Expert-Guided Schema-Based Structured Extraction from CONSORT Diagrams Using Vision-Language Models

Damian Stachura, Bartosz Przechera, Monika Opałek,
Ewelina Sadowska, Ewa Borowiack, and Artur Nowak
Evidence Prime, Krakow, Poland
damian.stachura@evidenceprime.com

Abstract

Visual-language models (VLMs) are rapidly advancing on tasks that require visual understanding of text, tables, plots, and diagrams. Yet extracting structured information from text-heavy scientific diagrams remains challenging, as it requires not only OCR but also recovery of layout, grouping, and flow relationships. We study this problem in the context of CONSORT flow diagrams, which summarize participant screening, randomization, follow-up, and analysis in randomized controlled trials. We introduce a 200-example benchmark of PubMed Central diagrams, annotated by a biomedical team specializing in systematic literature reviews and clinical evidence extraction, and evaluate schema-constrained CONSORT extraction across proprietary and open-weight model families. Using structure-aware metrics, we compare single-pass and stepwise extraction strategies. Expert-guided single-pass extraction performs best for proprietary frontier models, with Gemini 3 Pro achieving the strongest overall results, whereas stepwise prompting improves less capable open-weight models on challenging arm-level extraction. These results offer practical deployment guidance and suggest that high-quality schema-constrained extraction is feasible, but not yet solved.

1 Introduction

Scientific articles contain rich structured knowledge, but extracting it remains challenging because evidence is distributed across multiple modalities, including narrative text, tables, charts, and diagrams. Recent vision-language models (VLMs) and multimodal document understanding systems have improved the ability to process visually rich scientific documents, making it increasingly feasible to extract information directly from figures and other non-textual layouts (Hu et al., 2023; Fujitake, 2024; Pramanick et al., 2024; Wang et al., 2025). However, text-heavy visual artifacts remain difficult, as successful extraction requires not only reading embedded text but also understanding layout, grouping, topology, and numerical relationships between visual elements.

A particularly important case arises in randomized controlled trials (RCTs), where key study information is often presented visually. Among these, CONSORT flow diagrams are especially valuable because they summarize participant screening, exclusions, randomization, follow-up, and analysis populations (Moher et al., 2010; Hopewell et al., 2011, 2025). However, these diagrams are highly heterogeneous in wording, layout, and completeness, making structured extraction of patient-flow information a challenging multimodal problem. An example is shown in Figure 1.

Prior work on CONSORT has largely focused on reporting support rather than figure understanding, including diagram generation, checklist extraction, and reporting-quality assessment (O'Leary et al., 2019; Wang et al., 2020; Laï-king and Paroubek, 2025). While broader research has explored methodological and multimodal evidence extraction from RCT publications (Hoang et al., 2023; Ghosh et al., 2024; Ji et al., 2026), extracting structured patient-flow data directly from CONSORT diagrams remains underexplored.

To address this gap, we introduce a benchmark for structured extraction from CONSORT flow diagrams. We collaborate with biomedical experts to define a schema capturing key patient-flow information and annotate a dataset of diagrams with substantial visual and reporting variability. Using this benchmark, we evaluate multiple VLM-based extraction strategies, including end-to-end schema filling with different prompting styles and stepwise approaches that decompose the task into smaller, visually coherent subproblems.

Our contributions are as follows:

- We introduce a 200-example benchmark for

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 955–969

structured extraction from CONSORT flow diagrams, with schema-aligned annotations prepared in collaboration with biomedical experts.

- We define and report a structure-aware evaluation suite for nested CONSORT flow outputs, including group-level, leaf-level, strict field-level, and diagnostic error metrics.
- We provide a capability profile across proprietary and open-weight VLMs under single-pass and stepwise prompting, and derive practical model-regime guidance for deployment.

2 Related work

Prior automation around CONSORT has primarily focused on reporting support rather than figure understanding. For example, O'Leary et al. (2019) proposed automatic generation of CONSORT diagrams from clinical-trial databases, while Wang et al. (2020) introduced CONSORT-NLP for generating reporting checklists from article PDFs. More recent work has evaluated large language models for CONSORT-oriented reporting-quality assessment (Laï-king and Paroubek, 2025). Related research has explored information extraction from RCT publications more broadly, including methodological extraction (Hoang et al., 2023), PICO extraction (Ghosh et al., 2024), and structured evidence extraction pipelines such as TrialMind and LatteReview (Wang et al., 2024b; Rouzrokh et al., 2025). While these approaches demonstrate progress toward structured clinical evidence extraction, they do not address recovery of patient-flow information from CONSORT diagrams, where key evidence is expressed visually.

The closest technical work comes from diagram and visually rich document understanding. Prior research has highlighted the challenges of OCR and figure-text extraction in scientific documents (Kim and Yu, 2011), while more recent approaches combine OCR, element detection, and generative reasoning for diagram parsing (Arbaz et al., 2024; Deka and Devereux, 2025). Multimodal systems have further explored diagram understanding (Hu et al., 2023), layout-aware document processing (Fujitake, 2024), and schema-constrained extraction (Wang et al., 2024a). These studies suggest that extraction from text-heavy visuals benefits from explicit structure recovery rather than treating images as generic captioning targets.

A related line of work studies reasoning over diagrams and flowcharts. Benchmarks such as FlowchartQA and FlowVQA evaluate multimodal reasoning in flowchart settings (Tannert et al., 2023; Singh et al., 2024), while FlowLearn highlights persistent weaknesses of VLMs in OCR, counting, and structural reasoning (Pan et al., 2024). On the modeling side, approaches such as TextFlow and arrow-guided VLMs explicitly incorporate graph structure and directional relationships to improve flowchart understanding (Ye et al., 2024; Omasa et al., 2025). These findings are particularly relevant because CONSORT diagrams represent participant flow through structured, text-bearing graph elements.

Recent work also explores agentic and tool-augmented approaches to multimodal extraction. For example, Chen et al. (2025) propose a multi-agent system for extracting structured information from scientific graphics, while systems such as ViperGPT demonstrate how visual reasoning can be decomposed into executable tool calls (Surís et al., 2023). These approaches suggest that modular reasoning and explicit structure recovery may be beneficial for complex visual extraction tasks.

Despite these advances, no prior work specifically targets structured extraction of patient-flow data from published CONSORT diagrams, leaving a clear gap for dedicated benchmarks and systematic evaluation of extraction strategies.

3 Benchmark

To evaluate extraction performance, we constructed a dedicated benchmark for CONSORT flowchart data extraction. This task remains relatively underexplored, and publicly available benchmarks are scarce, which motivated us to curate our own evaluation dataset.

3.1 Data Acquisition

We collected candidate CONSORT flow diagrams from PubMed Central articles published under CC-BY 4.0 licenses, which permit figure reuse. We first assembled a pool of 500 diagrams by randomly sampling eligible literature. From this pool, we selected 180 examples for benchmark release through a focused curation pass designed to preserve visual diversity, including layout styles, branch depth, and labeling conventions, as well as difficulty diversity, ranging from simple linear flows to structurally complex cases such as nested branches and partial

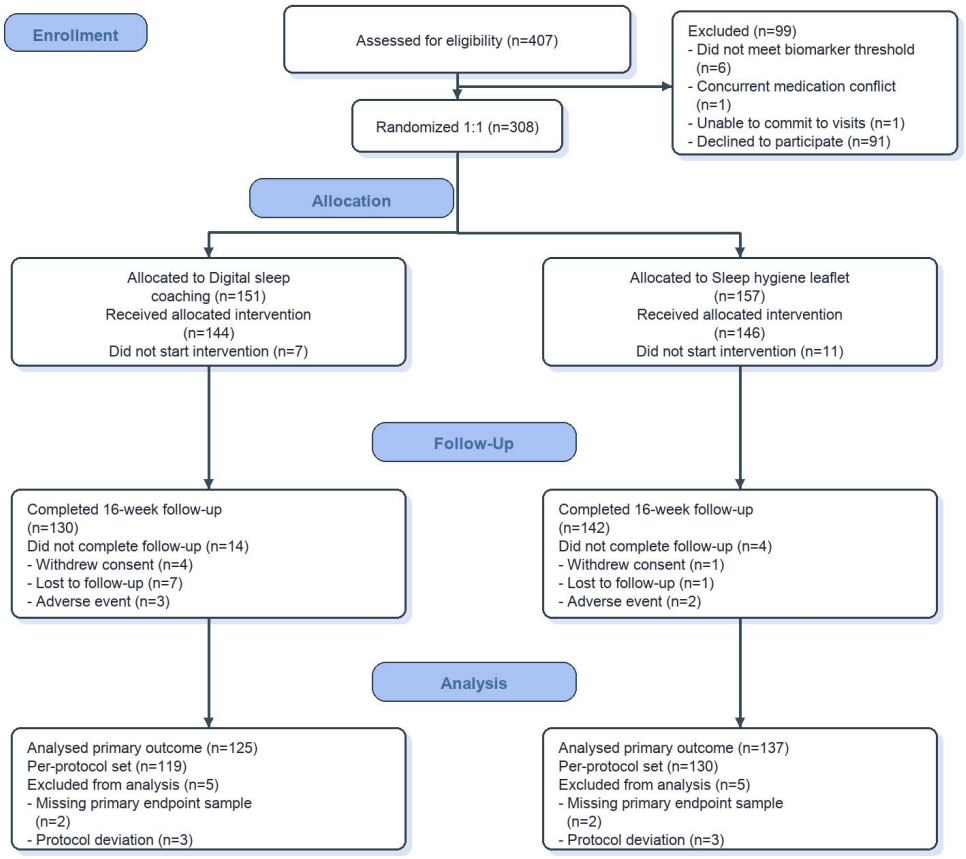

Figure 1: Synthetic CONSORT flowchart showing participant screening, randomization, follow-up, and analysis across intervention and control arms.

assessments. We focused on selecting only literature reporting RCTs. Additionally, we added 20 expert-curated examples that were considered challenging to reach a quick consensus on regarding how the data should be extracted. These 200 examples constitute the full evaluation benchmark used in this paper.

The 20 expert-curated diagrams are included as a deliberate stress-test subset to increase structural difficulty and probe failure modes under harder conditions. All systems are evaluated with the same fixed benchmark, schema, and scoring pipeline, so comparisons remain transparent and directly comparable across models and prompting regimes. We therefore interpret the reported results as a capability stress profile for structured CONSORT extraction rather than as a prevalence estimate for the full population of published CONSORT diagrams.

Each sample consists of a CONSORT flowchart paired with a structured representation following our schema. Annotation and verification were carried out by a team of three biomedical experts with background from masters and doctoral-level studies in biomedical domains and prior industry experience, specialized in systematic literature reviews and clinical evidence synthesis. The team aligned study-level and arm-level counts, exclusions, and analysis populations to canonical schema fields despite substantial variation in wording and layout.

3.2 Data Structure

CONSORT flowcharts are highly heterogeneous in both visual layout and textual content. Authors vary in how they organize diagrams, which counts they report, and the level of detail included in individual nodes. Despite this variability, most flowcharts can be mapped to a small number of recurring semantic components.

A typical CONSORT flowchart includes the following components:

- **Enrollment and screening:** the number of individuals assessed for eligibility, along with the number excluded prior to randomization and the reported reasons for exclusion.
- **Randomization and allocation:** the number of participants randomized and assigned to each study arm.
- **Follow-up and intervention status:** for each study arm, the number of participants who received the allocated intervention, did not receive it, completed treatment, discontinued treatment, or were lost to follow-up, together with corresponding reasons when reported.
- **Analysis:** the number of participants included in the primary outcome analysis for each arm, along with any exclusions from analysis and their reported reasons.

Our schema captures this structure at two levels. At the study level, it represents overall eligibility, pre-randomization exclusions, randomization, and the set of study arms. At the arm level, it captures participant trajectories within each arm, including allocation, treatment receipt, discontinuation, follow-up, and analysis. Supporting objects store finer-grained breakdowns of ineligibility, intervention discontinuation, analysis exclusions, and partial assessments. Both levels are illustrated in Figure 2 and summarized compactly in Table 1.

3.3 CONSORT Schema Fields

The target schema for CONSORT flow extraction is organized around a study-level object and a collection of per-arm objects. The fields we included in our schema were based on the CONSORT diagram schema presented in the literature (Hopewell et al., 2025) and we have chosen the fields whose experts flagged as most important in the process of random controlled trial analysis.

The study-level object tracks eligibility, exclusions, randomization, washout structure, and arm containers. Each arm object tracks allocation, treatment receipt and completion, follow-up losses, analysis populations, exclusions, and optional nested branches. A detailed field-by-field glossary for both levels is provided in Appendix A.

3.4 Data Annotation Protocol

The dataset was annotated using a two-stage procedure:

Component	Main content
Study-level flow	Eligibility, pre-randomization exclusions, randomization totals, washout indicator, arm containers, optional study-level analysis populations.
Arm identity	Arm label and ordering used for branch alignment.
Arm treatment flow	Allocation, treatment receipt/non-receipt, completion/non-completion, discontinuation reasons.
Arm follow-up and analysis	Follow-up losses, analysis populations, analysis-stage exclusions.
Optional advanced structure	Partial assessments and nested randomization branches.

Table 1: Compact summary of the CONSORT extraction schema used in the main experiments.

- Each expert independently mapped data from the CONSORT diagram to our predefined schema, which had been developed with the support of the same group of experts.
- Disagreements were adjudicated by the expert panel, and corrected labels were rechecked before dataset freeze.

4 Methods

Using this benchmark, we evaluated three VLM-based strategies for structured extraction from CONSORT flow diagrams in a zero-shot setting. The comparison was designed to isolate the effects of prompt scope, task decomposition, and post hoc verification while keeping the target output schema fixed.

- **Direct end-to-end extraction with basic prompt:** a single-pass baseline in which one VLM call using a basic system prompt receives the full diagram and optional caption and returns the complete target structure.
- **Direct end-to-end extraction with expert-guided prompt:** a single-pass baseline in which one VLM call with a prompt crafted by domain experts receives the full diagram and optional caption and returns the complete target structure.
- **Stepwise extraction:** a multi-pass strategy that reuses the same full diagram across a sequence of narrower prompts for study-level fields, washout detection, arm inventory, and per-arm flow extraction.

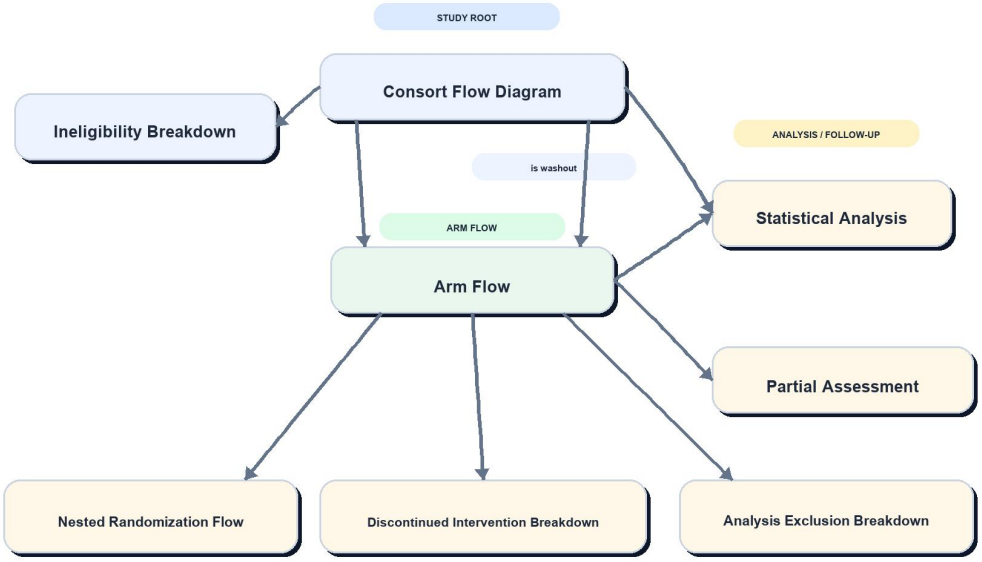

Figure 2: High-level schema of the CONSORT flow data model. The study-level CONSORT Flow Diagram captures overall enrollment and ineligibility information, a washout indicator, and arrays of Arm Flow objects before and after washout. Each study arm then records its own participant flow, including nested randomization, discontinuation, partial assessments, statistical analysis, and analysis-exclusion components.

4.1 Direct End-to-End Extraction

The first approach is the single pass setting. In this setting, a single VLM call is provided with the full CONSORT diagram, optionally accompanied by its caption, and is tasked with populating the complete CONSORT diagram structure in one step. This formulation serves as a direct benchmark of diagram understanding, as the model must jointly infer study-level counts, arm-level participant flows, and analysis populations without intermediate decomposition.

We evaluate two variants of prompting within this setting. The first uses a minimal prompt that guides the model to extract information according to the target schema, without additional domain-specific guidance. The second employs an extended prompt enriched with detailed definitions of schema fields and curated synonym lists, prepared by domain experts, to better align the model's interpretation with CONSORT reporting conventions.

4.2 Stepwise Extraction

The second approach is the stepwise setting. Instead of asking for the whole structure at once, the extractor runs a sequence of scoped passes over the same full diagram. The first pass extracts study-level enrollment and randomization fields, the second detects whether an explicit washout phase is present, and the third identifies the top-level arm inventory. The extractor then runs one additional pass for each arm to recover the detailed arm flow. These partial outputs are compiled deterministically into the final diagram structure. Importantly, this strategy does not crop the image into local regions. Rather, it keeps the full diagram available at every step and narrows the prompt scope to reduce structural ambiguity.

5 Experiments

This section presents the experimental setup and evaluation results for CONSORT diagram extraction.

The goal of the experiments is to characterize extraction capability across model families and scales, not only to rank a single best system. We therefore report structure-aware metrics, strict exact-match, and subgroup diagnostics to expose residual failure modes even when top-line scores are high.

5.1 Evaluation Metrics

We evaluate CONSORT diagram extraction using a structure-aware metric suite designed for nested participant-flow outputs. Since the task requires both accurate value extraction and correct assign-

ment of values to the appropriate arm, phase, or nested branch, our evaluation emphasizes structural correctness in addition to leaf-level accuracy.

Before scoring, both expected values and predicted outputs are normalized using the same schema semantics as the extraction system. This includes canonicalizing the CONSORT diagram representation, normalizing strings via case folding and whitespace collapsing, preserving exact integer counts, and resolving schema-level conventions such as washout structure. Object-valued lists are aligned by semantic identity rather than raw position, and objects representing arms, partial assessments, nested randomizations, exclusion reasons, and analysis populations are matched by their names.

We focus on three metrics that capture overall quality, structural correctness, and arithmetic coherence of extracted CONSORT flow representations.

CONSORT Group F1. This is the main structure-aware metric. It summarizes correctness across grouped components of the CONSORT diagram, including study-level structure, top-level arms, washout branches, nested randomizations, reason groups, partial assessments, and analysis populations. This metric is particularly important because a model may extract correct counts but still attach them to the wrong arm or branch.

Leaf Micro F1. This metric averages correctness over aligned atomic leaves. Numeric leaves are scored by exact match, while free-text identity fields such as arm names, assessment names, and reasons are evaluated using normalized string similarity. It reflects extraction quality at the value level once the hierarchical structure has been aligned.

Exact Field Accuracy. This metric measures the fraction of aligned schema fields that match exactly after normalization. It is stricter than the F1-style metrics because it does not award partial credit at the field level, but unlike full-example exact match it still captures partial success within a diagram.

Dataset-Level Aggregation. Metrics are computed per example and averaged over the evaluation set. We also report a normalized exact-match rate as a supplementary indicator, but not as a primary metric due to its stricter nature compared to the structure-aware measures above.

5.2 Technical Details

For the experiments, we evaluated all systems on the same 200-example benchmark. All runs used deterministic decoding with temperature 0 in a zero-shot setting. We did not provide figure captions during benchmarking, so the reported results reflect extraction from the diagram alone. For the stepwise system, the full image remained available at every stage.

Our primary model-selection objective was to measure the practical upper bound of capability under a fixed extraction protocol. We therefore used proprietary frontier models across size tiers to test what the strongest current systems can do on this task. The proprietary ablation in Table 2 shows a consistent pattern that expert-guided single-pass prompting is the strongest setting for these models.

We then evaluated open-weight models to estimate how far smaller openly available VLMs remain from proprietary systems on the same benchmark. This comparison is capability profiling under a common schema and evaluation pipeline, not a parameter-matched fairness claim. We also test smaller open-weight models because CONSORT extraction can be used as an agentic subtask in larger evidence-extraction workflows, where efficient local models may offer practical advantages.

Prompting setups are intentionally reported asymmetrically across regimes. For proprietary models, we run full basic/expert/stepwise ablations to establish the best prompting policy. For the newer open-weight 30B and Gemma runs, we focus on expert single-pass versus stepwise because this is the decision-relevant comparison after the proprietary ablation.

To assess the range of systems that can address this task, we used:

- GPT-5.1 and GPT-5 (OpenAI, 2025) mini with low reasoning effort deployed via Azure AI,
- Gemini 3 Flash Preview and Gemini 3 Pro Preview (DeepMind, 2025) with low thinking level accessed through the Google API,
- Qwen3-VL 8B Instruct (Bai et al., 2025) with 8-bit quantization (Ollama[1]), run locally on a MacBook M4 Max,
- Qwen3-VL 30B A3B Instruct (Bai et al., 2025) and Gemma-4-26B-A4B-it (Google

[1] https://ollama.com/

System	Mode	Group F1	Leaf Micro F1	Exact Field Acc.	Exact Match
Gemini 3 Pro	basic prompt	0.903	0.896	0.822	0.210
Gemini 3 Pro	expert prompt	**0.976**	**0.971**	**0.944**	**0.705**
Gemini 3 Pro	stepwise	0.903	0.891	0.809	0.155
Gemini 3 Flash	basic prompt	0.905	0.896	0.817	0.210
Gemini 3 Flash	expert prompt	0.964	0.957	0.926	0.625
Gemini 3 Flash	stepwise	0.907	0.895	0.818	0.165
GPT-5.1	basic prompt	0.874	0.843	0.743	0.050
GPT-5.1	expert prompt	0.872	0.832	0.732	0.065
GPT-5.1	stepwise	0.819	0.800	0.654	0.000
GPT-5 mini	basic prompt	0.850	0.812	0.675	0.055
GPT-5 mini	expert prompt	0.868	0.824	0.718	0.065
GPT-5 mini	stepwise	0.820	0.789	0.636	0.000

Table 2: Proprietary-model prompting ablation on the 200-example CONSORT benchmark (zero-shot, deterministic decoding, diagram-only input without figure captions). Expert-guided single-pass prompting is the strongest proprietary setting and is therefore used as the reference for open-weight comparisons.

Open-weight system	Mode	Group F1	Leaf Micro F1	Exact Field Acc.	Exact Match
Qwen3-VL 8B Instruct	expert prompt	0.701	0.593	0.375	0.000
Qwen3-VL 8B Instruct	stepwise	0.571	0.539	0.063	0.005
Qwen3-VL 30B A3B Instruct	expert prompt	0.601	0.495	0.379	0.000
Qwen3-VL 30B A3B Instruct	stepwise	0.633	0.611	0.435	0.005
Gemma-4-26B-A4B-it	expert prompt	0.635	0.541	0.068	0.000
Gemma-4-26B-A4B-it	stepwise	0.622	0.564	0.407	0.000

Table 3: Open-weight model comparison on the same 200-example benchmark and evaluation pipeline (zero-shot, deterministic decoding, diagram-only input).

DeepMind, 2026) via OpenRouter[2].

5.3 Results

Table 2 establishes prompting behavior on proprietary models. Across Gemini variants, expert-guided single-pass prompting clearly outperforms both basic and stepwise prompting on all primary metrics. Gemini 3 Pro with expert prompting is the strongest proprietary configuration overall, reaching 0.976 group F1, 0.971 leaf micro F1, 0.944 exact field accuracy, and 0.705 exact match.

Table 3 reports open-weight results for Qwen3-VL 8B Instruct, Qwen3-VL 30B A3B Instruct, and Gemma-4-26B-A4B-it in expert single-pass and stepwise modes. In this lower-capability regime, expert single-pass fails on complex arm-level extraction for Qwen3-VL 30B A3B Instruct and Gemma-4-26B-A4B-it. For Qwen3-VL 30B, stepwise improves Group F1 from 0.601 to 0.633, Leaf Micro F1 from 0.495 to 0.611, and Exact Field Accuracy from 0.379 to 0.435; for Gemma-4-26B-A4B-it, stepwise improves Leaf Micro F1 from 0.541 to 0.564 and Exact Field Accuracy from 0.068 to 0.407 (with a small Group F1 decrease from 0.635 to 0.622).

This suggests that decomposition can help smaller models, but the current full-context stepwise design is still limited. A promising next step is arm-localized extraction: first detect arms, then run extraction on per-arm crops instead of repeatedly presenting the full diagram context.

From a deployment perspective, these results support a regime-dependent policy: use expert single-pass prompting for frontier proprietary systems, and prefer stepwise prompting for smaller open-weight systems when arm-level complexity is high.

Despite strong top-line scores for the best systems, strict exact match remains below one for all completed models (best: 0.705), and the appendix subgroup diagnostics show persistent errors on structurally complex components. This supports the benchmark's utility as a non-trivial capability test rather than an already-solved task.

5.4 Error Analysis

We observe several recurring error types in our results. First, some errors arise from ambiguity in the source diagrams rather than from complete extraction failure. For example, the number of patients who received treatment is sometimes only inferable from related counts, such as the number allocated to treatment and the number who did

[2] https://openrouter.ai

not receive treatment, but is not stated explicitly. Similar ambiguity appears when assigning reasons to schema fields such as other reasons or missing data. In such cases, models differ in how well they recover the intended structure. Gemini performs best overall, especially when the correct field can be inferred despite slight mismatches between the diagram wording and the target schema.

Second, all models struggle with structurally complex diagrams, including nested randomization steps and partial assessments reported for example at selected follow-up time points. These cases require both accurate parsing of the study flow and correct attachment of local counts to the appropriate fields. Performance decreases for all models on such examples, although Gemini remains the most robust.

We also observe several model-specific failure modes. Qwen3-VL 8B Instruct frequently inserts 0 into fields that are not explicitly provided instead of leaving them unspecified. While this can leave the high-level structure looking superficially plausible, it still lowers exact field accuracy and constitutes a mild hallucination, especially because the prompts explicitly discourage this behavior. In some cases, Qwen also appears to infer the correct number of study arms but fails to generate the corresponding arm objects, resulting in an empty arm list.

Finally, some errors are caused by low-level text interpretation. Subscript and superscript notation in more complex diagrams can lead to transcription mistakes. In addition, Qwen and GPT models sometimes extract full descriptive phrases instead of normalized arm or assessment names, for example `allocated to control group` instead of `control group`. These errors are relatively minor, but they reduce output consistency and make downstream normalization more difficult.

6 Future Work

CONSORT flow diagrams are complex visual summaries, and our current benchmark captures the concepts that can be most reliably standardized across studies. Future versions should expand coverage to additional structured patterns, including cluster-level versus individual-level flows, subgroup analyses versus formal re-randomization, and multi-stage designs involving multiple study-wide randomization events.

We also plan to improve evidence grounding by detecting study arms and key nodes, extracting information from localized crops, and evaluating this approach across a broader set of open-weight VLMs.

7 Conclusion

This paper presents a benchmark and capability profile study of schema-constrained CONSORT diagram extraction. Our experiments show that structured extraction from CONSORT flow diagrams is feasible with modern VLMs when grounded in a fixed schema and evaluated with structure-aware metrics. Across completed systems, the best results come from direct single-pass extraction with an expert-guided prompt, which provides the strongest balance of structural fidelity, leaf-level accuracy, and exact field recovery. The top configuration achieves a CONSORT group F1 of 0.976, a leaf micro F1 of 0.971, an exact field accuracy of 0.944, and an exact-match rate of 0.705, indicating that end-to-end extraction can recover a large fraction of complete diagrams without intermediate decomposition.

Prompting behavior depends on model capability. For the strongest proprietary models, expert-guided single-pass extraction is consistently best. For smaller open-weight models, however, single-pass settings fail on complex arm-level information, and stepwise prompting can improve key metrics by narrowing each extraction step.

These findings support decomposition, but also show limits of the current full-context stepwise design, where every pass still sees the whole diagram. A more promising next step is arm-localized decomposition: detect study arms and key nodes first, then extract from per-arm crops before merging outputs. This should reduce irrelevant context and improve local OCR and attachment decisions, especially for smaller open-weight models. At the same time, the gap between high top-line accuracy and imperfect exact-match confirms that the benchmark remains challenging and informative for capability profiling.

Acknowledgments

This research was co-funded by the European Union – European Regional Development Fund (Programme: European Funds for a Modern Economy 2021-2027, grant no. FENG.01.01-IP.02-4479/23).

References

Abdul Arbaz, Heng Fan, Junhua Ding, Meikang Qiu, and Yunhe Feng. 2024. Genflowchart: Parsing and understanding flowchart using generative ai. In *Knowledge Science, Engineering and Management: 17th International Conference, KSEM 2024, Birmingham, UK, August 16–18, 2024, Proceedings, Part I*, volume 14884 of *Lecture Notes in Computer Science*, pages 99–111. Springer.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Yufan Chen, Ching Ting Leung, Bowen Yu, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. 2025. A multi-agent system enables versatile information extraction from the chemical literature. *arXiv preprint arXiv:2507.20230*.

Google DeepMind. 2025. Gemini 3: Flash and pro models. https://deepmind.google.

Pritam Deka and Barry Devereux. 2025. Structured extraction from business process diagrams using vision-language models. *arXiv preprint arXiv:2511.22448*.

Masato Fujitake. 2024. Layoutllm: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10219–10224, Torino, Italia. ELRA and ICCL.

Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. Alpapico: Extraction of pico frames from clinical trial documents using llms. *Methods*, 226:78–88.

Google DeepMind. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-20.

Linh Hoang, Yingjun Guan, and Halil Kilicoglu. 2023. Methodological information extraction from randomized controlled trial publications: a pilot study. In *AMIA Annual Symposium Proceedings*, pages 542–551.

Sally Hopewell, An-Wen Chan, Gary S Collins, Hróbjartsson A., D. Moher, K. F. Schulz, and 1 others. 2025. Consort 2025 statement: updated guideline for reporting randomised trials. *BMJ*, 388:e081123.

Sally Hopewell, Allison Hirst, Gary S. Collins, Sue Mallett, Ly-Mee Yu, and Douglas G. Altman. 2011. Reporting of participant flow diagrams in published reports of randomized trials. *Trials*, 12:253.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2023. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. *arXiv preprint arXiv:2311.18248*.

Changkai Ji, Yang Li, Yingwen Wang, Wen He, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2026. Evimmqa: Multimodal question answering for medical evidence extraction in systematic reviews. *Pattern Recognition*, page 113441. Available online 5 March 2026.

Daehyun Kim and Hong Yu. 2011. Figure text extraction in biomedical literature. *PLoS ONE*, 6(1):e15338.

Mathieu Laï-king and Patrick Paroubek. 2025. Evaluation of clinical trials reporting quality using large language models. *arXiv preprint arXiv:2510.04338*.

David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, PJ Devereaux, Diana Elbourne, Matthias Egger, and Douglas G Altman. 2010. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c869.

Teresa O'Leary, June Weiss, Benjamin Toll, Cynthia Brandt, and Steven L. Bernstein. 2019. Automated generation of consort diagrams using relational database software. *Applied Clinical Informatics*, 10(1):60–65.

Takamitsu Omasa, Ryo Koshihara, and Masumi Morishige. 2025. Arrow-guided vlm: Enhancing flowchart understanding via arrow direction encoding. *arXiv preprint arXiv:2505.07864*.

OpenAI. 2025. Gpt-5 technical report. https://openai.com.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. Flowlearn: Evaluating large vision-language models on flowchart understanding. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, pages 73–80. IOS Press.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118807–118833.

Pouria Rouzrokh, Bardia Khosravi, Parsa Rouzrokh, and Moein Shariatnia. 2025. Lattereview: A multi-agent framework for systematic review automation using large language models. *arXiv preprint arXiv:2501.05468*.

Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. 2024. Flowvqa: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational*

Linguistics: ACL 2024, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Simon Tannert, Marcelo G. Feighelstein, Jasmina Bogojeska, Joseph Shtok, Assaf Arbelle, Peter W. J. Staar, Anika Schumann, Jonas Kuhn, and Leonid Karlinsky. 2023. Flowchartqa: The first large-scale benchmark for reasoning over flowcharts. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 34–46, Ingolstadt, Germany. Association for Computational Linguistics.

Fan Wang, Richard L. Schilsky, David Page, Robert M. Califf, Kei Cheung, Xiaofei Wang, and Herbert Pang. 2020. Development and validation of a natural language processing tool to generate the consort reporting checklist for randomized clinical trials. *JAMA Network Open*, 3(10):e2014661.

Fei Wang, Yuewen Zheng, Qin Li, Jingyi Wu, Pengfei Li, and Luxia Zhang. 2024a. Chatschema: A pipeline of extracting structured information with large multimodal models based on schema. *arXiv preprint arXiv:2407.18716*.

Xingbo Wang, Samantha L. Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2025. Scidasynth: Interactive structured data extraction from scientific literature with large language model. *Campbell Systematic Reviews*, 21(4):e70073.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2024b. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv:2406.17755*.

Junyi Ye, Ankan Dash, Wenpeng Yin, and Guiling Wang. 2024. Beyond end-to-end vlms: Leveraging intermediate text representations for superior flowchart understanding. *arXiv preprint arXiv:2412.16420*.

A Detailed CONSORT Schema Glossary

This appendix provides a field-level description of the schema used for structured extraction from CONSORT diagrams. The goal is to make mapping decisions explicit for both annotation and evaluation.

A.1 Study-Level Structure

The study-level object captures trial-wide participant flow before arm-specific trajectories are resolved. It includes enrollment, pre-randomization filtering, randomization totals, optional post-randomization exclusions, washout structure, and global analysis-population records.

A.2 Arm-Level Structure

Each arm object captures the participant trajectory within one trial arm. The structure covers allocation, treatment receipt/completion, follow-up losses, analysis populations, exclusions, and optional internal branching such as nested randomization.

A.3 Expert Definition Alignment and Synonym Guidance

To improve annotation consistency and prompt robustness, we aligned key schema fields with expert-authored definitions and synonym inventories from our domain guidance artifact. Tables 4 and 5 summarize the mappings used during data preparation and error analysis.

Figures 3 and 4 provide structural companions to Tables 4 and 5, visualizing study-level and arm-level hierarchy components.

The expert definitions emphasize core enrollment and attrition fields, while advanced structural components in our schema (notably `partial_assessments`, `nested_randomizations`, and washout-specific branch pairing via `arms_after_washout`) require additional schema-specific normalization rules beyond direct synonym matching. We retain these components because they are required for full structural fidelity in complex CONSORT diagrams.

B Exploratory Data Analysis of Dataset

To complement the field glossary, we report corpus-level exploratory analysis of the 200-example dataset. This links schema semantics to empirical prevalence and sparsity patterns, and clarifies where extraction reliability is most sensitive to structural complexity.

B.1 Dataset Shape and Structural Prevalence

The benchmark contains $N = 200$ diagrams. The number of top-level arms per diagram is typically small (mean 2.33, median 2, range 0–9), and most examples are two-arm trials ($155/200$, 77.5%). Washout structure is uncommon (`is_washout`=true in $7/200$, 3.5%), with non-empty `arms_after_washout` in $8/200$ (4.0%). Partial assessment structures appear in $84/200$ diagrams (42.0%), while nested randomization is rare ($4/200$, 2.0%).

B.2 Field Coverage and Sparsity by Schema Level

Table 6 summarizes non-null prevalence for core schema fields.

C Additional Correctness Metrics

For detailed error analysis, we additionally report a broader set of diagnostic metrics in the appendix. Overall, the appendix metrics provide a more fine-grained view of model behavior and failure modes.

Structure-Specific Metrics. Arm inventory F1 evaluates whether the model identifies the correct set of top-level arms. Washout structure accuracy measures correct washout detection and correct matching of pre-washout and post-washout branches. Nested randomization accuracy evaluates whether internal randomization steps and their child arms are correctly represented.

Semantic Subgroup Metrics. Reason group F1 focuses on structured reason extraction, including ineligibility, non-receipt, discontinuation, and assessment-level reason groups. Analysis population F1 evaluates study-level and arm-level analysis populations, including named populations such as intention-to-treat, modified intention-to-treat, per-protocol, and other explicitly labeled analyses.

Numeric Error Metrics. For count-valued leaves, we report mean absolute error (MAE), and symmetric mean absolute percentage error (sMAPE):

$$\texttt{count\_mae} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|\,, \tag{1}$$

$$\texttt{count\_smape} = \frac{1}{N}\sum_{i=1}^{N} \frac{2\,|\hat{y}_i - y_i|}{|y_i| + |\hat{y}_i|}. \tag{2}$$

We summarize metrics from last three paragraphs in Table 7.

Edit-Based Diagnostics. We also summarize edit-operation statistics over the full validation set, including scalar count edits, scalar label edits, object insertions/deletions, reason insertions/deletions, and branch-repair operations. These diagnostics help characterize the types of errors made by each system and complement the aggregate quantitative metrics. We present them in Table 8.

Field	Expert definition	Synonym cues	Disambiguation rule
patients assessed for eligibility	All individuals screened/assessed for potential enrollment before allocation.	screened for eligibility; approached; assessed for enrolment/enrollment	Do not substitute enrolled or randomized counts when screening is unreported.
patients excluded before randomization total	All screened individuals not advancing to randomization for any pre-randomization reason.	patients excluded; ineligible; not enrolled; excluded during screening	Keep this as an aggregate pre-randomization total; do not merge post-randomization losses.
ineligibility.inclusionCriteriaNotMet	Excluded because inclusion criteria were not satisfied.	did not meet inclusion criteria; not eligible	Use only criterion-failure cases, not refusal or administrative reasons.
ineligibility.declinedToParticipate	Screened/approached individuals who declined consent/participation.	refused to participate; patient refusal; declined consent	Do not collapse into generic `other_reason` when refusal is explicit.
ineligibility.screeningFailure	Screening-stage failures preventing enrollment.	screening failure; not qualified at screening	Reserve for explicit screening-stage failures; not broader non-completion after randomization.
ineligibility.otherReason	Pre-randomization exclusions outside canonical categories.	other reasons; miscellaneous exclusions	Use only when no canonical ineligibility subtype applies.
patients randomized	Participants formally assigned by random allocation (including crossover sequences).	randomly assigned; underwent randomization; allocated to sequence	Prefer explicit randomization totals; avoid deriving from downstream treatment counts alone.
post randomization pre allocation exclusions	Participants excluded after randomization but before arm-specific allocation/treatment nodes.	excluded after randomization before allocation; withdrew before allocation; post-randomization exclusions	Use only for explicit between-stage losses; do not merge with pre-randomization exclusions or arm-level attrition.
is washout	Boolean indicator that a dedicated washout phase exists as a structural node.	washout period; washout phase; post-period washout	Set true only when washout is represented structurally in the flow, not when mentioned narratively.
statistical analysis	Study-level analyzed populations reported outside arm branches (e.g., ITT, mITT, safety, efficacy).	intention-to-treat; modified ITT; per-protocol; safety set; efficacy analysis set	Record only global analysis sets here; arm-specific analysis populations belong in arm objects.
arms	Ordered list of top-level arm flow objects before washout, or the sole arm list when no washout exists.	intervention arm; control arm; treatment group; sequence arm	Preserve diagram branch order to maintain deterministic alignment and downstream comparisons.
arms after washout	Ordered list of post-washout arm flows aligned one-to-one with pre-washout arm order.	post-washout arm; after washout sequence	Populate only when explicit washout splitting is present; keep positional correspondence with pre-washout arms.

Table 4: Study-level mapping between canonical schema fields and expert-authored definitions/synonym cues.

Canonical field	Condensed expert definition	High-value synonym cues	Disambiguation rule
arm name	Arm label as reported in the diagram (e.g., intervention, control, sequence name).	intervention arm; control arm; placebo arm; sequence AB/BA	Normalize superficial formatting only; preserve source arm semantics and role distinctions.
patients allocated to intervention	Participants assigned to an arm at allocation/randomization branch entry.	assigned to intervention; randomized per arm; allocated to sequence	Do not replace with received/completed treatment counts.
patients received allocated intervention	Allocated participants who initiated the assigned intervention.	received intended treatment; initiated assigned treatment	Exclude participants who never started treatment.
patients did not receive allocated intervention	Allocated participants who never initiated assigned treatment.	did not receive allocated intervention; never started treatment	Do not combine with treatment discontinuation after initiation.
patients completed treatment	Participants completing planned treatment course.	completers; completed protocol; completed intervention	Completion implies initiation; do not map follow-up completion to this field.
patients did not complete treatment	Participants who started but did not complete treatment.	drop-outs; withdrew from treatment; did not finish treatment	Keep distinct from non-receipt and analysis-stage exclusion.
discontinued intervention.*	Reason-structured treatment discontinuation after initiation.	due to adverse events; lack of efficacy; protocol violation; withdrawal	Attach only reasons linked to discontinuation nodes for that arm.
lost to follow up for primary outcome	Started treatment but missing primary-outcome follow-up data because participant could not be followed.	lost to follow-up; unable to contact; missed follow-up	Do not merge with explicit analysis exclusion categories unless diagram conflates them.
excluded from primary outcome analysis.*	Participants excluded at analysis stage with reason breakdown.	excluded from analysis; no outcome data; major protocol violation	Apply only to analysis-stage exclusions, not earlier flow attrition.
statistical analysis	Arm-specific analyzed populations (e.g., ITT, mITT, per-protocol, labeled sets).	analyzed population; ITT set; mITT set; per-protocol set; safety set	Keep distinct named populations when reported; avoid collapsing multiple analysis sets into one count.
partial assessments	Interim follow-up checkpoints (often timepoint-based) represented separately from final analysis totals.	week 12 assessed; month 6 follow-up; interim assessment	Represent each checkpoint as a separate structured object with its own count and optional reason groups.
nested randomizations	Secondary randomization branches within a top-level arm.	re-randomized; second randomization; split within arm	Create child branches only for explicit formal re-randomization steps, not ordinary subgroup reporting.

Table 5: Arm-level mapping between canonical schema fields and expert-authored definitions/synonym cues.

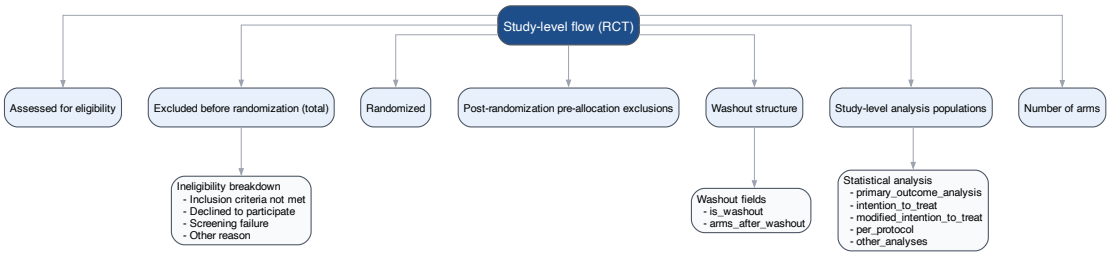

Figure 3: Study-level RCT patient-flow hierarchy aligned with Table 4.

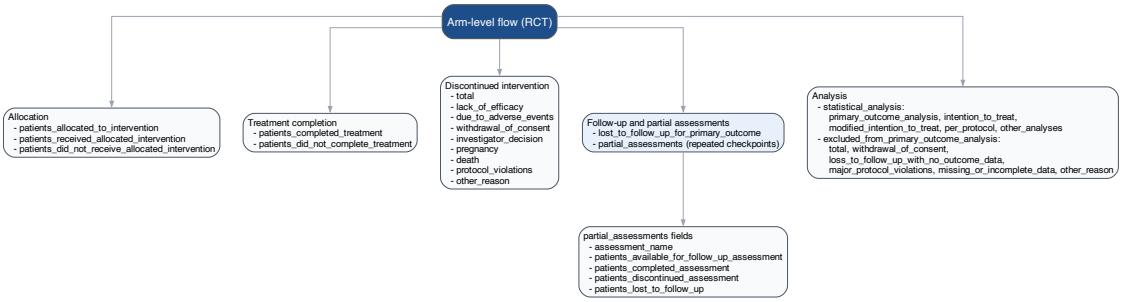

Figure 4: Arm-level RCT patient-flow hierarchy aligned with Table 5.

Study-level field	**Non-null (%)**	**Arm-level field (over all arms)**	**Non-null (%)**
patients_assessed_for_eligibility	89.0	patients_allocated_to_intervention	98.5
patients_excluded_before_randomization_total	82.5	patients_received_allocated_intervention	35.8
patients_randomized	94.5	patients_did_not_receive_allocated_intervention	23.6
post_randomization_pre_allocation_exclusions	5.0	patients_completed_treatment	14.3
statistical_analysis	9.0	patients_did_not_complete_treatment	3.0
ineligibility	83.0	discontinued_intervention	36.2
number_of_arms	100.0	lost_to_follow_up_for_primary_outcome	19.5
is_washout	100.0	statistical_analysis	63.0
		excluded_from_primary_outcome_analysis	21.2
		partial_assessments	42.0
		nested_randomizations	1.5

Table 6: Observed schema-field coverage in the 200-example gold-standard dataset used for all experiments. Study-level percentages are computed per diagram; arm-level percentages are computed over all arm objects.

Model	Setup	Study	Arm	Washout	Partial	Reasons	Analysis	Nested	MAE	sMAPE
Gemini 3 Pro	expert prompt	0.987	0.989	0.999	0.966	0.963	0.978	0.981	0.06	0.015
	basic prompt	0.975	0.933	0.998	0.797	0.863	0.904	0.963	3.10	0.042
	stepwise	0.937	0.921	0.994	0.874	0.851	0.906	0.959	8.67	0.067
Gemini 3 Flash	expert prompt	0.977	0.978	0.997	0.951	0.942	0.972	0.976	0.09	0.038
	basic prompt	0.976	0.929	0.998	0.802	0.862	0.908	0.964	3.07	0.041
	stepwise	0.953	0.925	0.993	0.874	0.856	0.906	0.959	8.64	0.037
GPT-5.1	expert prompt	0.917	0.881	0.993	0.838	0.761	0.912	0.988	2.08	0.034
	basic prompt	0.915	0.879	0.991	0.828	0.790	0.903	0.988	12.15	0.036
	stepwise	0.827	0.832	0.992	0.788	0.738	0.832	0.982	6.52	0.185
GPT-5 mini	expert prompt	0.952	0.877	0.993	0.802	0.776	0.872	0.984	9.40	0.039
	basic prompt	0.903	0.863	0.992	0.813	0.749	0.858	0.982	19.49	0.070
	stepwise	0.925	0.802	0.991	0.733	0.741	0.813	0.975	12.68	0.044
Qwen3-VL 8B Instruct	expert prompt	0.864	0.574	0.984	0.746	0.469	0.763	0.935	34.46	0.146
	stepwise	0.094	0.654	0.987	0.743	0.421	0.771	0.991	2.93	1.963
Qwen3-VL 30B A3B Instruct	expert prompt	0.807	0.407	0.967	0.671	0.428	0.499	0.991	27.16	0.280
	stepwise	0.816	0.669	0.987	0.373	0.498	0.534	0.932	19.14	0.267
Gemma-4-26B-A4B-it	expert prompt	0.624	0.654	0.987	0.743	0.363	0.640	0.991	43.27	0.602
	stepwise	0.699	0.684	0.986	0.679	0.363	0.493	0.986	48.51	0.116

Table 7: Section-level breakdown from aggregated evaluation reports on the 200-example benchmark under the same normalization and matching pipeline. Study through Nested are aggregated subsection scores, where higher is better. MAE and sMAPE are count-valued numeric error metrics, where lower is better. Partial-assessment scores are computed on the 42.0% of examples where that structure is present, and nested-randomization scores are computed on the 2.0% of examples where nested randomization appears.

Model	Setup	Count	Label	Obj+	Obj-	Reason+	Reason-	Branch	Total
Gemini 3 Pro	expert prompt	1.45	0.39	0.53	0.05	1.65	0.05	0.58	4.69
	basic prompt	4.82	1.30	1.43	0.70	5.50	0.69	2.28	16.70
	stepwise	4.28	1.77	1.77	0.38	6.12	0.37	2.33	17.01
Gemini 3 Flash	expert prompt	1.84	0.52	0.58	0.15	2.93	0.07	0.77	6.84
	basic prompt	4.78	1.35	1.47	0.72	5.62	0.69	2.34	16.96
	stepwise	4.16	1.67	1.61	0.43	5.95	0.37	2.23	16.40
GPT-5.1	expert prompt	6.26	3.18	2.28	0.79	9.00	0.42	3.57	25.48
	basic prompt	6.56	2.94	2.10	0.99	7.30	0.79	3.60	24.26
	stepwise	7.82	3.59	2.98	0.82	8.94	0.54	4.32	28.99
GPT-5 mini	expert prompt	6.75	3.30	2.43	0.68	9.87	0.50	3.39	26.90
	basic prompt	7.72	3.67	2.96	1.30	10.53	0.99	4.73	31.89
	stepwise	8.22	4.09	3.34	1.40	10.29	0.69	5.00	33.00
Qwen3-VL 8B Instruct	expert prompt	17.46	3.68	4.41	0.55	13.96	0.44	5.09	45.58
	stepwise	15.69	5.81	10.19	0.00	27.35	0.01	10.12	69.16
Qwen3-VL 30B A3B Instruct	expert prompt	21.36	3.31	3.19	1.62	12.42	0.85	5.20	47.93
	stepwise	14.53	3.68	4.08	2.66	19.05	0.86	6.84	51.68
Gemma-4-26B-A4B-it	expert prompt	13.50	5.79	10.15	0.01	27.47	0.00	10.12	67.03
	stepwise	12.25	3.14	3.38	1.08	25.73	0.02	4.53	50.12

Table 8: Mean edit operations per example from `final_metrics_jsons` on the 200-example benchmark; lower is better. Columns report scalar count edits, scalar label edits, object insertions/deletions, reason insertions/deletions, branch-repair operations, and total edit operations.