Enhancing Study-Level Inference from Clinical Trial Papers via Reinforcement Learning-Based Numeric Reasoning

Massimiliano Pronesti[1,2], Michela Lorandi[2], Paul Flanagan[2], Oisín Redmond[2], Anya Belz[2], Yufang Hou[1,3]
[1]IBM Research Europe - Ireland, [2]Dublin City University,
[3]IT:U Interdisciplinary Transformation University Austria
Correspondence: massimiliano.pronesti@ibm.com, yufang.hou@it-u.at

Abstract

Systematic reviews in medicine play a critical role in evidence-based decision-making by aggregating findings from multiple studies. A central bottleneck in automating this process is extracting numeric evidence and determining study-level conclusions for specific outcomes and comparisons. Prior work has framed this problem as a textual inference task by retrieving relevant content fragments and inferring conclusions from them. However, such approaches often rely on shallow textual cues and fail to capture the underlying numeric reasoning behind expert assessments. In this work, we conceptualise the problem as one of quantitative reasoning. Rather than inferring conclusions from surface text, we extract structured numerical evidence (e.g., *event counts* or *standard deviations*) and apply domain knowledge informed logic to derive outcome-specific conclusions. We develop a numeric reasoning system composed of a numeric data extraction model and an effect estimate component, enabling more accurate and interpretable inference aligned with the domain expert principles. We train the numeric data extraction model using different strategies, including supervised fine-tuning (SFT), and reinforcement learning (RL) with a new value reward model. When evaluated on the COCHRANEFOREST benchmark, our best-performing approach – using RL to train a small-scale number extraction model – yields up to a 21% absolute improvement in F1 score over retrieval-based systems and outperforms general-purpose LLMs of over 400B parameters by up to 9%. Our results demonstrate the promise of reasoning-driven approaches for automating systematic evidence synthesis.

1 Introduction

Systematic reviews are the cornerstone of evidence-based medicine, offering rigorous syntheses of available studies to guide clinical decision-making (Murad et al., 2016). A critical component

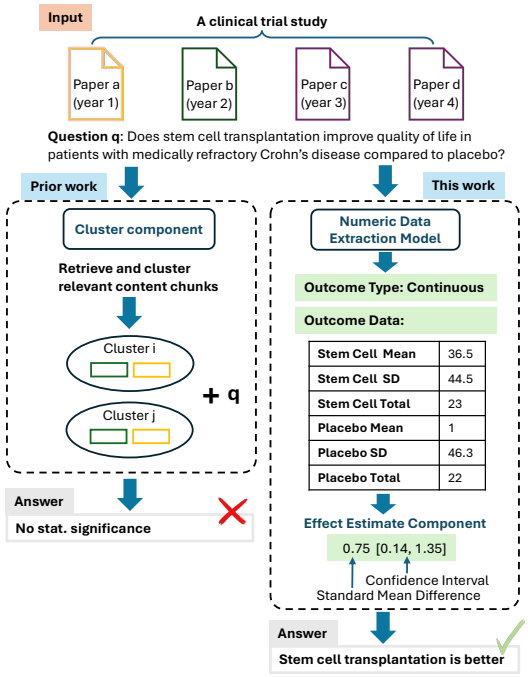

Figure 1: Example of estimating the intervention effect based on the extracted outcome data for a clinical study.

of systematic reviews is the extraction of study-level numeric evidence (e.g., *event counts* or *standard deviations*) from one or multiple corresponding clinical trial papers and derive the conclusions for each outcome and comparison under assessment. However, automating this extraction remains an open challenge. As illustrated in Figure 1 (prior work), previous studies (Pronesti et al., 2025) have primarily framed this task as a retrieval-based question answering problem: given a query about an outcome, systems retrieve relevant study fragments and infer conclusions based on the retrieved text.

Despite progress, such approaches fundamentally rely on surface-level textual cues, limiting their effectiveness. To illustrate these limitations, we plot in Figure 2 the relationship between evidence retrieval precision (x-axis) and the answer F1 score (y-axis) on 30 manually annotated instances

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30346–30362
November 4-9, 2025

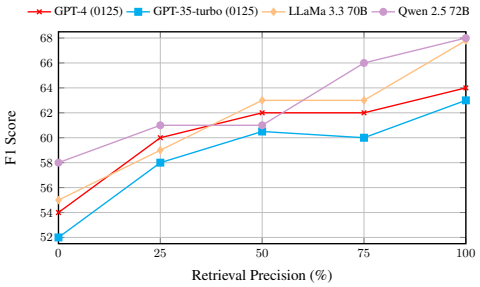

Figure 2: F1 score for predicting the correct answers on 30 instances from COCHRANEFOREST on different retrieval precision using 4 state-of-the-art LLMs.

(Appendix C) from the COCHRANEFOREST benchmark (Pronesti et al., 2025) across four state-of-the-art large language models (LLMs). Each instance consists of a trial study comprising one or more papers, a research question, its corresponding categorical answer, and an exhaustive annotation of all supporting textual evidence drawn from the entire paper. In general, while better retrieval precision correlates with improved performance, even perfect retrieval (100% precision) results in a modest maximum F1 score of 68%. Moreover, performance gains plateau quickly: increasing retrieval precision from 50% to 100% yields only 3-4% absolute improvement. This suggests that textual information alone is often insufficient to determine study conclusions, particularly when studies address multiple outcomes or consist of multiple publications.

These findings motivate a shift in perspective – from relying on surface-level textual cues to explicitly modelling the quantitative reasoning that underpins expert assessments in systematic reviews. A natural alternative is to adopt domain expert principles by extracting and interpreting the numerical evidence (e.g., *effect sizes* and *confidence intervals*) that supports each study's conclusion. This reframes the task from semantic retrieval to structured statistical inference. Recent work by Yun et al., 2024 has explored this direction by leveraging pretrained LLMs through prompting to extract quantitative results. While promising, this approach has so far been limited to individual trials and has not addressed the challenges posed by longer, heterogeneous full-text studies included in systematic reviews. Furthermore, the potential of custom-trained models optimised specifically for numerical reasoning and alignment with expert conclusions remains largely unexplored in this setting.

Concurrently, advances in supervised fine-tuning (SFT) and reinforcement learning (RL) have shown significant improvements in aligning model behaviour with complex reasoning objectives (Wei et al., 2022; Guo et al., 2025). Building on these insights, as shown in Figure 1 (this work), we propose a structured pipeline that extracts interpretable numerical evidence from full-text studies via a numeric data extraction model, and infers study-level conclusions using transparent rules through an effect estimate component—eschewing reliance on implicit textual signals. We train compact numeric data extraction models using a wide range of strategies, including SFT, SFT with intermediate reasoning traces, and RL with a novel value-based reward model. On two different datasets, our models outperform the prompting-based approaches based on big models proposed in Yun et al. (2024).

In addition, compared to the implicit text evidence reasoning approach (Pronesti et al., 2025), our models allow one to automatically generate the corresponding row of a forest plot (Section 2.2) from a full text study by directly extracting numerical evidence for each outcome measure. This represents a key step toward full automation of the systematic review process, bridging the gap between primary study reporting and meta-analytic synthesis (Wallace et al., 2010; Tsafnat et al., 2014).

In summary, our contributions are as follows: (1) we propose a novel pipeline that predicts study conclusions by extracting and statistically analysing numerical outcomes, instead of relying purely on textual retrieval; (2) we explore how different training strategies—standard SFT, SFT with intermediate reasoning traces, and RL—affect the reasoning abilities of compact language models on this task; (3) we develop custom models trained for this task, achieving up to 21% absolute improvement in F1 score compared to retrieval-based systems on COCHRANEFOREST.

2 Preliminaries

2.1 Systematic Reviews

A systematic review is a rigorous method of synthesising evidence from multiple studies that address a clearly formulated research question (Chandler et al., 2019). It follows a structured protocol for identifying, selecting, and appraising relevant research to minimise bias and yield reliable findings.

2.2 Forest Plots

Forest plots are visual tools commonly used in systematic reviews to display the estimated effects

from multiple studies on a common scale. Each study is typically represented as a point estimate (e.g., *mean difference*, *odds ratio*) with a confidence interval, and a vertical line indicating a null effect (e.g., 0 for differences, 1 for ratios). Forest plots help in assessing consistency across studies and interpreting the overall effect direction and magnitude. Examples are shown in Appendix A.

2.3 Data Extraction from Biomedical Studies

Biomedical studies, particularly randomised controlled trials (RCTs), typically report numerical outcome data for different treatments or interventions. These results—such as event or group size—may appear in tables, figures, or embedded within narrative text. Extracting this information is essential for downstream tasks such as study-level evidence synthesis and across-study meta-analysis.

Following Yun et al. (2024), we categorise extracted numerical data into two main outcome types: *binary outcomes*, which include the number of events and group sizes for both the intervention and comparator arms; and *continuous outcomes*, which consist of means, standard deviations, and group sizes for intervention and comparator groups.

Once extracted, these values can be used in standard meta-analytic methods to estimate treatment effects—such as mean differences, risk ratios, or odds ratios—along with their corresponding confidence intervals, forming the basis for deriving study-level conclusions. An example of estimating treatment effects is provided in Appendix A.

3 Methodology

As shown in Figure 1, our system consists of a fine-tuned numeric extraction model grounded in reasoning, alongside a rule-based effect size estimation component. In the following, we describe each part in detail.

3.1 Training Dataset Creation

Training Data Collection. The initial phase of our methodology involved the acquisition of high-quality, human-annotated data to train our models. Following the methodology described in Pronesti et al. (2025), we processed the Cochrane Database of Systematic Reviews (CDSR)[1] to identify systematic reviews containing non-paywalled full-text studies and at least one forest plot available in SVG format. Each forest plot was parsed to extract the underlying numerical data (i.e. the number of events and total participants in each group for binary outcomes, or the mean, standard deviation, and sample size for continuous outcomes), along with the point estimate, 95% confidence interval, and textual conclusion for each included study.

[1] https://www.cochranelibrary.com/cdsr/reviews

Dataset	Train	Test	Total	Avg tokens
CochraneForestExt	1864	208	2072	12109.2
CochraneForest	–	725	725	11688.7
RCTs	–	413	413	4364.9

Table 1: Datasets statistics. Train/test split only applies to CochraneForestExt. CochraneForest and RCTs are used for testing.

To increase dataset size and training diversity, we relaxed one of the constraints imposed in the original CochraneForest benchmark. Specifically, we no longer require that each forest plot contain at least two studies with differing conclusions. This change allows us to include more reviews while still preserving outcome-level heterogeneity across the dataset. In addition, to prevent data leakage, we explicitly excluded any study in CochraneForest from our training set.

The final training dataset consists of 2,072 examples, spanning 104 systematic reviews and 25.9 M tokens of full-text biomedical content. Each data point contains: (1) the full text of a study, (2) the outcome type (i.e., *binary* or *continuous*) as defined in the corresponding forest plot, (3) the set of numerical values to be extracted (e.g., group means or event counts), (4) the computed point estimate, (5) the 95% CI, and (6) the final conclusion assigned to the outcome in the forest plot. We refer to this dataset as CochraneForestExt.

Synthetic Data Annotation. To enrich the dataset with reasoning traces for SFT, we used Llama 3.1 405B (Grattafiori et al., 2024) with the system prompt shown in Figure 6 (Appendix), temperature of 0.7 and 2,048 tokens generation limit. An example data instance is provided in Table 6 (Appendix).

3.2 Numeric Data Extraction Model

3.2.1 SFT with CoT

We first adapt a pretrained language model to our task via supervised fine-tuning (SFT), which adapts a pretrained language model π_θ to reflect a domain-specific distribution $\mathcal{P}$. This is achieved by minimising the negative log-likelihood over a dataset

of example sequences, encouraging the model to increase the probability of desired outputs. In our context, the goal is to enable the model to map free-text descriptions of study results to structured outputs that capture outcome information in a standardised schema. Each training example consists of a biomedical study, paired with the corresponding outcome of interest (Table 6). The output is represented in YAML format and encodes either binary outcomes (with event counts and totals for intervention and comparator groups) or continuous outcomes (with group-wise means, standard deviations, and sample sizes). The model learns to generate these structured summaries conditioned on the corresponding textual evidence. That is, given prompt-target pairs $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$, the loss is computed as:

$$\mathcal{L}_{\text{cond}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}} \left[\sum_{t=1}^{m} \log \pi_\theta(y_t \mid \mathbf{x}, y_{<t}) \right]$$

This focus helps the model better learn task-relevant outputs without overfitting to input tokens (Wang et al., 2023; Chiang et al., 2023; Yu et al., 2024).

3.2.2 RL with Fine-grained Rewards

As an alternative to SFT, we explore Reinforcement Learning, which further refines LLMs by aligning model outputs with human preferences or reward signals. We adopt Group-Relative Policy Optimisation (GRPO) (Shao et al., 2024), which computes normalised rewards over a group of responses and reduces variance in learning.

In our setup, the model acts as a policy π_θ that takes as input a textual passage describing clinical study results and outputs a structured response. Each response consists of a thought process enclosed in a `<think>` tag, followed by a YAML object encoding outcome data, using the same schema adopted in SFT. For each passage $\mathbf{x}$, G candidate completions $\{y_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot \mid \mathbf{x})$ are sampled from the reference policy π_{old} to encourage robustness and diversity. These completions are scored using rule-based reward functions that evaluate factual correctness and adherence to format. The raw rewards R_i are then normalised across the group:

$$A_i = \frac{R_i - \mathbb{E}[R_j]}{\sqrt{\mathbb{V}[R_j]}}, \quad j \in \{i, ..., G\}$$

where $\mathbb{E}[R_j]$ and $\mathbb{V}[R_j]$ are respectively the mean and variance of the rewards for the group of responses. The policy is optimised using a clipped, KL-regularised objective that encourages agreement with high-reward behaviours while maintaining proximity to a reference model π_{ref}:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \Big(p_{i,t}(\theta) A_i, \text{clip}(p_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon) A_i \Big) - \beta\, \text{KL}[\pi_\theta \,\|\, \pi_{\text{ref}}] \Bigg]$$

where β governs the regularisation strength and $p_{i,t}(\theta)$ is the token-level probability ratio defined as follows:

$$p_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}$$

Reward Functions. In our setting, the model produces structured YAML outputs representing binary or continuous outcomes. To evaluate these outputs during RL, we define three rule-based reward functions based on format validity and numerical correctness. Let C_i be the model output, E_i the expected answer.

Correctness Reward (CR). This reward compares numerical values in C_i and E_i when the outcome type is correct. Let $V_i = \{v_1, \ldots, v_n\}$ and $\hat{V}_i = \{\hat{v}_1, \ldots, \hat{v}_n\}$ be the parsed numerical fields. Then:

$$R_{\text{CR}} = \frac{1 + \sum_{j=1}^{n} \mathbb{1}\{v_j \approx \hat{v}_j\}}{1+n}$$

where $v_j \approx \hat{v}_j$ means exact match for integers and absolute difference $< 10^{-3}$ for floats. If parsing fails or types mismatch, we set $R_{\text{CR}} = 0$.

Format Reward (FR). This reward checks whether C_i follows the expected structure. Let $\mathcal{F}$ denote the set of all valid formats (Appendix B), including required keys and outcome type. We define:

$$R_{\text{FR}} = \begin{cases} 1 & \text{if } \pi_\theta(x) \in \mathcal{F} \\ 0 & \text{otherwise} \end{cases}$$

This reward ensures that the output adheres to a valid YAML schema for the predicted outcome type.

Thought Format Reward (TFR). The TFR incentivises the model to adhere to a predefined output structure, such as the use of `<think>` tag.

$$R_{\text{TFR}} = \begin{cases} 1 & \text{if } \pi_\theta(x) \text{ matches thought pattern} \\ 0 & \text{otherwise} \end{cases}$$

Final Reward. The final reward is a weighted combination of format and correctness components:

$$R = 0.8 \cdot R_{\text{CR}} + 0.1 \cdot R_{\text{FR}} + 0.1 \cdot R_{\text{TFR}}$$

The weights prioritise factual accuracy while still incentivising structural correctness and robustness to formatting issues with proper reasoning traces. More details are provided in Appendix E.

3.3 Effect Estimate Component

After extracting the relevant numerical data, we compute standardised fixed-effect estimates based on the type of outcome (Hedges and Vevea, 1998).

Binary Outcomes. For event-based outcomes, we compute the risk ratio (RR) as $\text{RR} = \frac{a/(a+b)}{c/(c+d)}$, where a, b are the numbers of events and non-events in the treatment group, and c, d in the control group. To quantify uncertainty, we compute the 95% confidence interval on the log scale as $\log(\text{RR}) \pm 1.96 \cdot \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$, which is then exponentiated to return to the RR scale (see Appendix A for an example calculation).

Continuous Outcomes. For outcomes measured on a continuous scale, we compute the mean difference (MD) as $\bar{x}_T - \bar{x}_C$, where $\bar{x}_T$ and $\bar{x}_C$ are the group means in the treatment and control arms, respectively. The 95% CI is then computed as $\text{MD} \pm 1.96 \cdot \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}$, where s_T, s_C and n_T, n_C are respectively the standard deviations and samples sizes of the two groups (see Appendix A for an example calculation).

Deriving Study Conclusions. Study conclusions are determined directly from the 95% confidence interval of the effect estimate (Chang et al., 2022). For binary outcomes, if the confidence interval for the odds ratio lies entirely above 1, the study is classified as supporting the intervention; if it lies entirely below 1, it favors the control; and if it includes 1, the result is considered inconclusive. For continuous outcomes, the same logic applies with respect to the null value 0.

4 Experiments

4.1 Experimental setup

Training and Evaluation Datasets. For training, we use the dataset created as described in Section 3.1, with the reasoning traces for SFT, using 1864 samples for training and 208 for validation.

For evaluation, we use two datasets. The first is CochraneForest (Pronesti et al., 2025), which consists of 725 instances derived from 48 Cochrane Systematic Reviews and 220 forest plots. The second, introduced by Yun et al. (2024), contains 413 complete, human-annotated instances derived from 120 RCTs. Dataset statistics are reported in Table 1.

Evaluation Metrics. We report a range of metrics to assess both end-to-end performance and intermediate reasoning accuracy. For the end-to-end task of predicting the final conclusion label of each study, we report accuracy and F1 score. Following Yun et al. (2024), we also evaluate the quality of numerical extraction using exact match (EM), EM@1 (at least one positional match), as well as the mean squared error (MSE) of the computed point estimate derived from the extracted values. In addition, we measure the Error Impact Rate (EIR), defined as the ratio between the number of extraction errors that lead to a flipped conclusion and the number of total extraction errors at the study level.

$$\text{EIR} = \frac{\#\text{errors that flip the conclusion}}{\#\text{extraction errors}}$$

This metric provides an insight into how often extraction mistakes materially affect downstream decision-making. A high EIR indicates that even a few extraction errors can substantially alter the final prediction, whereas a low EIR suggests that the model's conclusions are more robust to imperfect inputs. Thus, EIR serves as a proxy for understanding the correlation between intermediate numerical accuracy and end-to-end reliability.

Training Setup. We conduct our training using the Qwen2.5 model family (Yang et al., 2025), specifically the 7B variant. Two distinct training regimes are explored: SFT and RL. In the results section, models are labelled with subscripts corresponding to the respective training strategy.

SFT is performed for 5 epochs with a batch size of 1 using a learning rate of 5×10^{-5} and the AdamW optimiser (Loshchilov and Hutter, 2017). For the RL setup, we adopt the GRPO algorithm (Shao et al., 2024), training for 3 epochs with a learning rate of 1×10^{-6}, batch size 1, and 16 sampled generations per batch. Additional details are provided in Appendix E.

Model Baselines. To validate our results, we compare a range of open- and closed-source models, with and without reasoning capabilities. All

Model	Think	CochraneForest						RCTs (Yun et al., 2024)					
		Acc	F1	EM	EM@1	MSE↓	EIR↓	Acc	F1	EM	EM@1	MSE↓	EIR↓
Pretrained LLMs													
GPT-4-0125	✗	71.4	73.7	28.4	72.1	0.88	0.31	71.2	65.3	34.5	70.3	0.69	0.44
Qwen2.5-7B	✗	69.7	65.4	14.9	66.4	1.40	0.39	63.2	59.7	20.3	73.8	1.19	0.50
Qwen2.5-14B	✗	76.2	73.7	27.4	73.6	0.72	0.30	69.8	66.4	28.6	75.8	0.91	0.44
Qwen2.5-72B	✗	82.2	80.3	42.5	81.7	0.63	0.26	**75.3**	**72.8**	**35.1**	**81.1**	**0.68**	**0.39**
Llama-3.1-8B	✗	68.7	66.6	15.9	62.2	2.16	0.33	62.7	54.7	14.8	66.5	1.30	0.47
Llama-3.1-70B	✗	79.4	78.3	31.5	78.7	0.59	0.28	65.5	60.8	30.0	74.2	1.45	0.44
Llama-3.1-405B	✗	**82.4**	**80.8**	**44.3**	**82.4**	**0.43**	**0.23**	70.3	66.5	33.6	76.7	0.80	0.41
DeepSeek-Qwen-7B	✓	59.7	54.3	11.2	58.3	2.83	0.50	53.8	45.7	9.4	53.8	3.68	0.61
DeepSeek-Qwen-14B	✓	65.4	61.2	19.9	66.4	1.29	0.44	60.2	55.1	11.2	60.3	2.41	0.58
DeepSeek-Qwen-32B	✓	74.0	71.6	28.6	73.5	0.58	0.28	68.8	65.0	29.1	78.2	0.72	0.43
Our Models													
Qwen2.5-7B-SFT	✓	74.5	70.1	28.4	79.1	0.51	0.29	71.5	68.4	30.2	80.0	0.64	0.36
Qwen2.5-7B-RL	✓	**81.6**	**80.1**	**42.2**	**81.4**	**0.40**	**0.24**	**79.3**	**76.4**	**41.2**	**89.7**	**0.49**	**0.28**

Table 2: Evaluation results across models on two datasets. We report Accuracy and F1 (label prediction), EM, EM@1, EIR and MSE (numerical extraction).

models are evaluated in zero-shot settings with prompt and hyperparameters shown in Appendix B.

We include two main model families: Qwen 2.5 (Yang et al., 2025) and Llama 3.1 (Grattafiori et al., 2024). In addition, we benchmark DeepSeek-R1 and the distilled Qwen models derived from it. We exclude distilled models afferent to the Llama family because of their limited context size. For closed-source models, we use GPT-4-0125.

In addition, we compare performances on the end-to-end task against the two best RAG baselines for this task: URCA (Pronesti et al., 2025), which clusters retrieved passages based on their embedding vectors and filters relevant information from each cluster given the query; and GraphRAG (Edge et al., 2024), which builds a graph-based text index by summarising closely related entities from the source documents.

4.2 Main Results

Comparison with Pretrained Baselines. Table 2 presents a performance comparison of pretrained and fine-tuned language models on the CochraneForest and RCTs datasets. Models vary in size, architecture, and training approach, allowing us to examine the relationship between model scale, reasoning capability, and task-specific performance. Among the pretrained baselines, Llama-3.1-405B achieves the best performance on CochraneForest, closely followed by Qwen2.5-72B, which performs competitively on both datasets, topping the RCTs benchmark on several metrics. Llama-3.1-70B and Qwen2.5-14B also rank highly, while Qwen2.5-7B and Llama-3.1-8B underperform relative to their larger variants.

Fine-tuning improves performance significantly. Qwen2.5-7B-SFT outperforms all pretrained models of similar or larger size. The Qwen2.5-7B-RL model establishes a new state-of-the-art on both datasets, surpassing all baselines—including the 405B model—in nearly every metric on RCTs and closely matching it on CochraneForest, despite being nearly 58x smaller. In addition, when compared directly to its pretrained counterpart, Qwen2.5-7B-RL shows large gains in EM (+20.9), accuracy (+16.1), and F1 (+16.7) on the RCTs dataset, highlighting the effectiveness of RL.

Notably, we observe that reasoning capabilities do not always lead to improved performance. In fact, all the distilled DeepSeek models with reasoning perform worse than their non-reasoning counterparts. These results suggest that general reasoning ability alone is insufficient for complex domain-specific tasks. Instead, explicit task supervision and structured reasoning training appear necessary to guide models in applying reasoning capabilities effectively.

Comparison with RAG Baselines. Table 3 compares the best-performing numbers-based models with two strong retrieval-augmented generation (RAG) baselines —URCA and GraphRAG—on the CochraneForest dataset. Both RAG methods underperform compared to direct numerical rea-

soning. The fine-tuned `Qwen2.5-7B-RL` achieves an absolute gain of over 20 F1 points compared to the best RAG baseline, highlighting the limitations of retrieval in settings where precise numerical grounding and structured inference are required. Even the zero-shot numbers-based approach outperforms the RAG methods, suggesting that factual retrieval alone is insufficient without robust reasoning over structured content.

Method	F1	Acc
RAG baselines		
URCA (Pronesti et al., 2025)	60.2	58.8
GraphRAG (Edge et al., 2024)	59.6	57.5
Numerical baselines		
Qwen2.5-7B (zero-shot)	69.7	65.4
Qwen2.5-7B-SFT	74.5	70.1
Qwen2.5-7B-RL	**81.6**	**80.1**

Table 3: Performance comparison between RAG-based and numbers-based approaches on COCHRANEFOREST using `Qwen2.5-7B-Instruct`.

4.3 Ablation Studies

To assess the impact of different input modalities and training strategies, we conduct a series of ablation experiments (Table 4) on the RL- and SFT-trained models. These include both data ablations and training ablations.

In the data ablations, we isolate the contribution of different components of the input. We evaluate model performance when provided with: (i) only textual context, where all tables are removed; (ii) only tables, where we exclude surrounding text and provide the model with structured numerical content; and (iii) only the top retrieved chunks obtained with URCA (Pronesti et al., 2025), restricting access to 10 evidence passages per query.

In the training ablations, we analyse the effect of intermediate reasoning and reward design. To evaluate the role of reasoning traces, we compare the SFT model to a version trained without Chain-of-Thought (CoT), which predicts the final answer directly without intermediate steps. To assess the impact of the reward function, we compare the RL model to another version trained using the `EX` reward, which only provides a positive signal when all predicted values are correct.

$$R_{\text{EX}} = \mathbb{1}\left\{\bigwedge_{j=1}^{n}(v_j \approx \hat{v}_j)\right\}$$

We observe that among the data ablations, the models trained only on tables achieve the best performance, confirming the importance of structured numerical evidence in this task. However, there is a consistent drop in performance across all settings when models only have access to a single input type compared to the full input (text + tables), indicating that both textual and tabular information are necessary for accurate reasoning. The `URCA` setting, which corresponds to the output of a retrieval-based pipeline, yields the lowest scores. This aligns with the findings in Table 3 and further highlights the limitations of existing retrieval-augmented generation approaches in this setting, primarily due to their inability to recover the precise numerical content required for grounded inference.

Model	Input	F1	Acc
Input Data Ablations			
`Qwen2.5-7B-SFT`	`text`	62.3	60.1
	`tables`	70.4	66.5
	`urca`	56.7	54.0
`Qwen2.5-7B-RL`	`text`	65.8	63.2
	`tables`	**73.1**	**71.6**
	`urca`	59.2	56.8
Training Ablations			
`Qwen2.5-7B`	`text + tables`	69.7	65.4
`+ SFT-no-CoT`	`text + tables`	71.2	68.3
`+ SFT` **(ours)**	`text + tables`	74.5	70.1
`+ RL-EX`	`text + tables`	79.7	78.8
`+ RL` **(ours)**	`text + tables`	**81.6**	**80.1**

Table 4: Ablation study on model inputs and training supervision on COCHRANEFOREST. `urca` refers to the top 10 chunks retrieved by its namesake RAG approach.

In the training ablations, removing CoT supervision during SFT (`SFT-no-CoT`) results in lower performance compared to full SFT, showing that intermediate reasoning traces help guide the model's learning process. When using RL, we find that the dense reward signal (R_{CR}) significantly outperforms the variant trained with the EX reward (R_{EX}). This confirms that sparse rewards are less effective at shaping model behaviour than denser, fine-grained signals.

4.4 Analyses

Impact of the Thought Process. We want to evaluate whether the reasoning processes generated by our trained models (RL and SFT) are logically sound, provide meaningful explanations for the extracted numbers, and enable traceability back to the input papers. To this end, we manually annotate the outputs of both models on 30 biomedical studies, for which human annotators had previously identified the sources of the correct numbers (see

Thought Process	Reasoning Label	Qwen2.5-7B-RL		Qwen2.5-7B-SFT	
		Count	Fraction	Count	Fraction
✓	Correct reasoning with traceability	42	31.82%	34	25.76%
✓	Correct reasoning without traceability	39	29.55%	24	18.18%
✓	Correct reasoning, incorrect number	37	28.03%	41	31.06%
✗	Correct number, incorrect reasoning	3	2.72%	13	9.85%
✗	Copy without reasoning	6	4.55%	0	0.00%
✗	Hallucinated	5	3.79%	20	15.15%
✗	Missing reasoning	0	0.00%	0	0.00%
	Total	132	100%	132	100%

Table 5: Distribution of annotated reasoning labels for Qwen2.5-7B-RL and and Qwen2.5-7B-SFT. ✓ indicates correct reasoning. ✗ indicates incorrect reasoning.

Appendix C), assigning a single label to each reasoning process. Labels distinguish whether the reasoning correctly supports the individual number, whether it enables traceability to the source, and whether it hallucinates or lacks explanation. A complete list of labels and annotation criteria is provided in Appendix D.

Results (Table 5) show that the RL model produces a higher fraction of correct reasoning processes overall compared to the SFT model (61.37% vs 43.94%, summing the first two rows). This suggests that reinforcement learning improves the model's ability to provide plausible and well-structured explanations for numerical outputs. Notably, the RL model achieves a substantially lower rate of hallucinated reasoning (3.79% vs 15.15%), highlighting a significant improvement in factual grounding. It also shows a reduced frequency of incorrect reasoning behind correct numbers (2.72% vs 9.85%), indicating better alignment between the generated explanations and the numerical outputs.

These findings suggest that reinforcement learning not only improves the factual accuracy of model outputs, but also enhances the traceability and reliability of the underlying reasoning process.

Reward Dynamics. Figure 3 shows the reward dynamics during RL training. Thought Format Reward (TFR) and Format Reward (FR) increase rapidly and plateau early, indicating that the model quickly learns the reasoning scaffold and YAML schema. This behaviour also suggests that the model inherently learns to predict the correct outcome type at an early stage. By contrast, Correctness Reward (CR) improves more gradually and continues rising after TFR/FR have stabilised, showing that numerical correctness requires longer training. These dynamics are consistent with our main results and design choices: structural aspects

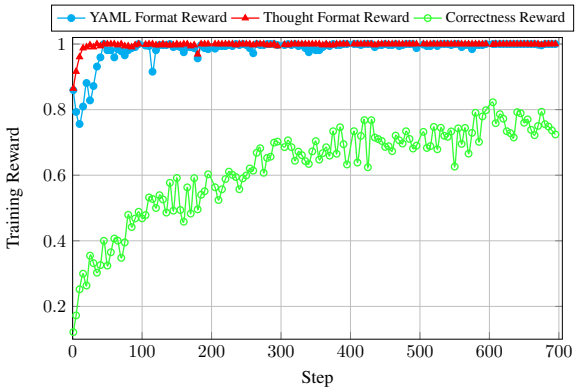

Figure 3: Rewards dynamics. Format rewards plateau quickly during training, whereas the correctness reward improves more gradually, suggesting that LLMs naturally learn to follow structural patterns while still improving on quality rewards.

are mastered quickly, whereas correctness is more demanding and depends on the model's ability to apply reasoning across each component of the final output. Overall, the reward dynamics illustrate how reinforcement signals guide the model from surface-level alignment toward substantive reasoning ability.

Qualitative Example. We present a qualitative example from CochraneForest (Figure 9, in appendix), comparing the reasoning and outputs of Qwen2.5-7B-SFT and Qwen2.5-7B-RL on a challenging case where the exact values for the comparator and outcome are not explicitly reported in the biomedical study. While the SFT model fails to extract the relevant information, the RL model demonstrates stronger reasoning capabilities: it identifies population percentages from one of the study's tables and correctly maps them to the total number of participants previously extracted, thereby analytically inferring the desired values. Notably, even our human annotators initially strug-

gled to locate the correct numbers in this scenario.

5 Related work

Biomedical Information Extraction. Recent efforts in biomedical information extraction have made significant advances. Wadhwa et al. (2023) introduced a generative approach for extracting interventions, comparing and outcomes from RCTs. While effective, we extend this by focusing on numerical evidence and reasoning.

Lehman et al. (2019) demonstrated that accurately extracting evidence is the primary bottleneck for determining treatment efficacy. O'Doherty et al. (2024) showed using abstracts alone rather than structured PICO data can worsen synthesis quality, underscoring the importance of data representation. Our model processes the full content of the paper, providing a more comprehensive basis. Our work can be seen as an example of scientific argument mining at the global discourse level (Al Khatib et al., 2021) and contributes to the broader agenda of AI for Science (Eger et al., 2025).

LLMs for Evidence Synthesis and Numerical Reasoning. While LLMs have shown promise in biomedical text mining, they face challenges synthesising complex evidence and numerical reasoning. Shaib et al. (2023) found that GPT-3 struggled with multi-document synthesis and biased effect reporting. Nye et al. (2020) introduced a framework to map evidence and infer conclusions. Yun et al. (2024) assessed how effectively LLMs can extract numerical data, noting challenges with continuous outcomes and distinguishing similar measures. Wang et al. (2025) introduced a method that integrates LLMs with code generation to extract clinical study outcomes. However, their evaluation was restricted to selected cancer-related reviews, and the reliance on structured prompts and domain-specific heuristics raises questions about scalability to broader therapeutic areas. Lai et al. (2025) evaluated LLMs for data extraction and risk of bias assessment in complementary medicine, reporting high performance, but their focus on a relatively narrow and domain-specific set of trials limits the generalizability of the findings. Similarly, Sun et al. (2024) assessed LLMs for automated data extraction from randomized trials, finding encouraging results, yet performance was uneven across outcome types, with continuous measures proving particularly challenging. These studies highlight that while LLMs show high potential, existing work focuses on narrow domains, curated benchmarks, or simplified tasks, leaving open the question of how well they generalise to the diverse evidence encountered in real-world systematic reviews.

Reinforcement Learning for Numerical Evidence Extraction. Previous work has proven that SFT can be used to improve models such as BERT (Devlin et al., 2019) for biomedical text mining (Lee et al., 2019; Xie et al., 2022). Reinforcement learning has improved model alignment for complex reasoning objectives (Wei et al., 2022; Lambert et al., 2024; Guo et al., 2025). Recent work (Lin et al., 2025) demonstrates that RL with verifiable rewards can improve model performance on clinical reasoning tasks like EHR-based calculations and trial matching. To the best of our knowledge, we are the first to design and apply RL-based approaches specifically to the extraction of numerical evidence from biomedical studies for use in systematic reviews.

6 Conclusion

In this paper we presented a quantitative reasoning framework for automating evidence extraction in systematic reviews, shifting away from shallow textual inference toward structured numeric understanding. By directly modelling the process domain experts follow, we move closer to automating a key step in evidence-based medicine. Our proposed system, combining SFT and RL numeric extraction models with reasoning over effect estimates, significantly outperforms retrieval-based baselines and even large-scale language models on both the CochraneForest and the RCTs benchmarks. These results affirm the value of reasoning-driven supervision for complex scientific tasks and point to RL as a promising avenue for developing reliable and domain-aligned extraction systems in evidence-based medicine.

Beyond improving performance, our approach highlights the importance of aligning machine reasoning with expert workflows, ensuring interpretability and trustworthiness. Looking forward, this work opens opportunities for integrating automated evidence extraction into clinical decision support pipelines, ultimately reducing the manual burden on researchers and accelerating the translation of medical knowledge into practice. Future research may extend this framework to other scientific domains where quantitative reasoning is core.

7 Limitations

While this work makes meaningful contributions toward reasoning-based evidence extraction, several limitations remain. First, the system assumes that all relevant numerical evidence is explicitly or implicitly reported and cleanly extractable, which is often not the case in real-world clinical studies. Missing values and inconsistent formats can hinder reliable extraction. At present, the model does not attempt to identify or flag missing information, which limits its utility in incomplete or noisy settings. Training models to detect and explicitly report missing or uncertain values is a critical direction for future development. Additionally, our approach currently focuses on a limited set of outcome types and uses predefined logical rules, which may constrain generalisation to more complex or diverse clinical scenarios. Addressing these limitations will be key to deploying robust systems for real-world systematic review automation.

References

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. 2019. *Hoboken: Wiley*, 4.

Yaping Chang, Mark R Phillips, Robyn H Guymer, Lehana Thabane, Mohit Bhandari, Varun Chaudhary, and RETINA study group Wykoff Charles C. 7 8 Sivaprasad Sobha 9 Kaiser Peter 10 Sarraf David 11 Bakri Sophie 12 Garg Sunir J. 13 Singh Rishi P. 14 15 Holz Frank G. 16 Wong Tien Y. 17 18. 2022. The 5 min meta-analysis: understanding how to read and interpret a forest plot. *Eye*, 36(4):673–675.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatGPT quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *Preprint*, arXiv:2502.05151.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The LLaMa 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Larry V Hedges and Jack L Vevea. 1998. Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486.

Hugging Face. 2025. Open R1: A fully open reproduction of deepseek-r1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Honghao Lai, Jiayi Liu, Chunyang Bai, Hui Liu, Bei Pan, Xufei Luo, Liangying Hou, Weilong Zhao, Danni Xia, Jinhui Tian, and 1 others. 2025. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digital Medicine*, 8(1):74.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. 2025. Training LLMs for EHR-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. 2016. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.

Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace. 2020. Trialstreamer: Mapping and browsing medical evidence in real-time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69, Online. Association for Computational Linguistics.

James O'Doherty, Cian Nolan, Yufang Hou, and Anya Belz. 2024. Beyond abstracts: A new dataset, prompt design strategy and method for biomedical synthesis generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 358–377, Bangkok, Thailand. Association for Computational Linguistics.

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. 2025. Query-driven document-level scientific evidence extraction from biomedical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28034–28051, Vienna, Austria. Association for Computational Linguistics.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhuanlan Sun, Ruilin Zhang, Suhail A Doi, Luis Furuya-Kanamori, Tianqi Yu, Lifeng Lin, and Chang Xu. 2024. How good are large language models for automated data extraction from randomized trials? *MedRxiv*, pages 2024–02.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic reviews*, 3:1–15.

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C. Wallace. 2023. Jointly extracting interventions, outcomes, and findings from rct reports with LLMs. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 754–771. PMLR.

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11:1–11.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Xiao Yu, Qingyang Wu, Yu Li, and Zhou Yu. 2024. LIONs: An empirically optimized approach to align language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8732–8753, Miami, Florida, USA. Association for Computational Linguistics.

Hye Sun Yun, David Pogrebitskiy, Iain J Marshall, and Byron C Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. In *Machine Learning for Healthcare Conference*. PMLR.

A Forest Plots and Effect Size Estimation

Forest plots are a standard way to visualise the results of individual studies and their synthesis in a meta-analysis. Each row typically represents one study, showing its estimated treatment effect and the corresponding 95% confidence interval (CI). The point estimate is usually plotted as a square (with size proportional to the study's weight), and the CI as a horizontal line. The vertical line represents the line of no effect: 1 for ratios (e.g., risk ratio or odds ratio), and 0 for differences (e.g., mean difference).

At the bottom of the plot, a diamond often represents the pooled effect estimate from all included studies. This visualisation helps readers assess not only the overall direction and magnitude of the effect but also the consistency (heterogeneity) across studies.

Below we illustrate how to interpret and compute effect estimates and confidence intervals for the two common outcome types: binary outcomes with risk ratios (Figure 4), and continuous outcomes with mean differences (Figure 5).

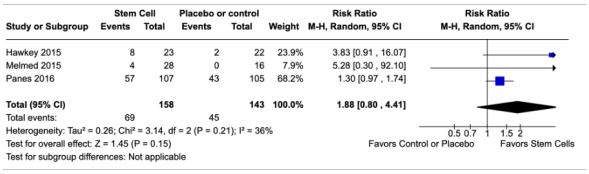

Figure 4: A forest plot assessing clinical remission (binary outcome) in patients affected by medically refractory Crohn's disease treated with stem cells versus control.

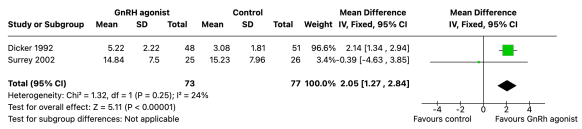

Figure 5: A forest plot comparing GnRH agonist versus no agonist (placebo) in women with endometriosis, assessing a continuous outcome (average number of oocytes per woman).

Binary Outcomes and Risk Ratios. Binary outcomes refer to variables with two possible states, such as event/no event. A commonly used effect measure in this context is the *risk ratio* (RR), which compares the proportion of events in the intervention group to that in the comparator group:

$$\text{RR} = \frac{a/n_1}{c/n_2}$$

where a and n_1 are the number of events and total participants in the intervention group, and c and n_2 are those in the comparator group.

Example. In the **Hawkey 2015** study (Figure 4):

- Intervention (stem cells): 8 events out of 23 participants
- Comparator (placebo): 2 events out of 22 participants

The point estimate is:

$$\text{RR} = \frac{8/23}{2/22} \approx \frac{0.3478}{0.0909} \approx 3.83$$

To compute the 95% confidence interval, we use the standard error of $\log$:

$$\text{SE} = \sqrt{\frac{1}{8} - \frac{1}{23} + \frac{1}{2} - \frac{1}{22}} \approx \sqrt{0.536} \approx 0.732$$

The CI on the log scale is:

$$\log(3.83) \pm 1.96 \cdot 0.732 \approx (-0.091, 2.777)$$

Exponentiating the bounds gives the 95% CI for the RR:

$$95\%\ \text{CI} \approx (e^{-0.091}, e^{2.777}) \approx (0.91, 16.07)$$

Continuous Outcomes and Mean Differences. Continuous outcomes are measured on a numerical scale, such as a score or a lab value. The typical effect measure is the *mean difference* (MD), which is the arithmetic difference in average outcome values between groups:

$$\text{MD} = \bar{x}_1 - \bar{x}_2$$

where $\bar{x}_1$ and $\bar{x}_2$ are the group means.

Example. In the **Dicker 1992** study (Figure 5):

- Intervention (GnRH agonist): mean = 5.22, SD = 2.22, $n = 48$
- Comparator (control): mean = 3.08, SD = 1.81, $n = 51$

The point estimate is:

$$\text{MD} = 5.22 - 3.08 = 2.14$$

The standard error is computed as:

$$\text{SE} = \sqrt{\frac{2.22^2}{48} + \frac{1.81^2}{51}} \approx 0.408$$

Then, the 95% confidence interval is:

$$2.14 \pm 1.96 \cdot 0.408 \approx 2.14 \pm 0.80 = (1.34, 2.94)$$

B Prompts

The prompts used for synthetic data annotation and for training are shown in Figure 6 and 7, respectively. For training, a temperature of 0.7 and 2,048 tokens as maximum output length are used.

C Manual Annotations Identifying Number Sources

Annotation Instances. Two master's degree students in NLP and co-authors of this work annotated 30 studies paired with outcomes from the COCHRANEFOREST dataset. These studies were randomly sampled from the whole dataset, and each student annotated fifteen studies with an overlap of five studies for cross-comparison. For each study, the annotators were instructed to locate and mark all textual or tabular spans that contained the numerical evidence supporting the reported outcomes. In the case of binary outcomes, this involved identifying four values: the event count and total group size for both the intervention and comparator arms. For continuous outcomes, six values were annotated: the mean, standard deviation, and group size for each group. Each annotated number was linked to one or more corresponding spans from the original study, enabling fine-grained traceability. An example annotation instance is shown in Figure 8.

Inter-annotator Agreement. To assess annotation consistency, we computed inter-annotator agreement (IAA) on the five studies annotated by both annotators. Agreement was evaluated at the span level: a match was counted when both annotators identified overlapping text spans referring to the same numerical value (e.g., intervention group size or mean outcome). We report a span-level F1 score of 0.70, indicating substantial agreement. Disagreements were primarily due to differences in span boundaries or the inclusion of surrounding contextual phrases.

Prompt for synthetic data annotation

{study_content}

The above is a study from a medical systematic review. Your task is to produce a reasoning to explain how to extract the relevant numbers for the following intervention, comparator and outcome:

intervention: {intervention}
comparator: {comparator}
outcome: {outcome}

The expected output is:

{target_value}

You need to first explain how to infer the outcome type. Then, which numbers to extract and why. Finally, point to where these numbers appear in the study.

Figure 6: Prompt for the synthetic data annotation with reasoning traces.

Prompt for training and inference

Articles: {articles}

Question: Based on the given trial articles, what is the outcome type and corresponding numerical data for the following Comparison and Outcome?

Comparison: {comparison}
Outcome: {outcome}

First, determine and output the outcome_type as either: binary or continuous

Then, provide the extracted data in format as follows:
If the outcome is binary, use this format:

outcome_type: binary
intervention:
events: NUMBER total: NUMBER
comparator:
events: NUMBER total: NUMBER

If the outcome is continuous, use this format:

outcome_type: continuous
intervention:
mean: NUMBER standard_deviation: NUMBER group_size: NUMBER
comparator:
mean: NUMBER standard_deviation: NUMBER group_size: NUMBER

Use post-intervention data when both pre and post are available. If multiple timepoints are reported, choose the one closest to the timepoint of interest, or the latest available. Think about it step by step.

Figure 7: Prompt for training and inference.

D Manual Annotations for Extraction Model Outputs

After some pilot studies, we use the following labels to categorize a model's thinking process:

1. *Correct reasoning with traceability*, the reasoning correctly supports the generated numbers and provides sufficient clues to identify their location in the input papers;
2. *Correct reasoning without traceability*, the reasoning correctly supports the generated numbers, but the source of the number cannot be located based on the explanation;
3. *Correct reasoning, incorrect number*, the extracted number is incorrect, but the reasoning is sensible and supports the outcome.
4. *Correct number, incorrect reasoning*, the extracted number is correct, but the reasoning process is incorrect, misleading, or does not support the outcome;
5. *Copy without reasoning*, the numbers are copied from the input papers with no understanding or reasoning steps leading to the final answer;
6. *Hallucinated*, the reasoning process refers to data, methods, or results not present in the input papers;
7. *Missing reasoning*, no reasoning process is generated.

E Hyperparameters and APIs

We executed all the experiments either via API or on our own cluster. We used the paid-for OpenAI API to access GPT-3.5-turbo and GPT-4. On the other hand, we hosted and trained the open-source models used in this paper on a distributed cluster.

SFT is performed for 5 epochs with a batch size of 1 (due to the large size of the input data) using a learning rate of 5×10^{-5} and the AdamW optimiser (Loshchilov and Hutter, 2017). For the RL setup, we adopt the GRPO algorithm (Shao et al., 2024), training for 3 epochs with a learning rate of 1×10^{-6}, batch size 1, and 16 sampled generations per batch. Both training protocols leverage gradient accumulation with 8 accumulation steps. All experiments are conducted using the Open-R1 framework (Hugging Face, 2025) on 8 NVIDIA A100 GPUs, each equipped with 80GB of memory. Models have been served for inference with the vllm framework (Kwon et al., 2023).

The weights of the final reward model have been set to maximise the importance of the correctness reward (0.8) while still ensuring the model follows the YAML format reward (0.1) and the already learnt thought format reward present in the base model (0.1). A tuning of such weights revealed that assigning equal weights to the 3 components leads to a degradation of the performance.

F Scientific Artefacts and Licensing

In this work, we used the following scientific artefacts. LLaMa 3.1 is licensed under a commercial license[2]. GPT-4 is licensed under a commercial license[3]. Qwen2.5 is licensed under the Apache 2.0 license[4]. DeepSeek models are licensed under the MIT license[5]. Mining text and data from the Cochrane library is permitted for non-commercial research through the Wiley API.[6]. The usage of the listed artefacts is consistent with their licenses.

[2] https://llama.meta.com/doc/overview
[3] https://openai.com/policies/terms-of-use
[4] https://qwenlm.github.io/blog/qwen3
[5] https://api-docs.deepseek.com/news/news250120
[6] https://www.cochranelibrary.com/help/access

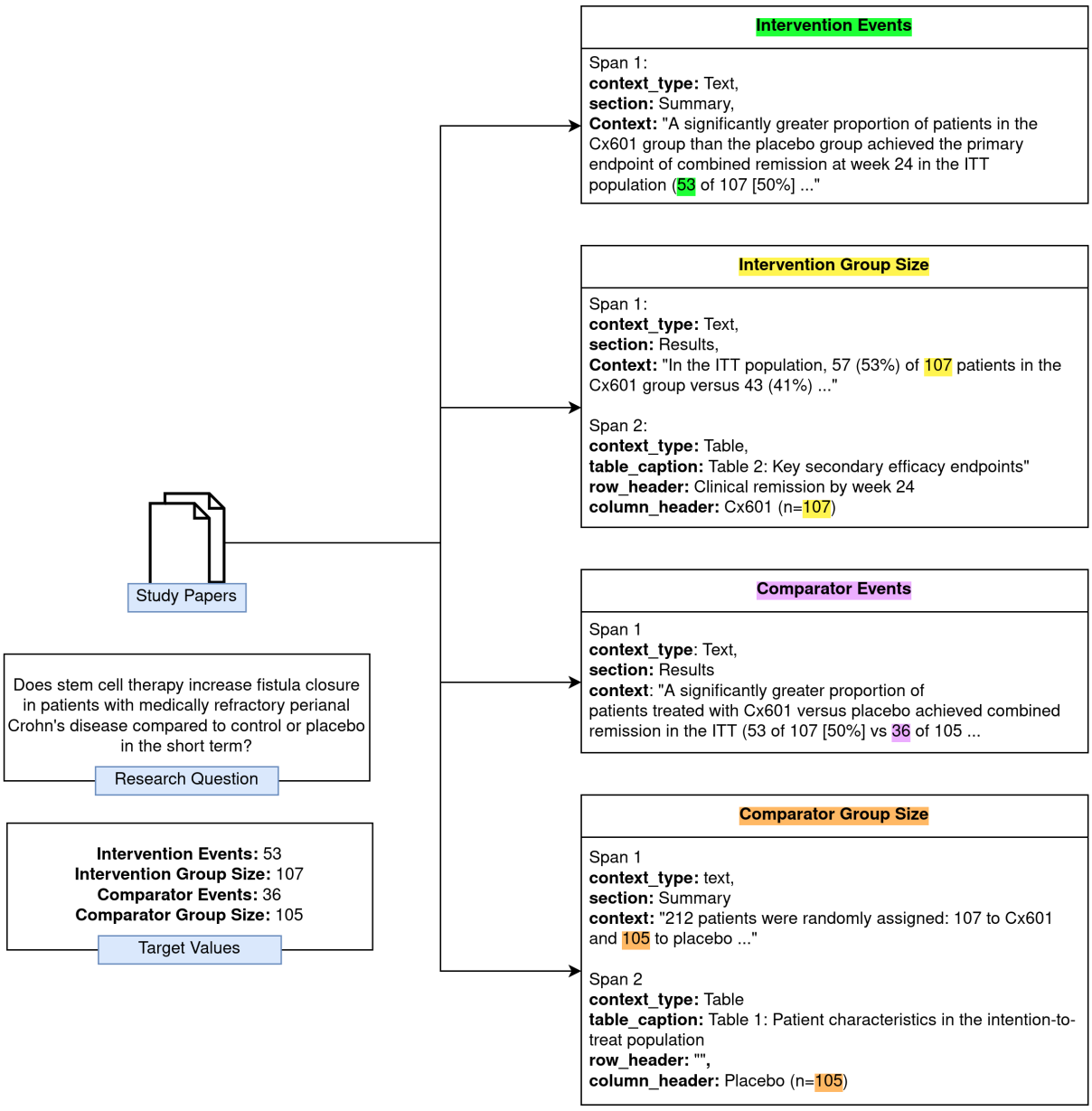

Figure 8: Example of multi-span numerical data annotations for a given study

Input

Original article

Table 3. Number of patients with treatment-emergent AEs (n = 449)*

	Placebo (n = 113)	Belimumab 1.0 mg/kg (n = 114)	Belimumab 4.0 mg/kg (n = 111)	Belimumab 10.0 mg/kg (n = 111)	Belimumab All active (n = 336)
≥1 serious AE	19.5	18.4	13.5	16.2	16.1

Question

Based on the given trial articles, what is the outcome type and corresponding numerical data for the following Comparison and Outcome?
Comparison: Belimumab 10 mg/kg vs placebo
Outcome: Participants with at least one serious adverse event

Output

SFT

<think>
To determine the conclusion for the outcome "Deaths" and comparison "Belimumab 10 mg/kg vs placebo", we first need to identify the type of outcome. Since the outcome is "Deaths", it is a binary outcome (yes/no, died/survived).
We then look for the relevant numbers in the study. The numbers are found in the text: for the intervention (Belimumab 10 mg/kg), there is 1 death out of 111 patients, and for the comparator (placebo), there are 0 deaths out of 113 patients.
</think>

outcome_type: binary
intervention:
events: 10.0 total: 111.0
comparator:
events: 5.0 total: 113.0

RL

<think>
1. First, I need to identify if the outcome "Participants with at least one serious adverse event" is binary or continuous. Since it is counting the number of participants in a specific event, it is binary.
2. Next, I will look for the relevant data in the article for the comparison of Belimumab 10 mg/kg vs placebo.
3. From Table 3, I find the following data:
- For Belimumab 10 mg/kg: 16.2% of patients had at least one serious adverse event.
- For Placebo: 19.5% of patients had at least one serious adverse event.
</think>

outcome_type: binary
intervention:
events: 17 total: 111
comparator:
events: 22 total: 113

Figure 9: Example of a complex example from COCHRANEFOREST. The SFT model fails at inferring the correct numerical information, and its thought mismatches its structured output. On the other hand, the model trained with GRPO correctly infers the number of events, reasoning on the previously extracted total number of participants. **Interestingly, even our human annotators initially struggled to extract the right numbers**.

Hawkey 2015	
Text	Importance Case reports and series suggest hematopoietic stem cell transplantation (HSCT) may benefit some patients with Crohn disease.Objective To evaluate the effect of autologous HSCT on refractory Crohn disease. ...
Outcome	Clinical remission
Comparison	Stem cells versus Control
Outcome type	Binary
Intervention events	8
Intervertion total	23
Comparator events	2
Comparator total	22
Point estimate	3.83
95% CI	[0.91, 16.07]
Target value	outcome_type: binary intervention: events: 8 total: 23 comparator: events: 2 total: 22
Thought	To extract the relevant numbers, we first infer the outcome type as "binary" because the outcome "Clinical remission" is a yes/no or present/absent type of outcome. The relevant numbers for this outcome are found in Table 2: "Final Assessments", under the outcome "CDAI <150 last 3 mo". Specifically, for the intervention group (Stem cells), we have 8 events (patients in clinical remission) out of 23 total patients (34.8%). For the comparator group (Control), we have 2 events out of 22 total patients (9.1%).

Table 6: The corresponding entry in COCHRANEFORESTEXT for Hawekey 2015 in Figure 4. Thought and target value are contatenated for SFT as follows: `<think>{thought}</think>{target_value}`.