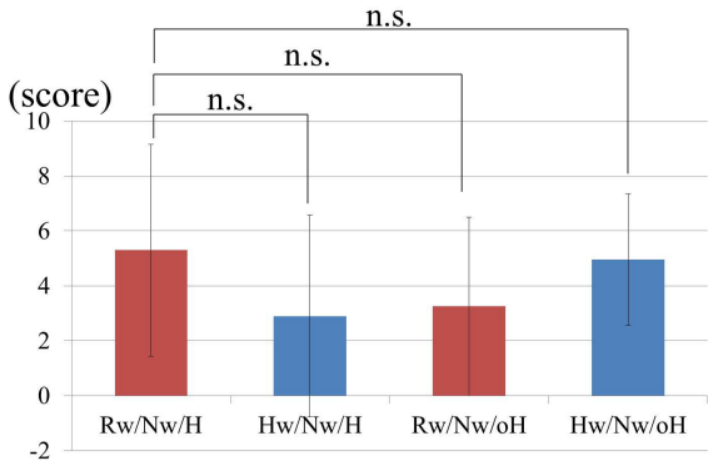

Fig. 12. Comparison of the four groups with nodding

Figure 3.3: Relevant figure from Paper 75 illustrating a difficult figure-reading case. Both the central values and the error bars must be inferred visually, which likely contributed to incorrect extraction of both the mean and SD values.

4

Discussion

Overall, extraction quality varied with input representation, source type, and model. In the Multimodal setting, it also depended on the crop-generation pipeline. In this study, Multimodal therefore refers to the native PDF plus a cropped visual appendix generated through YOLO-based crop detection and SigLIP-based filtering, rather than to an image-only alternative to the PDF. Markdown performed strongly in many settings, row-construction failures remained a major source of error, and a non-trivial portion of low-scoring cases was traceable to GT issues, source-paper inconsistencies, or ambiguities in how the evaluated rows mapped onto the reported study structure rather than to pure model failure. The relatively stable recall across conditions suggests that the models retrieved a comparable share of the GT values, whereas the larger variation in precision indicates that performance differences were driven more by false positive extractions and incorrect structural assignments than by simple omission of values. Across the full evaluation corpus, Markdown was often the strongest or near-strongest representation, figure papers remained the most difficult category, and performance improved strongly when models were also given the target condition in advance. Taken together, these findings suggest that much of the remaining error may reflect row construction and study-structure interpretation rather than numerical reading alone.

4.1. Answering the research questions

For the *main research question*, the answer is therefore twofold. First, the results suggest that input representation is associated with differences in extraction accuracy, but not in a uniform or model-independent way. Second, performance improved strongly when the relevant target condition was provided in advance, meaning that the model no longer had to decide which study arm, outcome, comparison, or assessment interval formed each target row. Together with the conditional analysis based on exact row-count agreement, this indicates that an important part of the extraction difficulty lay in structural recovery rather than in reading individual values.

For *RQ1*, no single input representation was universally best across all source types. At the aggregate level, Claude Opus 4.6 with Markdown input achieved the strongest overall result, but the category-level pattern was more mixed: the strongest text-category result came from PDF input, the strongest table-category result from Markdown input, and the strongest figure-category result from Markdown input. The answer to RQ1 is therefore that input representation mattered descriptively, but there was no universal best input representation across text, tables, and figures. The strong performance on text-dominant papers is also not surpris-

ing, since explicitly reported numerical results are generally easier for current LLMs to interpret than values that must be visually estimated from figures or reconstructed from more complex table structure.

A plausible explanation for the strong Markdown results is that Markdown may reduce layout noise and present symbolic content in a more linearized form, which could make tables and semi-structured reporting easier for language models to parse [19, 20]. Importantly, however, the Markdown condition in this study should not be interpreted as plain extracted text. The Markdown files were generated using LlamaParse in "Agentic Plus" mode with the "Agentic Specialized Chart Parsing" add-on, meaning that tables were restructured into a more linearized textual format and figures could be accompanied by generated descriptive summaries. Its strong performance may partly reflect this upstream conversion of tabular and visual content into text, not only the intrinsic advantage of Markdown as a file format. This is also relevant for interpreting the Multimodal condition: because the original PDF already contained the figures and tables, the Multimodal condition tested whether an additional cropped visual appendix improved extraction beyond the native PDF representation. The weaker-than-expected advantage of this condition suggests that simply appending detected crops is not necessarily sufficient to solve figure-based extraction. At the same time, figure-based papers remained the hardest category across all conditions, indicating that visual interpolation, legend resolution, and proxy-statistic interpretation remained major bottlenecks.

For *RQ2*, the evidence is clearest on the targeted rerun subset. The important interpretation is not that a shorter prompt was inherently better, but that performance improved when the model was also given the intended target condition. In this setting, part of the original task was removed: the model no longer had to infer which condition, outcome, comparison, or assessment interval should define each evaluated row. Together with the exact row-count analysis, this supports the interpretation that structural identification errors were a major contributor to low performance in these cases. This pattern is consistent with recent diagnostic work showing that LLM-based evidence extraction degrades when the task requires stable binding between variables, roles, statistical methods, and effect-size information, rather than only recognizing isolated entities [26]. In the present benchmark, the analogous failure point was the construction of the evaluated paper-condition row within a source article. This conclusion should therefore be interpreted as diagnostic rather than population-level, since the rerun subset was specifically selected for persistent structural failures.

For *RQ3*, the models showed descriptive performance differences across input representations and structurally challenging formats. Claude Opus 4.6 was the strongest and most consistent model overall, especially in the Markdown and Multimodal settings, while Gemini 3.1 Pro was also competitive and achieved the strongest figure-category result when paired with Markdown. GPT-5.2, however, warrants a more nuanced interpretation than its overall ranking suggests. Its performance improved strongly once the target condition was provided or once exact row-count agreement had already been achieved. GPT-5.2 therefore illustrates a broader pattern in the benchmark: across models, a substantial share of the remaining error appeared to stem from evaluated-row classification and condition-structure recovery rather than from reading individual numeric values once the relevant row had been identified.

The statistical comparisons further clarify how strongly these findings can be stated. At the paper level, the input representation comparisons did not reach the conventional significance threshold. At the cell level, there was limited evidence for input representation differences, and after Bonferroni correction only the Claude Opus 4.6 PDF versus Multimodal comparison remained statistically significant. The results should therefore be interpreted mainly as descrip-

tive for RQ1 and RQ3. By contrast, RQ2 is supported more directly by the targeted rerun and conditional analyses, although those analyses remain diagnostic rather than population-level. These results also position the present benchmark relative to recent work on automated meta-analysis data extraction. Jansen et al. found that LLMs can approach expert accuracy when extracting effect sizes, numbers of included studies, and numbers of students from educational meta-analyses after expert-adjudicated benchmark construction [25]. The present task was different and more source-level: values had to be recovered from primary articles, linked to the correct condition and assessment interval, and in some cases estimated from figures. The lower and more input representation dependent performance observed here therefore does not contradict those higher-accuracy results; rather, it suggests that extraction accuracy depends strongly on the granularity of the target and on whether the relevant study-condition association must be reconstructed from the source paper. This interpretation also fits with Li et al., who argue that central statistical results for meta-analysis remain high-risk fields that require structured outputs, source linkage, and human verification, even when LLMs can assist with extraction [24].

The specific contribution of the present study relative to this prior work is therefore not a claim that one model or one input representation solves meta-analytic extraction. Rather, the contribution is a more fine-grained diagnostic separation of the extraction problem. The results show that performance depends not only on the model, but also on the reporting source, the input representation, and whether the target row has already been structurally identified. This matters because a system can appear capable at the aggregate level while still failing in specific high-risk situations, such as figure-based SD extraction or condition-row linking. The benchmark therefore extends prior work by showing that the remaining bottleneck is not only numerical extraction, but also the construction and verification of the meta-analytic row itself.

4.2. Implications for automated meta-analysis

A practical implication of this work is that LLM-based extraction may support meta-analysis in two complementary ways: during the construction of a new meta-analysis and during the verification of an existing extraction dataset. In a newly conducted meta-analysis, the GT values are not known in advance. In that setting, an LLM could act as a co-extractor by drafting candidate rows, proposing numerical values, identifying possible source locations, and flagging uncertain cases for human review. In an existing or manually extracted meta-analysis, the same type of system could act as an independent second reader by highlighting discrepancies between the published source papers and the current extraction table.

This distinction is important because the present benchmark used verified GT values for evaluation, but the model was not given those GT values during single-pass extraction. In one response, the model had to identify the relevant study arm, outcome, assessment interval, and the five evaluated numerical values while adhering to a fixed JSON schema. The fact that meaningful extraction was possible at all is therefore notable. It suggests that LLMs may be useful both as accelerators in new extraction workflows and as quality-control tools for checking existing meta-analytic datasets.

During source-paper verification and subsequent targeted error analysis, 14 of the 965 scored cells in the original meta-analytic extraction dataset were found to be incorrect, corresponding to approximately 1.45% of all scored cells. These corrections did not all have the same practical severity: a small correction in n is different from a swapped row or incorrect condition assignment, because the latter can disrupt row alignment and make otherwise correct model outputs appear incorrect across multiple scored cells. In that sense, automated extraction may be valuable not only as an accelerator, but also as a second-reader mech-

anism that helps surface hidden inconsistencies in manually extracted datasets. This also argues for auditable extraction design. EviSearch is a useful point of comparison because it treats clinical evidence extraction from native PDFs as a multi-agent, human-in-the-loop workflow with per-cell provenance, reconciliation, and human verification rather than as an unverified single model response [13]. The same principle applies here: for effect-size-relevant values, the most useful system is likely one that returns the value, the source location, the row interpretation, and a confidence or disagreement signal for human checking. A related lesson comes from agentic scientific reproduction work, where extraction, reimplementation, deterministic evaluation, and error attribution are treated as separable steps, and where discrepancies may originate not only from agents but also from underspecified reporting in the original papers [89]. This parallels the present qualitative analysis, where GT corrections, paper inconsistencies, and task-definition ambiguities were themselves important findings rather than merely noise in the benchmark. More broadly, large-scale AI-assisted peer-review and meta-reviewer-assistance studies show that LLMs can already support scientific evaluation workflows, but those workflows differ from the stricter numerical extraction task studied here [11, 12]. For meta-analysis, the safer implication is therefore not to replace expert extraction outright, but to use LLMs as structured, auditable assistants that can pre-fill candidate values and highlight disagreements for review.

4.3. Limitations and future work

This study has several limitations. First, there was a prompt-related limitation in the main experimental setup. In a subset of papers, the prompt encouraged the models to generate more rows than desired relative to the evaluated row structure. More concretely, the single-pass prompt asked the models to produce a relatively rich JSON output that included not only the five evaluated numerical values, but also additional descriptive fields and study-structure information. In papers containing multiple outcomes, subgroup splits, delayed post-tests, or secondary metrics, this likely encouraged exhaustive extraction of all plausible rows rather than strict adherence to the narrower target used for evaluation. This effect was likely reinforced by the formatted JSON example and the explicit study-structure fields in the prompt, which together encouraged the reconstruction of all plausible conditions and rows. After the main experiments had already been completed, exploratory prompt changes produced somewhat better results on some of these cases. However, the observed improvements were modest, and this study was designed as an extraction benchmark rather than as a prompt-engineering study. For that reason, the full experiment was not rerun with the revised prompt. The measured gains in the target-condition variant may therefore partially reflect that the rerun bypassed a prompt-induced row-selection problem, rather than only an inherent structural limitation of the models.

Paper 74 illustrates a related prompt-sensitivity issue for figure-based SD extraction. In this case, most models left the SD values empty, whereas GPT-5.2 with Multimodal input attempted to extract them and recovered several values correctly, with only a small tolerance error. This suggests that a more aggressive instruction, for example one that encourages the model to always attempt an estimate when a value is visually available, could reduce false negatives and increase true positives. At the same time, it would likely also increase false positives by encouraging unsupported or inaccurate estimates. The effect on F1 would therefore depend on the balance between these changes: F1 would improve only if the gain from additional true positives and fewer false negatives outweighed the loss in precision caused by additional false positives. This means that prompt strictness is not merely a stylistic choice, but directly affects the precision-recall trade-off in this task.

A further evaluation consideration concerns the automated row-alignment procedure. The matching algorithm made the scoring reproducible and avoided manual judgement during evaluation, but it was still an approximation of the underlying semantic alignment problem. In papers with repeated sample sizes, repeated means, or several similar rows, numerical overlap may not uniquely identify the intended target row. The subsequent two-field validity threshold reduced the risk of accepting coincidental one-value matches, but it could not remove all possible ambiguity. Future work could compare this automated alignment rule with independently adjudicated human row alignment, or with an optimization-based alignment procedure that explicitly handles ties and repeated values.

Second, a related evaluation limitation concerns the tolerance rules. For the mean and SD fields, a percentage threshold gives much more absolute room for large values than for small values. For example, under the $\pm 4\%$ figure tolerance used in this study, a value of 100 would allow an absolute deviation of 4, whereas a value of 0.2 would allow only 0.008. This can make small SD values or SD values derived from visually estimated statistics especially difficult to score as correct, even when the model's estimate is close in practical terms. Papers 74 and 122 illustrate related aspects of this issue: Paper 74 shows how visually small error bars can make SD extraction sensitive to small visual differences, while Paper 122 shows how small dispersion values can make percentage-based scoring especially strict.

The sample-size field was treated with a stricter $\pm 0.2\%$ tolerance, which was intended to function as near-exact matching. This was a methodological choice because n is normally reported as an integer count, whereas means and SDs may be rounded, converted from other statistics, or visually estimated. At the same time, scoring n together with the four continuous fields means that all five fields contributed equally to the F1 metric, even though their downstream consequences for effect-size synthesis differ. Errors in means and SDs directly affect the estimated standardized effect, whereas errors in n primarily affect weighting, uncertainty, and, depending on the formula, small-sample correction. This limitation does not invalidate the tolerance-based scoring procedure, but it means that cell-level F1 should be interpreted as an extraction-reliability metric rather than as a direct measure of downstream meta-analytic effect-size error.

Third, all models were run with a provider-specific "medium" reasoning configuration, but these settings are not directly comparable across companies. Although this design reduced arbitrary within-provider variance, it does not guarantee equivalence of actual inference effort across OpenAI, Anthropic, and Google systems. In hindsight, a more controlled comparison might have focused on a resource-based measure such as output token budget or another provider-agnostic constraint. Because proprietary "medium" reasoning settings obscure the actual underlying token budgets and compute cycles, observed performance differences between models may partially reflect unequal compute expenditure rather than purely architectural capabilities.

Fourth, the targeted rerun analysis for RQ2 was diagnostic rather than population-level. The 17-paper subset was specifically chosen because it exhibited persistent structural failures in the single-pass setup (Table 3.2). This makes the target-condition rerun highly informative about a major failure mode, but it should not be interpreted as a direct estimate of expected gains across all papers.

Fifth, the study evaluated realistic single-run performance, with each model run once per paper and per input representation. This design reflects an applied extraction scenario, but it does not measure run-to-run variance. Some errors may therefore reflect stochastic instability rather than stable model behavior. Repeated runs per condition would have allowed a more

complete estimate of uncertainty.

Sixth, the benchmark was derived from a single meta-analysis in the domain of social robotics. Although this provided a thematically coherent and carefully verifiable corpus, reporting conventions, table layouts, figure styles, and outcome structures may differ substantially across other scientific domains. As a result, the relative advantage of Markdown, the frequency of row-construction failures, and the difficulty of figure-based extraction observed here may depend partly on domain-specific reporting practices. The present findings should therefore be interpreted as evidence from one coherent application domain rather than as a fully general estimate of LLM extraction performance across scientific literature.

Seventh, real papers often distribute relevant information across multiple reporting formats at once, such as text, tables, and figures within the same result section. Although this benchmark included all three source types, the category-based analysis still simplifies this mixed-format reality to some extent. Future work could evaluate extraction under explicitly mixed-format conditions in which the relevant benchmark row must be reconstructed from evidence spread across multiple representations within the same paper.

Eighth, the final analyzed source categories were not perfectly balanced. The original selection procedure was designed to balance source categories at 19 papers and 67 statistical rows each, but post-selection verification revealed that one paper originally assigned to the Figure category also reported the target values in a table. Retaining this paper as a Figure paper would have made the figure-category analysis misleading, because the relevant values were not exclusively figure-based. Reassigning it to the Table category or replacing it with another paper after model outputs had already been inspected could also have introduced a different problem, namely an outcome-informed change to the dataset composition. The paper was therefore excluded from the reported analyses. The final dataset consequently contained 18 Figure papers and 59 Figure rows, compared with 19 papers and 67 rows in each of the Text and Table categories. This slightly reduces the symmetry of the category comparison, but it is a more transparent choice than retaining a source-category-ineligible paper. The figure-category estimates should nevertheless be interpreted with this imbalance in mind.

One additional observation may be relevant for future work. In several figure papers, especially bar plots, the first bar closest to the y-axis appeared to be the most accurately extracted, or sometimes the only bar for which the model attempted an SD estimate. This was not formally quantified in the present study, so it should be treated as an exploratory observation rather than as a result. Still, it may point to a positional bias in chart interpretation and would be worth testing explicitly in future work.

5

Conclusion

This study evaluated frontier Large Language Models on meta-analytic data extraction from scientific articles in which relevant results were reported in text, tables, and figures. The results showed that extraction performance depended jointly on input representation, source structure, and model choice. Across the full evaluation, Markdown was often the strongest or near-strongest input representation, but no single representation was universally best across text-, table-, and figure-dominant papers. Claude Opus 4.6 with Markdown achieved the highest overall F1-score (0.8114, 95% CI: [0.7903, 0.8310]), while figure-based papers remained the most difficult category. The Multimodal condition should be interpreted as the native PDF plus an additional cropped visual appendix generated through YOLO-based detection and SigLIP-based filtering, not as an image-only alternative to the PDF.

The clearest finding concerned structural difficulty. On the 17-paper rerun subset, performance improved in all 12 model-input representation combinations when the models were also given the target condition, with gains ranging from $+0.3341$ to $+0.4845$ F1. Likewise, when performance was evaluated only on papers with exact row-count agreement, conditional F1-scores rose to between 0.8660 and 0.9444. These results suggest that much of the remaining error was associated with identifying the correct target row, not only with reading the numbers.

Taken together, the findings suggest that current frontier models can support meta-analytic extraction across text, tables, and figures, but that fully autonomous extraction remains brittle when target rows must first be identified from complex study structures. The main contribution of this study is therefore diagnostic: it separates the effects of input representation, reporting source, target-row construction, and numerical extraction within the same source-paper corpus. Automated extraction is best positioned as a parallel system alongside manual meta-analysis, where it can accelerate extraction and help surface hidden inconsistencies in human-generated datasets. This position is consistent with recent evidence-synthesis work that favors targeted, auditable, and human-in-the-loop automation for high-risk statistical extraction tasks [24, 25, 13]. At the same time, these conclusions should be interpreted in light of the study's prompt sensitivity, provider-specific reasoning settings, diagnostic rerun subset, automated alignment procedure, and single-run design. Future progress will likely depend less on a single universally optimal input representation and more on modular workflows that separate target-row identification from value extraction, especially for figure-heavy and otherwise structurally difficult papers.

Data Availability

The Python script and data used for this analysis are available in a repository linked here: GitHub Repository. Because the paper files were too large for the GitHub repository, they are provided separately as a ZIP archive in Google Drive: Papers.zip. In this zip folder, the following items can be found:

1. All PDFs of the papers
2. All markdown text of the papers
3. All images in the papers (in different directories: keep, delete, tables and figures), where 'keep' is the directory with images used in the evaluation.

Finally, the raw results have been put in an HTML file and can be found here: comparison_results_with_extraction.html

References

[1] Gene V. Glass. “Primary, Secondary, and Meta-Analysis of Research”. In: *Educational Researcher* 5.10 (1976), pp. 3–8. DOI: 10.3102/0013189X005010003.

[2] Mark Starr et al. “The origins, evolution, and future of The Cochrane Database of Systematic Reviews”. In: *International Journal of Technology Assessment in Health Care* 25.S1 (2009), pp. 182–195. DOI: 10.1017/S026646230909062X.

[3] John P. A. Ioannidis. “The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses”. In: *The Milbank Quarterly* 94.3 (2016), pp. 485–514. DOI: 10.1111/1468-0009.12210.

[4] Annette M. O’Connor et al. “A question of trust: can we build an evidence base to gain trust in systematic review automation technologies?” In: *Systematic Reviews* 8.1 (2019), p. 143. DOI: 10.1186/s13643-019-1062-0.

[5] Rohit Borah et al. “Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry”. In: *BMJ Open* 7.2 (2017), e012545. DOI: 10.1136/bmjopen-2016-012545.

[6] Per Olav Vandvik, Romina Brignardello-Petersen, and Gordon H Guyatt. “Living cumulative network meta-analysis to reduce waste in research: a paradigmatic shift for systematic reviews?” In: *BMC Medicine* 14.1 (2016), p. 59. DOI: 10.1186/s12916-016-0596-4.

[7] Guy Tsafnat et al. “Systematic review automation technologies”. In: *Systematic Reviews* 3.1 (2014), p. 74. DOI: 10.1186/2046-4053-3-74.

[8] Iain J. Marshall and Byron C. Wallace. “Toward systematic review automation: a practical guide to using machine learning tools in research synthesis”. In: *Systematic Reviews* 8.1 (2019), p. 163. DOI: 10.1186/s13643-019-1074-9.

[9] Takehiko Oami, Yohei Okada, and Taka-aki Nakada. “Optimal large language models to screen citations for systematic reviews”. In: *Research Synthesis Methods* 16.6 (2025), pp. 859–875. DOI: 10.1017/rsm.2025.10014.

[10] Christian Cao et al. “Automation of Systematic Reviews with Large Language Models”. In: *medRxiv* (2025). DOI: 10.1101/2025.06.13.25329541. eprint: https://www.medrxiv.org/content/early/2025/06/13/2025.06.13.25329541.full.pdf. URL: https://www.medrxiv.org/content/early/2025/06/13/2025.06.13.25329541.

[11] Joydeep Biswas et al. *AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot*. 2026. arXiv: 2604.13940 [cs.AI]. URL: https://arxiv.org/abs/2604.13940.

[12] Eftekhar Hossain et al. “LLMs as Meta-Reviewers’ Assistants: A Case Study”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 7763–7803. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.395. URL: https://aclanthology.org/2025.naacl-long.395/.

[13] Naman Ahuja et al. *EviSearch: A Human in the Loop System for Extracting and Auditing Clinical Evidence for Systematic Reviews*. 2026. arXiv: 2604.14165 [cs.CL]. URL: https://arxiv.org/abs/2604.14165.

[14] OpenAI. *Introducing GPT-5.2*. https://openai.com/index/introducing-gpt-5-2/. Released December 11, 2025. Dec. 2025.

[15] Anthropic. *Introducing Claude Opus 4.6*. https://www.anthropic.com/news/claude-opus-4-6. Released February 5, 2026. Feb. 2026.

[16] Google DeepMind. *Gemini 3.1 Pro: The Next Generation of Native Multimodality*. https://deepmind.google/technologies/gemini/. Released February 19, 2026. Feb. 2026.

[17] Artificial Analysis. *AA-Omniscience: Knowledge and Hallucination Benchmark*. https://artificialanalysis.ai/evaluations/omniscience. Dec. 2025.

[18] Ka Yin Chin et al. "Assessing the Performance of LLMs in Multimodal Information Extraction for Biological Research". In: *bioRxiv* (2025). DOI: 10.1101/2025.10.29.685285.

[19] Yuan Sui et al. "Table meets LLM: Can Large Language Models Understand Structured Table Data?" In: *arXiv preprint arXiv:2405.14764* (2024).

[20] Zhehao Zhang, Yan Gao, and Jian-Guang Lou. "E5: Zero-shot Hierarchical Table Analysis using Augmented LLMs via Explain, Extract, Execute, Exhibit and Extrapolate". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*. Association for Computational Linguistics, 2024, pp. 1244–1258. URL: https://aclanthology.org/2024.naacl-long.68.

[21] Lei Huang et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 42 (2024).

[22] Ivan Turan and Taylor Sparks. "Revolutionizing Scientific Figure Decoding: Benchmarking LLM Data Extraction Performance". In: *ChemRxiv* (2025). Preprint. DOI: 10.26434/chemrxiv-2025-pcq8d.

[23] Yucheng Han et al. "ChartLlama: A Multimodal LLM for Chart Understanding and Generation". In: *arXiv preprint arXiv:2311.16483* (2024).

[24] Lingbo Li, Anuradha Mathrani, and Teo Susnjak. "What level of automation is "good enough"? A benchmark of large language models for meta-analysis data extraction". In: *Research Synthesis Methods* (2026), pp. 1–22. DOI: 10.1017/rsm.2025.10066.

[25] Thorben Jansen et al. "Automated Data Extraction by Large Language Models: Assessing Accuracy in Comparison to Human Experts Using the Example of Visible Learning". In: *Educational Psychology Review* 38.38 (2026). DOI: 10.1007/s10648-026-10136-5. URL: https://doi.org/10.1007/s10648-026-10136-5.

[26] Zhiyin Tan and Jennifer D'Souza. *Diagnosing Structural Failures in LLM-Based Evidence Extraction for Meta-Analysis*. 2026. arXiv: 2602.10881 [cs.CL]. URL: https://arxiv.org/abs/2602.10881.

[27] Joost de Winter et al. "Social Robots: A Meta-Analysis of Learning Outcomes". In: *Frontiers in Robotics and AI* 12 (2025), p. 1735198. DOI: 10.3389/frobt.2025.1735198. URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1735198/full.

[28] Glenn Jocher and Jing Qiu. *Ultralytics YOLO11*. Version 11.0.0. 2024. URL: https://github.com/ultralytics/ultralytics.

[29] Emmy Rintjema et al. “A Robot Teaching Young Children a Second Language”. en. In: (2018), pp. 219–220. DOI: 10.1145/3173386.3177059. URL: https://doi.org/10.1145/3173386.3177059.

[30] Jacqueline M. Kory Westlund et al. “Children use non-verbal cues to learn new words from robots as well as people”. en. In: *International Journal of Child-Computer Interaction* 13 (2017), pp. 1–9. DOI: 10.1016/j.ijcci.2017.04.001. URL: https://doi.org/10.1016/j.ijcci.2017.04.001.

[31] Minoo Alemi and Nafiseh Sadat Haeri. “How to Develop Learners’ Politeness: A Study of RALL’s Impact on Learning Greeting by Young Iranian EFL Learners”. en. In: (2017), pp. 88–94. DOI: 10.1109/icrom.2017.8466206. URL: https://doi.org/10.1109/icrom.2017.8466206.

[32] Takamasa Iio et al. “Improvement of Japanese adults’ English speaking skills via experiences speaking to a robot”. en. In: *Journal of Computer Assisted Learning* 35.2 (2018), pp. 228–245. DOI: 10.1111/jcal.12325. URL: https://doi.org/10.1111/jcal.12325.

[33] Karina R. Liles. *Ms. An (Meeting Students’ Academic Needs): Engaging Students in Math Education*. en. 2019. DOI: 10.1007/978-3-030-22341-0_50. URL: https://doi.org/10.1007/978-3-030-22341-0_50.

[34] Daniel Leyzberg, Aditi Ramachandran, and Brian Scassellati. “The Effect of Personalization in Longer-Term Robot Tutoring”. en. In: *ACM Transactions on Human-Robot Interaction* 7.3 (2018), pp. 1–19. DOI: 10.1145/3283453. URL: https://doi.org/10.1145/3283453.

[35] Jan de Wit et al. “The Effect of a Robot’s Gestures and Adaptive Tutoring on Children’s Acquisition of Second Language Vocabularies”. en. In: (2018), pp. 50–58. DOI: 10.1145/3171221.3171277. URL: https://doi.org/10.1145/3171221.3171277.

[36] Leigh Levinson et al. “Learning in Summer Camp with Social Robots: A Morphological Study”. en. In: *International Journal of Social Robotics* 13.5 (2020), pp. 999–1012. DOI: 10.1007/s12369-020-00689-y. URL: https://doi.org/10.1007/s12369-020-00689-y.

[37] Thomas W. Draper and Wanda W. Clayton. “Using a Personal Robot to Teach Young Children”. en. In: *The Journal of Genetic Psychology* 153.3 (1992), pp. 269–273. DOI: 10.1080/00221325.1992.10753723. URL: https://doi.org/10.1080/00221325.1992.10753723.

[38] Anna-Maria Velentza, Nikolaos Fachantidis, and Ioannis Lefkos. “Learn with surprize from a robot professor”. en. In: *Computers & Education* 173 (2021), p. 104272. DOI: 10.1016/j.compedu.2021.104272. URL: https://doi.org/10.1016/j.compedu.2021.104272.

[39] Maria Blancas-Muñoz et al. *Hints vs Distractions in Intelligent Tutoring Systems: Looking for the proper type of help*. en. 2018. DOI: 10.48550/arxiv.1806.07806. URL: http://arxiv.org/abs/1806.07806.

[40] Nicole Salomons et al. ““We Make a Great Team!”: Adults with Low Prior Domain Knowledge Learn more from a Peer Robot than a Tutor Robot”. en. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2022), pp. 176–184. DOI: 10.1109/hri53351.2022.9889441. URL: https://doi.org/10.1109/hri53351.2022.9889441.

[41] Chih-Chien Hu, Yu-Fen Yang, and Nian-Shing Chen. “Human–robot interface design – the ‘Robot with a Tablet’ or ‘Robot only’, which one is better?” In: *Behaviour & Information Technology* 42.10 (2023), pp. 1590–1603. DOI: 10.1080/0144929X.2022.2093271. eprint: https://doi.org/10.1080/0144929X.2022.2093271. URL: https://doi.org/10.1080/0144929X.2022.2093271.

[42] Anna-Maria Velentza, Efthymia Kefalouka, and Nikolaos Fachantidis. *Socially Assistive Robot in Sexual Health: Group and Individual Student-Robot Interaction Activities Promoting Disclosure, Learning and Positive Attitudes*. en. 2024. DOI: 10.48550/arxiv.2407.13030. URL: http://arxiv.org/abs/2407.13030.

[43] James Kennedy et al. “Heart vs hard drive: Children learn more from a human tutor than a social robot”. en. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016), pp. 451–452. DOI: 10.1109/hri.2016.7451801. URL: https://doi.org/10.1109/hri.2016.7451801.

[44] James Kennedy, Paul Baxter, and Tony Belpaeme. “The Robot Who Tried Too Hard”. en. In: (2015), pp. 67–74. DOI: 10.1145/2696454.2696457. URL: https://doi.org/10.1145/2696454.2696457.

[45] James Kennedy et al. *Higher Nonverbal Immediacy Leads to Greater Learning Gains in Child-Robot Tutoring Interactions*. en. 2015. DOI: 10.1007/978-3-319-25554-5_33. URL: https://doi.org/10.1007/978-3-319-25554-5_33.

[46] Anara Sandygulova et al. “CoWriting Kazakh”. en. In: (2020), pp. 113–120. DOI: 10.1145/3319502.3374813. URL: https://doi.org/10.1145/3319502.3374813.

[47] Patrícia Alves-Oliveira et al. “Boosting children’s creativity through creative interactions with social robots”. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2016, pp. 591–592. DOI: 10.1109/HRI.2016.7451871.

[48] Kai-Yi Chin, Zeng□Wei Hong, and Yen□Lin Chen. “Impact of Using an Educational Robot-Based Learning System on Students’ Motivation in Elementary Education”. en. In: *IEEE Transactions on Learning Technologies* 7.4 (2014), pp. 333–345. DOI: 10.1109/tlt.2014.2346756. URL: https://doi.org/10.1109/tlt.2014.2346756.

[49] Alice Rosi et al. “The use of new technologies for nutritional education in primary schools: a pilot study”. en. In: *Public Health* 140 (2016), pp. 50–55. DOI: 10.1016/j.puhe.2016.08.021. URL: https://doi.org/10.1016/j.puhe.2016.08.021.

[50] Huili Chen, Hae Won Park, and Cynthia Breazeal. “Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children’s learning and emotive engagement”. en. In: *Computers & Education* 150 (2020), p. 103836. DOI: 10.1016/j.compedu.2020.103836. URL: https://doi.org/10.1016/j.compedu.2020.103836.

[51] Wilma C. M. Resing et al. “Dynamic testing: Can a robot as tutor be of help in assessing children’s potential for learning?” en. In: *Journal of Computer Assisted Learning* 35.4 (2019), pp. 540–554. DOI: 10.1111/jcal.12358. URL: https://doi.org/10.1111/jcal.12358.

[52] Joseph E. Michaelis and Bilge Mutlu. “Supporting Interest in Science Learning with a Social Robot”. en. In: (2019), pp. 71–82. DOI: 10.1145/3311927.3323154. URL: https://doi.org/10.1145/3311927.3323154.

[53] Nichola Lubold et al. *Comfort with Robots Influences Rapport with a Social, Entraining Teachable Robot*. en. 2019. DOI: 10.1007/978-3-030-23204-7_20. URL: https://doi.org/10.1007/978-3-030-23204-7_20.

[54] Eun-Ja Hyun et al. “Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children”. en. In: (2008), pp. 187–192. DOI: 10.1109/roman.2008.4600664. URL: https://doi.org/10.1109/roman.2008.4600664.

[55] James Kennedy, Paul Baxter, and Tony Belpaeme. “Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children”. en. In: *International Journal of Social Robotics* 7.2 (2014), pp. 293–308. DOI: 10.1007/s12369-014-0277-4. URL: https://doi.org/10.1007/s12369-014-0277-4.

[56] Takayuki Kanda et al. “Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial”. en. In: *Human-Computer Interaction* 19.1 (2004), pp. 61–84. DOI: 10.1207/s15327051hci1901&2_4. URL: https://doi.org/10.1207/s15327051hci1901&2_4.

[57] H. S. Hsiao et al. ““iRobiQ”: the influence of bidirectional interaction on kindergarteners’ reading motivation, literacy, and behavior”. en. In: *Interactive Learning Environments* 23.3 (2012), pp. 269–292. DOI: 10.1080/10494820.2012.745435. URL: https://doi.org/10.1080/10494820.2012.745435.

[58] Anna Riedmann, Philipp Schaper, and Birgit Lugrin. “Integration of a social robot and gamification in adult learning and effects on motivation, engagement and performance”. In: *AI & SOCIETY* 39.1 (Feb. 2024), pp. 369–388. ISSN: 1435-5655. DOI: 10.1007/s00146-022-01514-y. URL: https://doi.org/10.1007/s00146-022-01514-y.

[59] Gwo-Dong Chen et al. “Digital Learning Playground: supporting authentic learning experiences in the classroom”. en. In: *Interactive Learning Environments* 21.2 (2012), pp. 172–183. DOI: 10.1080/10494820.2012.705856. URL: https://doi.org/10.1080/10494820.2012.705856.

[60] Alireza Esfandbod et al. “Fast mapping in word-learning: A case study on the humanoid social robots’ impacts on Children’s performance”. en. In: *International Journal of Child-Computer Interaction* 38 (2023), p. 100614. DOI: 10.1016/j.ijcci.2023.100614. URL: https://doi.org/10.1016/j.ijcci.2023.100614.

[61] Jamy Li et al. “Social robots and virtual agents as lecturers for video instruction”. en. In: *Computers in Human Behavior* 55 (2015), pp. 1222–1230. DOI: 10.1016/j.chb.2015.04.005. URL: https://doi.org/10.1016/j.chb.2015.04.005.

[62] Zeng-Wei Hong et al. “Utilizing robot-tutoring approach in oral reading to improve Taiwanese EFL students’ English pronunciation”. en. In: *Cogent Education* 11.1 (2024). DOI: 10.1080/2331186x.2024.2342660. URL: https://doi.org/10.1080/2331186x.2024.2342660.

[63] Erin Walker et al. “The Effects of Physical form and Embodied Action in a Teachable Robot for Geometry Learning”. en. In: (2016), pp. 381–385. DOI: 10.1109/icalt.2016.129. URL: https://doi.org/10.1109/icalt.2016.129.

[64] James Kennedy et al. “Social robot tutoring for child second language learning”. en. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016), pp. 231–238. DOI: 10.1109/hri.2016.7451757. URL: https://doi.org/10.1109/hri.2016.7451757.

[65] Zhanel Zhexenova et al. “A Comparison of Social Robot to Tablet and Teacher in a New Script Learning Context”. en. In: *Frontiers in Robotics and AI* 7 (2020), p. 99. DOI: 10.3389/frobt.2020.00099. URL: https://doi.org/10.3389/frobt.2020.00099.

[66] Jun Yin et al. “The Influence of Robot Social Behaviors on Second Language Learning in Preschoolers”. en. In: *International Journal of Human-Computer Interaction* 40.7 (2022), pp. 1600–1608. DOI: 10.1080/10447318.2022.2144828. URL: https://doi.org/10.1080/10447318.2022.2144828.

[67] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. “Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools”. en. In: *International Journal of Humanoid Robotics* 11.03 (2014), p. 1450022. DOI: 10.1142/s0219843614500224. URL: https://doi.org/10.1142/s0219843614500224.

[68] Shruti Chandra et al. “Children’s peer assessment and self-disclosure in the presence of an educational robot”. en. In: (2016), pp. 539–544. DOI: 10.1109/roman.2016.7745170. URL: https://doi.org/10.1109/roman.2016.7745170.

[69] Melissa Donnermann, Philipp Schäper, and Birgit Lugrin. “Integrating a Social Robot in Higher Education – A Field Study”. en. In: (2020), pp. 573–579. DOI: 10.1109/ro-man47096.2020.9223602. URL: https://doi.org/10.1109/ro-man47096.2020.9223602.

[70] Thorsten Schodde et al. “Adapt, Explain, Engage—A Study on How Social Robots Can Scaffold Second-language Learning of Children”. en. In: *ACM Transactions on Human-Robot Interaction* 9.1 (2019), pp. 1–27. DOI: 10.1145/3366422. URL: https://doi.org/10.1145/3366422.

[71] Rosa Lanzilotti et al. “Social Robot to teach coding in primary school”. en. In: 2017 (2021), pp. 102–104. DOI: 10.1109/icalt52272.2021.00038. URL: https://doi.org/10.1109/icalt52272.2021.00038.

[72] Koki Suzuki and Masayoshi Kanoh. “Effectiveness of a robot for supporting expression education”. en. In: (2015), pp. 498–501. DOI: 10.1109/taai.2015.7407119. URL: https://doi.org/10.1109/taai.2015.7407119.

[73] Koki Suzuki and Masayoshi Kanoh. “Investigating Effectiveness of an Expression Education Support Robot That Nods and Gives Hints”. en. In: *Journal of Advanced Computational Intelligence and Intelligent Informatics* 21.3 (2017), pp. 483–495. DOI: 10.20965/jaciii.2017.p0483. URL: https://doi.org/10.20965/jaciii.2017.p0483.

[74] Felix Jimenez et al. “Effect of collaborative learning with robot that prompts constructive interaction”. In: *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* 2013 (Oct. 2013), pp. 2983–2988. DOI: 10.1109/smc.2014.6974384.

[75] Ryo Yoshizawa, Felix Jimenez, and Kazuhito Murakami. “Proposal of a Behavioral Model for Robots Supporting Learning According to Learners’ Learning Performance”. en. In: *Journal of Robotics and Mechatronics* 32.4 (2020), pp. 769–779. DOI: 10.20965/jrm.2020.p0769. URL: https://doi.org/10.20965/jrm.2020.p0769.

[76] Kohei Okawa et al. “Proposal of Learning Support Model for Teacher-Type Robot Supporting Learning According to Learner’s Perplexed Facial Expressions”. en. In: *Journal of Robotics and Mechatronics* 36.1 (2024), pp. 168–180. DOI: 10.20965/jrm.2024.p0168. URL: https://doi.org/10.20965/jrm.2024.p0168.

[77] Elly A. Konijn et al. “Social Robots for (Second) Language Learning in (Migrant) Primary School Children”. en. In: *International Journal of Social Robotics* 14.3 (2021), pp. 827–843. DOI: 10.1007/s12369-021-00824-3. URL: https://doi.org/10.1007/s12369-021-00824-3.

[78] Takayuki Kanda, Michihiro Shimada, and Satoshi Koizumi. “Children learning with a social robot”. en. In: (2012), pp. 351–358. DOI: 10.1145/2157689.2157809. URL: https://doi.org/10.1145/2157689.2157809.

[79] Fumihide Tanaka and Shizuko Matsuzoe. “Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning”. en. In: *Journal of Human-Robot Interaction* (2012), pp. 78–95. DOI: 10.5898/jhri.1.1.tanaka. URL: https://doi.org/10.5898/jhri.1.1.tanaka.

[80] Jan de Wit et al. “Varied Human-Like Gestures for Social Robots”. en. In: (2020), pp. 359–367. DOI: 10.1145/3319502.3374815. URL: https://doi.org/10.1145/3319502.3374815.

[81] Iris Howley et al. “Effects of social presence and social role on help-seeking and learning”. en. In: (2014), pp. 415–422. DOI: 10.1145/2559636.2559667. URL: https://doi.org/10.1145/2559636.2559667.

[82] Vivien Lin et al. “Enhancing EFL vocabulary learning with multimodal cues supported by an educational robot and an IoT-Based 3D book”. In: *System* 104 (2022), p. 102691. ISSN: 0346-251X. DOI: https://doi.org/10.1016/j.system.2021.102691. URL: https://www.sciencedirect.com/science/article/pii/S0346251X21002451.

[83] Maryam Alimardani et al. *Assessment of Engagement and Learning During Child-Robot Interaction Using EEG Signals*. en. 2021. DOI: 10.1007/978-3-030-90525-5_59. URL: https://doi.org/10.1007/978-3-030-90525-5_59.

[84] Junko Kanero et al. “When Even a Robot Tutor Zooms: A Study of Embodiment, Attitudes, and Impressions”. en. In: *Frontiers in Robotics and AI* 8 (2021), p. 679893. DOI: 10.3389/frobt.2021.679893. URL: https://doi.org/10.3389/frobt.2021.679893.

[85] LlamaIndex. *LlamaParse*. Version 2.0. AI-powered document parsing engine. 2025. URL: https://cloud.llamaindex.ai.

[86] Armaggheddon. *YOLO11 Document Layout Analysis*. Fine-tuned on DocLayNet dataset. Accessed: 30-01-2026. 2025. URL: https://github.com/Armaggheddon/yolo_doc_layout/tree/yolo11.

[87] Xiaohua Zhai et al. “Sigmoid Loss for Language Image Pre-Training”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 11975–11986.

[88] Nancy A. Obuchowski. “On the comparison of correlated proportions for clustered data”. In: *Statistics in Medicine* 17.13 (July 1998), pp. 1495–1507. DOI: 10.1002/(SICI)1097-0258(19980715)17:13<1495::AID-SIM863>3.0.CO;2-I.

[89] Benjamin Kohler et al. *Read the Paper, Write the Code: Agentic Reproduction of Social-Science Results*. 2026. arXiv: 2604.21965 [cs.AI]. URL: https://arxiv.org/abs/2604.21965.

A

Include and Exclude

Note: The bracketed numbers in Table 2.1 are bibliography reference numbers. Appendix A uses the original study numbering from the source screening file; these numbering systems are therefore not identical.

Table A.1: Overview of study inclusion and exclusion decisions

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
1	Text	Rintjema (2018)	Include		Explicit
2	Table	Mazzoni (2015)	Exclude	SDs calculated using raw data	
3	Text	Chandra (2015)	Exclude	Figure/text only; gains reported	
4	Text	Kory Westlund (2017)	Include		Explicit
5	Text	Kennedy (2016)	Include		Implicit
6	Figure	Alemi (2014)	Include		Explicit
7	Text	Alemi (2017)	Include		Explicit
8	Text	Mubin (2019)	Exclude	Improvement percentages	
9	Table	Chin (2014)	Include		Explicit
10	Text	Iio (2019)	Include		Explicit
11	Table	Rosi (2016)	Include		Explicit
12	Text	Park (2019)	Could include		Explicit
13	Table	Conti (2020)	Could include		Explicit
14	Table	Vogt (2019)	Could include		Explicit
15	Table	Kennedy (2016)	Include		Implicit
16	Text	Kennedy (2015)	Include		Implicit
17	Figure	Ramachandran (2018)	Exclude	Raw-data calculation	

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
18	Table	Chen (2020)	Include		Explicit
19	Figure	Chandra (2020)	Exclude	Means in text; SE in figure	
20	Text	Chandra (2020)	Exclude	Simulation-based	
21	Figure	Chandra (2016)	Include		Explicit
22	Table	Liebens (2019)	Could include		Explicit
23	Text	Liles (2019)	Include		Explicit
24	Figure	Ramachandran (2019)	Exclude	Post from figure raw data	
25	Text	Leyzberg (2018)	Include		Explicit
26	Table	Resing (2019)	Include		Explicit
27	Table	Michaelis (2019)	Include		Explicit
28	Text	De Wit (2018)	Include		Explicit
29	Figure	Ramachandran (2017)	Exclude	Pre in text; post from figure raw data	
30		Molenaar (2021)	Exclude	Supplementary information	
31	Table	Lubold (2018)	Include		Explicit
32	Text	Kennedy (2015)	Include		Implicit
33	Table	Hyun (2008)	Include		Explicit
34	Table	Kennedy (2015)	Include		Explicit
35	Table	Lee (2010)	Could include		Explicit
36	Text	Liles (2015)	Exclude	Simulation-based	
37	Table	Park (2011)	Exclude	Simulation-based	
38	Table	Kennedy (2017)	Exclude	Implicit; 95% CI in text/table	Implicit
39	Table	Kennedy (2017)	Exclude	Implicit; 95% CI in text/table	
40	Text	Hong (2016)	Could include		Explicit
41	Figure	Donnermann (2020)	Include		Explicit
42	Figure	Konijn (2022)	Include		Implicit
43	Table	Zhexenova (2020)	Include		Implicit
44	Figure	Kanda (2012)	Include		Implicit
45	Text	Konijn (2020)	Could include		Explicit
46	Text	Levinson (2021)	Include		Explicit
47		Baxter (2017)	Exclude	Supplementary information	
48	Figure	Schodde (2019)	Include		Explicit
49	Table	Hoorn (2021)	Exclude	Supplementary information	

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
50	Text	Draper (1992)	Include		Explicit
51	Table	Han (2008)	Could include		Explicit
52	Figure	Tanaka (2012)	Include		Implicit
53	Table	Moorlag (2021)	Could include		Explicit
54	Text	Yueh (2020)	Exclude	Simulation-based	
55	Text	Komatsubara (2014)	Could include		Explicit
56	Text	Sandygulova (2020)	Include		Implicit
57	Figure	Lanzilotti (2021)	Include		Explicit
58	Text	Park (2017)	Exclude	Simulation-based	
59	Text	Banaeian (2021)	Could include		Explicit
60	Table	Van den Berghe (2019)	Could include		Explicit
61	Text	Kory-Westlund (2019)	Could include		Explicit
62	Figure	De Wit (2020)	Include		Implicit
63	Text	Alves-Oliveira (2019)	Could include		Explicit
64	Text	Alves-Oliveira (2019)	Could include		Explicit
65	Table	Wang (2013)	Could include		Explicit
66	Figure	Movellan (2009)	Exclude	SD from figure interpolation	
67	Text	Eimler (2010)	Could include		Explicit
68	Table	Kanda (2004)	Include		Explicit
69	Table	Wu (2015)	Could include		Explicit
70	Text	Velentza (2021)	Include		Explicit
71		Alves-Oliveira (2020)	Exclude	Supplementary info; raw data elsewhere	
72	Figure	Howley (2014)	Include		Implicit
73	Table	Kennedy (2017)	Exclude	Simulation-based	
74	Figure	Suzuki (2015)	Include		Explicit
75	Figure	Suzuki (2017)	Include		Explicit
76	Figure	Bhat (2016)	Exclude	Pre in text; raw data in figures	
77	Table	De Haas (2021)	Could include		Explicit

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
78	Table	De Haas (2020)	Could include		Explicit
79	Table	Hsiao (2015)	Include		Explicit
80	Table	Lubold (2018)	Exclude	Paper unavailable; SD/SE unclear	
81	Text	Blancas-Muñoz (2018)	Include		Explicit
82	Text	Ramachandran (2019)	Could include		Explicit
83	Table	Coninx (2016)	Exclude	Raw data in Table 2	
84	Table	Resing (2020)	Could include		Explicit
85	Table	Chen (2020)	Could include		Explicit
86	Table	Al Hakim (2022)	Could include		Explicit
87		Elgarf (2022)	Exclude	Supplementary info; raw data elsewhere	
88	Text	Salomons (2022)	Include		Explicit
89	Figure	Lin 2022	Include		Implicit
90	Table	McIntosh (2022)	Exclude	SD formula mismatch	
91	Text	Alves-Oliveira (2020)	Include		Implicit
92	Table	Riedmann (2024)	Include		Explicit
93	Figure	Norman (2022)	Exclude	Figure unreadable	
94	Text	Hu (2023)	Include		Explicit
95	Table	Chen (2013)	Include		Explicit
96	Table	Hsieh (2022)	Could include		Explicit
97	Text	Zhanatkyzy (2022)	Could include		Explicit
98	Table	Yin (2024)	Include		Implicit
99	Table	Steinhaeusser (2022)	Could include		Explicit
100	Text	Donnermann (2022)	Could include		Explicit
101	Text	Lubbers (2020)	Could include		Explicit
102	Table	Zachrisson (2021)	Could include		Explicit
103	Table	Zhang (2023)	Could include		Explicit
104	Text	Tang (2023)	Exclude	Pre in text; post in figure	

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
105	Text	Peura (2023)	Could include		Explicit
106	Table	Pasalidou (2023)	Exclude	Personal communication	
107	Table	Lu (2023)	Could include		Explicit
108	Table	Esfandbod (2023)	Include		Explicit
109	Table	Alemi (2023)	Could include		Explicit
110	Table	Bravo Perucho (2023)	Exclude	Raw data in table	
111	Table	Hsieh (2024)	Could include		Explicit
112	Figure	Miyauchi (2020)	Exclude	Mixed sources: table/figure	
113	Figure	Miyauchi (2020)	Exclude	Mixed sources: table/figure	
114	Table	Rosenthal-von der Pütten (2016)	Could include		Explicit
115	Table	Donnermann (2024)	Could include		Explicit
116	Figure	Verhelst (2023)	Exclude	Raw data via personal communication	
117	Table	Iio (2024)	Could include		Explicit
118	Table	Yuliani (2024)	Could include		Explicit
119	Table	Wei (2021)	Could include		Explicit
120	Figure	Kanero (2022)	Exclude	Mixed sources: target values also reported in Table 8	
121	Table	Tsai (2019)	Could include		Explicit
122	Figure	Alimardani (2021)	Include		Implicit
123	Table	Chang (2024)	Could include		Explicit
124	Table	Ponce (2019)	Exclude	Raw-data calculation	
125	Table	Jao (2024)	Could include		Explicit
126	Figure	Kanero (2021)	Include		Implicit
127	Text	Velentza (2024)	Include		Explicit

Study no.	Category	Citation	Include / Exclude	Exclude reason	Implicit / Explicit
128	Text	Manikutty (2024)	Exclude	Personal communication	
129	Text	Velentza (2024)	Could include		Explicit
130	Text	Velentza (2024)	Could include		Explicit
131	Text	Kagkelidou (2024)	Could include		Explicit
132	Table	Hsiao (2024)	Could include		Explicit
133	Table	Li (2016)	Include		Explicit
134		Liu (2024)	Exclude	Personal communication	
135		Van Minkelen (2020)	Exclude	Personal communication	
136	Table	Hong (2024)	Include		Explicit
137		Oralbayeva (2023)	Exclude	Personal communication	
138		Shakerimov (2023)	Exclude	Personal communication	
139		Sarmonov (2023)	Exclude	Personal communication	
140	Text	Gruson (2019)	Could include		Explicit
141	Figure	Jimenez (2013)	Include		Explicit
142	Table	Jimenez (2020)	Exclude	Raw data in table	
143	Text	Donnermann (2021)	Could include		Explicit
144	Table	Darmawansah (2024)	Exclude	Simulation-based	
145	Table	Walker (2016)	Include		Explicit
146	Figure	Yoshizawa (2020)	Include		Explicit
147	Figure	Okawa (2024)	Include		Explicit

B

Prompt Specifications

This appendix details the exact instructions used for the Large Language Models across the three input representations used in the study: Native PDF, Markdown, and Multimodal.

The **Prompt Templates** (Sections B.2 and B.3) remain static, except for the variables `{VISUAL_RULES}` and `{INPUT_CONTEXT}` in both prompts, and `{formatted_targets}` and `{formatted_json_example}` in the extraction prompt, which change depending on the input representation used.

Here, `{formatted_targets}` contains a bullet-point list of all target conditions supplied to the model, while `{formatted_json_example}` contains the full target JSON structure in which these conditions are already instantiated as keys.

B.1. Input Representation Variable Definitions

The following content is injected into the placeholders found in the prompt templates below.

input representation 1: Native PDF

`{INPUT_CONTEXT}:`

The input consists of the PDF file of the paper.

`{VISUAL_RULES}:`

Use only the content found within the provided PDF file of the paper.

input representation 2: Markdown / LlamaParse

`{INPUT_CONTEXT}:`

The input consists of the Markdown text of the paper.
- Tables: Formatted as Markdown grids.
- Figures: Descriptive text summaries of all figures from the paper.

`{VISUAL_RULES}:`

Use only the content found within the provided Markdown text of the paper.

input representation 3: Multimodal / YOLO

`{INPUT_CONTEXT}:`

The input consists of the PDF file of the paper and PNG files of the tables and figures of the paper.
- Tables and figures: Pre-processed PNG files containing only the data-rich tables and figures (e.g., results tables, bar charts) extracted from the PDF file of the paper.

`{VISUAL_RULES}:`

Use only the content found within the provided PDF file of the paper and the provided PNG files of the tables and figures of the paper.

B.2. Single-Pass Prompt Template

Used to identify all eligible rows and extract their associated values in a single-pass.

System Prompt: You are an expert Methodologist and precise Data Extraction Specialist.

Prompt Content

```
TASK
From INPUT_CONTEXT, extract all eligible entries and output ONE valid JSON object (JSON only).
The data might be available in the text, tables, or figures of INPUT_CONTEXT.

Eligible entry = 1 Condition × 1 Outcome Measure × 1 post-test timepoint (i.e., one row in the output JSON per such entry).
Never create an entry where the Condition Name is a timepoint label (e.g., "pre-test", "post-test", "T0", "T1").
Pre-test values are attributes (M_pre, SD_pre) of the same entry and must NOT appear as separate keys.

OUTCOME MEASURES
Include: objective cognitive/learning performance outcomes (e.g., recall, comprehension, accuracy, test scores, etc.).
Exclude: perceived learning/attitudes/affect (motivation, anxiety, etc.).
If both subscales and total are reported for the same construct, extract subscales only.
If multiple post-tests: keep separate; append timepoint label to Outcome Measure so keys are unique.

KEYS
Key = "[Condition Name] - [Outcome Measure]".
Condition Name:
- Between-subjects: reported group name.
- Within-subjects/crossover: reported condition/phase name.
Gender-split reporting: make ONE key (no gender in key) and aggregate values (below).

VALUES (per key)
Stats:
- M_pre, SD_pre, M_post, SD_post, n.
- Post-test only: M_pre/SD_pre = null.
- Within/crossover: if one global pretest at study start, reuse it for all conditions; if condition-specific pretests exist, use
      matching one.
Missing/uncalculable => null.

SD derivation (only for a mean of same condition+timepoint, not differences/models):
If SD is not readily available, but other summary measures are (e.g., SE, 95% CI, or IQR), derive SD from these. If derived,
      append "; derived" to source_mean_sd.

Gender aggregation (same condition/outcome/timepoint):
If the same Outcome Measure is reported for males and females separately, merge these data.

Sample size n (use first available; if n_pre != n_post use n_post):
1) outcome+timepoint+condition n
2) timepoint+condition analyzed n
3) condition assigned/randomized n
4) else null

DESCRIPTORS (null if missing/unclear; use condition-level else study-level if clearly applies)
- country: ISO 3166-1 alpha-3
- robot_model: robot model/platform; null if non-robot condition
- learning_subject_posttest: subject/domain for this outcome/timepoint
- participants_trained_at_a_time: 1, 2, or "Group"(>2)
- presence_human_tutor: 1 if teaching/tutoring during training; 0 if not; null if unclear
- total_duration_training_min: total minutes across sessions (per-session*#sessions); range only => null
- number_training_sessions: integer
- posttest_delay_days: days from end of training to this post-test (weeks*7); immediate => 0; qualitative only => null
- similarity_pre_vs_post: 1 same test (items/order changes ok), 2 different; null if unclear
- test_measures_scoring_range: plain text description + range(s); if no numeric range => "scoring range not reported"
- mean_age_years: months/12; range only => null
- proportion_male: in [0,1] (percent/100; or count_male/total)
- source_mean_sd: "Text p.X" / "Table Y" / "Figure Z" (+ "; derived" if applicable). ASCII only.

VISUAL RULES
{VISUAL_RULES}

OUTPUT FORMAT (JSON only; no commentary/no markdown; double quotes; no trailing commas)
{
  "[Condition Name] - [Outcome Measure]": {
    "M_pre": null, "SD_pre": null, "M_post": null, "SD_post": null, "n": null,
    "country": null,
    "robot_model": null,
    "learning_subject_posttest": null,
    "participants_trained_at_a_time": null,
    "presence_human_tutor": null,
    "total_duration_training_min": null,
    "number_training_sessions": null,
    "posttest_delay_days": null,
    "similarity_pre_vs_post": null,
    "test_measures_scoring_range": null,
    "mean_age_years": null,
    "proportion_male": null,
    "source_mean_sd": null
  }
}

INPUT_CONTEXT
{INPUT_CONTEXT}
```

B.3. Extraction Prompt

This prompt was used for papers in which most models produced an incorrect number of data rows. In these cases, only the extraction step was run in order to assess whether low performance stemmed primarily from row identification errors or from the numerical extraction itself.

System Prompt: You are a precise Data Extraction Specialist.

Prompt Content

```
TASK
From INPUT_CONTEXT, extract all target keys and output ONE valid JSON object (JSON only).
The data might be available in the text, tables, or figures of INPUT_CONTEXT.

Target keys:
{formatted_targets}

VALUES (per key)
Stats:
- M_pre, SD_pre, M_post, SD_post, n.
- Post-test only: M_pre/SD_pre = null.
- Within/crossover: if one global pretest at study start, reuse it for all conditions; if condition-specific pretests exist, use
      matching one.
Missing/uncalculable => null.

SD derivation (only for a mean of same condition+timepoint, not differences/models):
If SD is not readily available, but other summary measures are (e.g., SE, 95% CI, or IQR), derive SD from these. If derived,
      append "; derived" to source_mean_sd.

Gender aggregation (same condition/outcome/timepoint):
If the same Outcome Measure is reported for males and females separately, merge these data.

Sample size n (use first available; if n_pre != n_post use n_post):
1) outcome+timepoint+condition n
2) timepoint+condition analyzed n
3) condition assigned/randomized n
4) else null

DESCRIPTORS (null if missing/unclear; use condition-level else study-level if clearly applies)
- country: ISO 3166-1 alpha-3
- robot_model: robot model/platform; null if non-robot condition
- learning_subject_posttest: subject/domain for this outcome/timepoint
- participants_trained_at_a_time: 1, 2, or "Group"(>2)
- presence_human_tutor: 1 if teaching/tutoring during training; 0 if not; null if unclear
- total_duration_training_min: total minutes across sessions (per-session*#sessions); range only => null
- number_training_sessions: integer
- posttest_delay_days: days from end of training to this post-test (weeks*7); immediate => 0; qualitative only => null
- similarity_pre_vs_post: 1 same test (items/order changes ok), 2 different; null if unclear
- test_measures_scoring_range: plain text description + range(s); if no numeric range => "scoring range not reported"
- mean_age_years: months/12; range only => null
- proportion_male: in [0,1] (percent/100; or count_male/total)
- source_mean_sd: "Text p.X" / "Table Y" / "Figure Z" (+ "; derived" if applicable). ASCII only.

VISUAL RULES
{VISUAL_RULES}

OUTPUT FORMAT (JSON only; no commentary/no markdown; double quotes; no trailing commas)
{formatted_json_example}

INPUT_CONTEXT
{INPUT_CONTEXT}
```

C

Supplementary Results, Statistical Comparisons, and Qualitative Case Inventories

This appendix reports supplementary visualizations, category-level performance tables, full numerical statistical comparisons, and detailed qualitative case inventories that support the main results section. All reported analyses are based on the final analyzed dataset after excluding Study 120, which was removed because the target values were also available in a table and therefore violated the single-source category rule.

C.1. Supplementary Visualizations

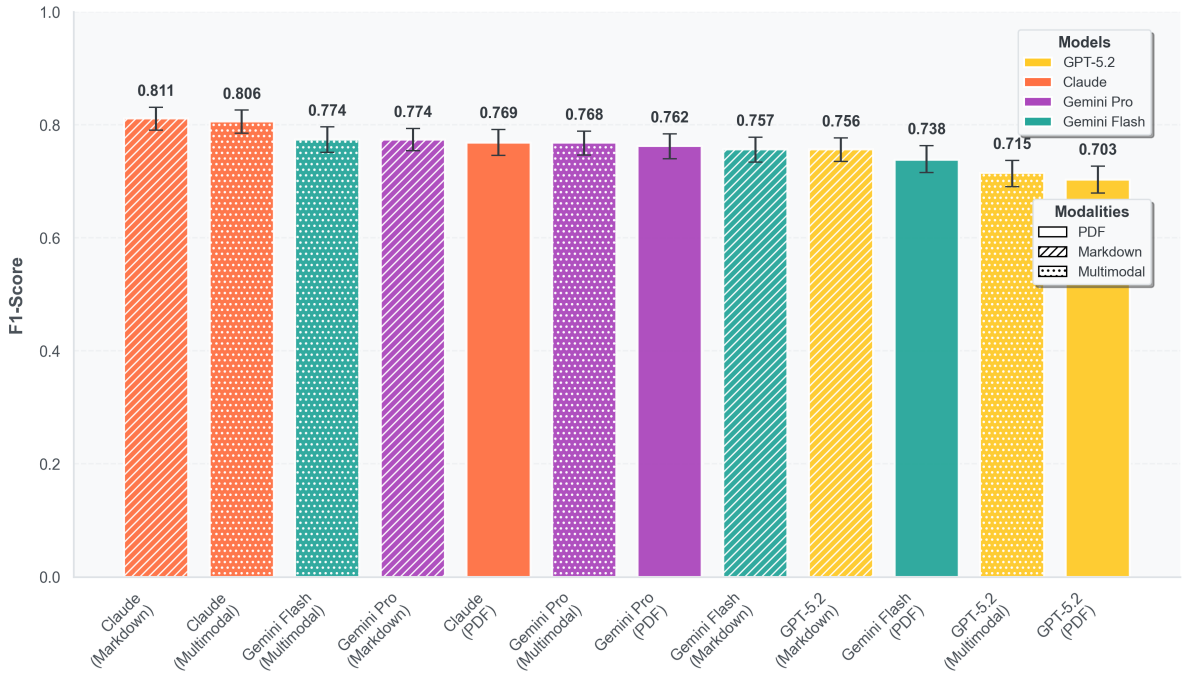

Figure C.1: Ranked overall F1-scores across all single-pass model-input representation conditions on the final analyzed dataset. Error bars denote 95% confidence intervals.

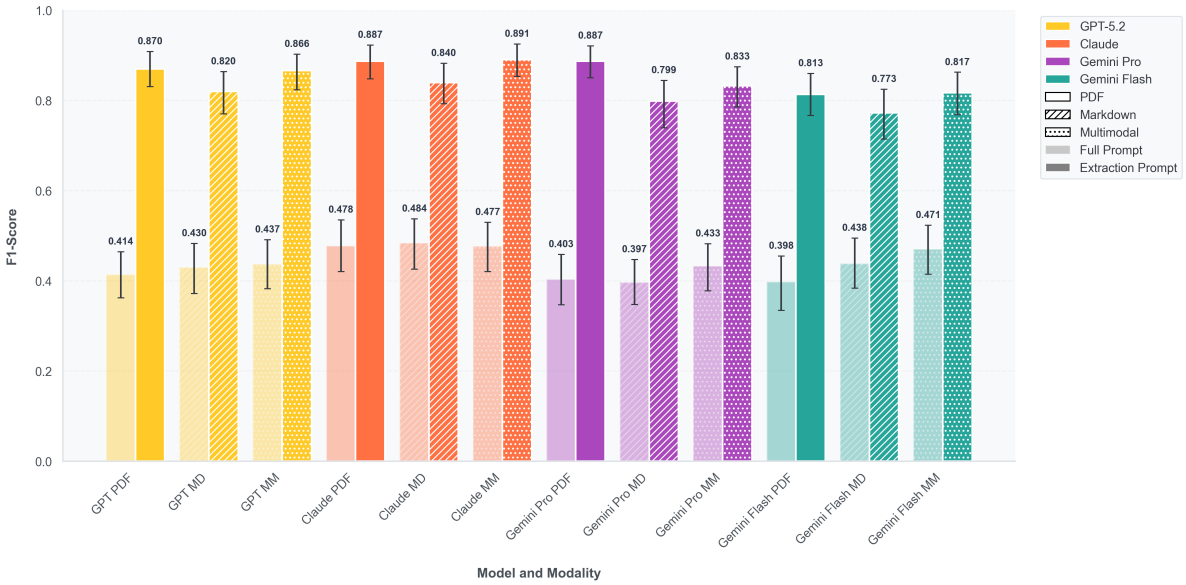

Figure C.2: Comparison of single-pass and target-condition extraction F1-scores on the 17-paper rerun subset. Error bars denote 95% confidence intervals.

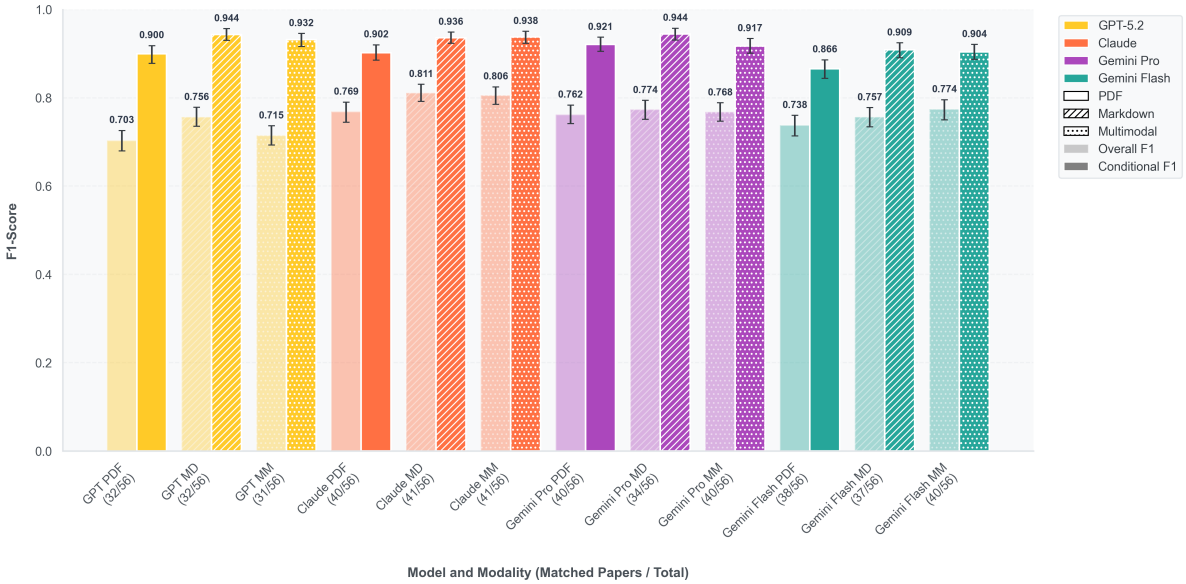

Figure C.3: Comparison of overall F1-scores and conditional F1-scores restricted to papers with exact row-count agreement in the final analyzed dataset. Error bars denote 95% confidence intervals.

	Wilcoxon *p*	Obuchowski *p*
GPT-5.2 PDF vs Markdown	0.0910	0.0716
GPT-5.2 PDF vs Multimodal	0.1803	0.0518
GPT-5.2 Markdown vs Multimodal	0.9498	0.4579
Claude PDF vs Markdown	0.0521	**0.0328**
Claude PDF vs Multimodal	0.1155	**0.0144**
Claude Markdown vs Multimodal	0.6029	1.0000
Gemini Pro PDF vs Markdown	0.9227	0.1001
Gemini Pro PDF vs Multimodal	0.3925	0.6171
Gemini Pro Markdown vs Multimodal	0.7646	0.3301
Gemini Flash PDF vs Markdown	0.3130	0.0985
Gemini Flash PDF vs Multimodal	0.0739	**0.0221**
Gemini Flash Markdown vs Multimodal	0.9341	0.7224

Figure C.4: Heatmap of pairwise input representation comparisons within each model group, showing paper-level Wilcoxon p-values and cell-level Obuchowski adjusted McNemar p-values.

C.2. Supplementary Category-Level Performance Table

Table C.1: Category-specific F1-scores for all single-pass model-input representation conditions in the final analyzed dataset. The Text and Table categories each contain 19 papers and 67 rows; the Figure category contains 18 papers and 59 rows after exclusion of Study 120.

Condition	Text	Table	Figure
GPT-5.2 PDF	0.8186	0.7217	0.4588
GPT-5.2 Markdown	0.7895	0.7604	0.6984
GPT-5.2 Multimodal	0.8391	0.7060	0.5532
Claude Opus 4.6 PDF	0.8992	0.8098	0.4885
Claude Opus 4.6 Markdown	0.8462	0.8431	0.7050
Claude Opus 4.6 Multimodal	0.8735	0.8333	0.6589
Gemini 3.1 Pro PDF	0.8331	0.7962	0.5856
Gemini 3.1 Pro Markdown	0.8031	0.7857	0.7075
Gemini 3.1 Pro Multimodal	0.8487	0.8027	0.5742
Gemini 3.1 Flash Lite Preview PDF	0.8078	0.7799	0.5671
Gemini 3.1 Flash Lite Preview Markdown	0.7821	0.7983	0.6531
Gemini 3.1 Flash Lite Preview Multimodal	0.8320	0.7977	0.6527

C.3. Detailed Statistical Comparisons

This section reports the full numerical results of the paired statistical comparisons between input representations within each model group. Paper-level differences were evaluated using the Wilcoxon signed-rank test on per-paper F1-scores. Cell-level differences were evaluated using Obuchowski's adjusted McNemar test for clustered matched-pair data.

Table C.2: Paper-level Wilcoxon signed-rank comparisons between input representations within each model group in the final analyzed dataset.

Model	Comparison	Mean F1 A	Mean F1 B	p
GPT-5.2	PDF vs Markdown	0.6795	0.7415	0.09096
GPT-5.2	PDF vs Multimodal	0.6795	0.7218	0.1803
GPT-5.2	Markdown vs Multimodal	0.7415	0.7218	0.9498
Claude Opus 4.6	PDF vs Markdown	0.7344	0.7851	0.05214
Claude Opus 4.6	PDF vs Multimodal	0.7344	0.7936	0.1155
Claude Opus 4.6	Markdown vs Multimodal	0.7851	0.7936	0.6029
Gemini 3.1 Pro	PDF vs Markdown	0.7475	0.7532	0.9227
Gemini 3.1 Pro	PDF vs Multimodal	0.7475	0.7564	0.3925
Gemini 3.1 Pro	Markdown vs Multimodal	0.7532	0.7564	0.7646
Gemini 3.1 Flash Lite Preview	PDF vs Markdown	0.6886	0.7290	0.313
Gemini 3.1 Flash Lite Preview	PDF vs Multimodal	0.6886	0.7421	0.07388
Gemini 3.1 Flash Lite Preview	Markdown vs Multimodal	0.7290	0.7421	0.9341

Table C.3: Cell-level Obuchowski adjusted McNemar comparisons between input representations within each model group in the final analyzed dataset.

Model	**Comparison**	Z	p
GPT-5.2	PDF vs Markdown	-1.8014	0.07164
GPT-5.2	PDF vs Multimodal	-1.9446	0.05182
GPT-5.2	Markdown vs Multimodal	0.7423	0.4579
Claude Opus 4.6	PDF vs Markdown	-2.1340	0.03284
Claude Opus 4.6	PDF vs Multimodal	-2.4475	0.01439
Claude Opus 4.6	Markdown vs Multimodal	0.0000	1
Gemini 3.1 Pro	PDF vs Markdown	-1.6444	0.1001
Gemini 3.1 Pro	PDF vs Multimodal	-0.5000	0.6171
Gemini 3.1 Pro	Markdown vs Multimodal	0.9739	0.3301
Gemini 3.1 Flash Lite Preview	PDF vs Markdown	-1.6523	0.09847
Gemini 3.1 Flash Lite Preview	PDF vs Multimodal	-2.2880	0.02214
Gemini 3.1 Flash Lite Preview	Markdown vs Multimodal	-0.3552	0.7224

C.4. Detailed Qualitative Case Inventories

C.4.1. GT Issues

Table C.4: GT issues identified during source-paper verification in the final analyzed dataset.

Study no.	Citation	Target condition	Column	Old value	New value	Remark
26	Resing et al. (2019)	Sham Agent - Schematic-picture completion (Total correct)	n	25	26	Control group size corrected
26	Resing et al. (2019)	Sham Agent - Schematic-picture completion (Number of body parts)	n	25	26	Control group size corrected
41	Donnermann et al. (2020)	Control Group - Digital Media Exam Grade	SD_{post}	1.96	1.00	Full error bar used as SD
41	Donnermann et al. (2020)	Robot Group (Pepper) - Digital Media Exam Grade	SD_{post}	1.01	0.51	Full error bar used as SD
43	Zhexenova et al. (2020)	Human Agent - New Script (Kazakh Latin) Conversion Test	M_{post}	14.82	14.63	Post mean corrected
43	Zhexenova et al. (2020)	Human Agent - New Script (Kazakh Latin) Conversion Test	SD_{post}	5.0618	5.33	Post SD corrected
57	Lanzilotti et al. (2021)	Robot Group (Pepper) - Computational Thinking (Coding) Exercises	n	42	23	Sample size corrected
91	Alves-Oliveira (2020)	Sham (Control) - Creativity in storytelling: Number of questions generated in 4 min about an image shown; CREA (learning gain)	SD_{post}	8.0327	6.3504	GT used 32 instead of N in SE-to-SD calculation
122	Alimardani et al. (2021)	Virtual Interface Group - L2 Vocabulary Matching Test	SD_{post}	0.3947	0.156	Whisker estimate used instead of IQR
122	Alimardani et al. (2021)	Robot Group (NAO) - L2 Vocabulary Matching Test	SD_{post}	0.4577	0.178	Whisker estimate used instead of IQR
126	Kanero et al. (2021)	Robot Group (NAO) - L2 Vocabulary Production Test (Immediate)	n	49	50	Voice tutor delayed sample size corrected
126	Kanero et al. (2021)	Robot Group (NAO) - L2 Vocabulary Receptive Test (Immediate)	n	49	50	Voice tutor delayed sample size corrected
145	Walker et al. (2016)	Virtual interface vTAG - 11 knowledge questions & Robot - 11 knowledge questions	n / condition label	10	13	GT rows were swapped
145	Walker et al. (2016)	Virtual interface vTAG - 11 knowledge questions & Robot - 11 knowledge questions	n / condition label	13	10	GT rows were swapped

C.4.2. Task-Definition and Paper-Structure Ambiguities

Table C.5: Observed task-definition and paper-structure ambiguity cases.

Study	Citation	Ambiguity category	Observed issue
1	Rintjema et al. (2018)	Row overgeneration	Task performance was extracted as additional rows.
4	Westlund et al. (2017)	Alternative metrics / combined arms	Alternative or secondary metrics were included and treatment arms were combined.
28	De Wit et al. (2018)	Combined treatment arms	Treatment arms were merged rather than kept in the GT row structure.
56	Sandygulova et al. (2020)	Boys and girls merged	Boys and girls were not merged but should have been merged to fit the ground truth
70	Velentza et al. (2021)	Invalid control group	Extracted comparisons lacked a valid control group or relied on a secondary metric.
91	Alves-Oliveira (2020)	Row overgeneration	Task performance was extracted as additional rows.
94	Hu et al. (2023)	Ineligible outcome structure	Knowledge acquisition was post-only and robot-only, without a valid pre-post or robot-control comparison.
27	Michaelis and Mutlu (2019)	Alternative metrics	Alternative or secondary metrics were extracted.
68	Kanda et al. (2004)	Delayed follow-up mismatch	A delayed post-test was present in the paper but did not align with the target row selection.
79	Hsiao et al. (2015)	Alternative metrics	Alternative or secondary metrics were extracted.
92	Riedmann et al. (2024)	Pre-test treated as condition	Pre-test measurements were extracted as separate conditions.
136	Hong et al. (2024)	Combined treatment arms	Treatment arms were merged
44	Kanda et al. (2012)	Ambiguous outcome grouping	“Robot construction kits” was treated as a separate row despite overlap with a broader achievement outcome.
75	Suzuki and Kanoh (2017)	Missing post-test data	Relevant post-test measurements were not available in the paper itself.
141	Jimenez and Kanoh (2013)	Alternative metrics	Alternative or secondary metrics were extracted.
146	Yoshizawa et al. (2020)	Alternative metrics	Alternative or secondary metrics were extracted.
147	Okawa et al. (2024)	Alternative metrics	Alternative or secondary metrics were extracted.

C.4.3. Genuine Model Extraction Failures

Table C.6: Observed genuine model extraction failures in the final analyzed dataset.

Study	Citation	Failure type	Observed issue
21	Chandra et al. (2016)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
34	Kennedy et al. (2015)	Dense-table value selection	Extraction failed for many values; the paper contains many means and SDs within a single table, making it difficult for the model to identify the correct ones.
42	Konijn et al. (2022)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
43	Zhexenova et al. (2020)	Derived-statistic mismatch	The extracted SD calculation differed from the GT procedure; the GT averaged subgroup SDs directly across subgroups, whereas models often used a pooled overall SD based on subgroup means, SDs, and sample sizes.
44	Kanda et al. (2012)	Imprecise figure-based extraction	SD values were not extracted correctly, and the extracted mean was close to the target but fell outside the tolerance. The figure was small and visually imprecise, with values shown only to one decimal place.
46	Levinson et al. (2021)	Shared pre-test value misassignment	A pre-test value applied to two rows, but models often assigned it to only one of them.
48	Schodde et al. (2019)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
52	Tanaka and Matsuzoe (2012)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
62	De Wit et al. (2020)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
72	Howley et al. (2014)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
74	Suzuki et al. (2015)	Missing SD values	SD values were frequently not extracted.

Study	Citation	Failure type	Observed issue
75	Suzuki and Kanoh (2017)	High-noise document extraction failure	Extraction failed for most target values, even under the target-condition extraction prompt. The paper contains substantial visual and textual noise, including many graphs and competing numeric values, and models often selected values from the text that were not the target values.
89	Lin et al. (2022)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
122	Alimardani et al. (2021)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
126	Kanero et al. (2021)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
141	Jimenez et al. (2013)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
146	Yoshizawa et al. (2020)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.
147	Okawa et al. (2024)	Missing SD values	No SD values were extracted, or multiple SD values were extracted incorrectly.

Table C.7 audits SD-related mismatches in figure-based cases from the final analyzed dataset where the interpretation of error bars or derived dispersion values contributed to the extraction difficulty. Study 120 was not included in this audit because it was excluded from the final analyzed dataset after post-selection verification showed that the target values were also available in a table. The column “Error-bar meaning clearly communicated” indicates whether the source paper clearly stated what the error bars represented, for example SD, SE, 95% CI, or another statistic. The column “Absolute error amplified by conversion” indicates whether a reported statistic such as SE, 95% CI, or IQR had to be converted to SD. In such cases, a small absolute interpolation error in the visually read statistic can become a larger absolute error after conversion to SD. Because the present scoring used a percentage-based tolerance, this conversion did not by itself change whether a value was classified as correct or incorrect when the same proportional error was preserved. However, it illustrates why a fixed absolute tolerance would be inappropriate for these cases: the acceptable absolute error depends on the scale of the original statistic, the conversion formula, and, in some cases, the sample size.

Table C.7: Paper-level audit of SD-related mismatches in figure-based extraction failures in the final analyzed dataset, focusing on whether the meaning of the error bars was explicitly stated and whether conversion from another dispersion statistic amplified the absolute error.

Study number	Cite	Reason for SD mismatch	Error-bar meaning clearly communicated	Absolute error amplified by conversion
42	Konijn et al. (2022)	95% CI was used, and this was clearly communicated in the source paper. The numerical tolerance was strict for this case, because after conversion a small interpolation error in the visually read value produced a larger absolute difference in the derived SD.	Yes	Yes
44	Kanda et al. (2012)	Only Gemini 3.1 Flash Lite Preview with Multimodal input attempted to extract the post-test SD. The values were taken from the wrong figure, namely Figure 7(b) instead of Figure 7(a).	No	No
48	Schodde et al. (2019)	Gemini 3.1 Flash Lite Preview with Multimodal input recovered the values correctly. Other models appeared to use the learning gain mentioned in the text and figures, which was easier to identify but did not match the target values. Gemini 3.1 Flash Lite Preview with PDF input did attempt to extract the figure data, but the resulting values fell outside the tolerance.	No	No
52	Tanaka and Matsuzoe (2012)	Some models filled in 0 instead of `null`. This likely happened because the pre-test selected only verbs that children had answered incorrectly, but this value was inferred from the study design rather than reported and should therefore have been `null`. Gemini 3.1 Flash Lite Preview with PDF input appeared to extract SD values directly, whereas the SE values should have been read from the figures and converted to SD. The figure did not explicitly state that the error bars represented SE.	No	No

Continued on next page

72	Howley et al. (2014)	The figure clearly communicated that the error bars represented SE, and the conversion was straightforward for a human reader. However, after converting SE to SD, only a small amount of visual interpolation error was possible before the derived SD exceeded the tolerance. For example, Claude Opus 4.6 with Markdown input returned a post-test SD of 18.3 where the GT required 15.5349. Converted back to SE using $n = 21$, this corresponds to approximately 3.99 for the model output and 3.39 for the GT. This is a small absolute visual difference, but not within the percentage tolerance after conversion. The two bars furthest from the y-axis were also interpolated outside the tolerance, raising the question whether bars further from the y-axis are harder for models to read accurately.	Yes	Yes
74	Suzuki et al. (2015)	The error bars in the figure were very thin. Only GPT-5.2 with Multimodal input attempted to extract the SD values and recovered several of them correctly, with only one small tolerance error. Many models likely left the SD values empty because they were being cautious. A more permissive prompt could therefore reduce false negatives by encouraging more extraction attempts, but it could also increase false positives by encouraging unsupported estimates.	No	No
75	Suzuki and Kanoh (2017)	The main problem was that the target condition was not specified clearly enough. Some values were extracted from the text rather than from the correct figure. The target condition should have made clearer that the score difference between follow-up and advance activity had to be used.	No	No
122	Alimardani et al. (2021)	The mismatch was caused by interpolation error followed by IQR-to-SD conversion. A small error in the approximated IQR led to a larger absolute error in the derived SD. This does not necessarily argue against the current percentage-based tolerance, but it does show why a fixed absolute error threshold would be inappropriate for these cases.	Yes	Yes
126	Kanero et al. (2021)	Only Gemini 3.1 Flash Lite Preview with Multimodal input attempted the extraction. The output showed a condition-linking mix-up: the selected score was higher and visually plausible, but it was not linked to the correct condition. This was the only case of this specific issue in the audit. The remaining mismatches in this paper appeared to be mainly tolerance-related.	Yes	Yes

Continued on next page

147	Okawa et al. (2024)	SE appeared to be assumed rather than explicitly given. Claude Opus 4.6 with PDF input appeared to calculate two values based on a 95% CI interpretation. Although “confidence interval” was mentioned once in the text, this interpretation did not fit the relevant figure, so the mismatch likely reflected an incorrect 95% CI-based inference.	No	No

C.4.4. Paper Inconsistencies

Table C.8: Observed source-paper inconsistencies identified during manual review.

Study	Citation	Inconsistency type	Observed issue
6	Alemi et al. (2014)	Graph-table inconsistency	The data reported in the graph did not match the data reported in the table.
81	Blancas-Muñoz et al. (2018)	Sample-size inconsistency	The reported sample size was inconsistent across the paper; the methodology and results sections reported different values ($n = 13$ and $n = 17$).