Systematic Literature Reviews With Two Multi-Agentic Systems And Human-In-The-Loop

Zexin Ren
George Washington University
Department of Statistics
zxren10@gwu.edu

Zixuan Zhao
George Washington University
Department of Statistics
zixuan.zhao@gwmail.gwu.edu

Qiyun Li
University of San Fransisco
Department of Epidemiology and Biostatistics
qil016@ucsd.edu

Yawen Wu
George Washington University
Department of Statistics
yawen@gwu.edu

Lanjing Wang
University of Washington
Department of Biomedical Informatics and Medical Education
lanjingw@uw.edu

Renjie Luo
George Washington University
Department of Statistics
rluo92@gwmail.gwu.edu

Yi Xu
University of Melbourne
School of Population and Global Health
yix11@student.unimelb.edu.au

Qing Guo
HopeAI, Inc
Statistical Innovation
Qing.Guo@hopeai.co

Jin Shi
HopeAI, Inc
Statistical Innovation
bubbles@hopeai.co

En Xie
HopeAI,Inc
Engineering Group
en@hopeai.co

Feifang Hu*
George Washington University
Department of Statistics
feifang@gwu.edu

Qian Shi
Mayo Clinic
Department of Quantitative Health Sciences
shi.qian2@mayo.edu

Abstract

Systematic literature review of clinical trials drives regulatory decision-making, but conventional screening and extraction are time-consuming, labor-intensive, and vulnerable to study selection bias. We propose two fit-to-purpose multi-agentic systems (MAS) for systematic literature review, with human-in-the-loop. The screening MAS uses multiple LLM agents with heterogeneous personas and multi-round cross-review, and uniformly improves accuracy over a single-LLM baseline. The extraction MAS combines standardization, an iterative correction loop, and retrieval-based context control to ensure accuracy and scalability. Both MAS are specifically designed to support Human-In-The-Loop which is essential for clinical decisions. The novelty of the proposed approach lies in the system architecture rather than in any single foundation tools: the system can naturally benefit from future improvements in the underlying tools, for instance, stronger LLM agents, retrieval engines, image recognition methods, etc. As a real-world application, a

published network meta-analysis is reproduced by the MAS. The result recovers all trials from the original study and identifies additional eligible trials missed by manual review, leading to updated clinical conclusions.

1 Introduction

Before initiating new randomized controlled trials (RCT), a comprehensive analysis of existing literature on similar trials, especially those involving the current standard of care, is crucial for sample size determination. The Food and Drug Administration (FDA) guideline FDA [2018] explicitly highlights the critical role of meta-analysis in drug safety, noting that *"Regulatory decisions related to drug safety are generally taken after considering the totality of available evidence, which may include meta-analytic findings, as well as other factors..."* In practice, systematic literature reviews are time-consuming, labor intensive and may suffer from screening bias. In fact, it takes up to 12 months on average per comprehensive systematic review Borah et al. [2017]. A conventional systematic review, or equivalently, meta-analysis, on RCTs consists of two major steps: a thorough screening of all relevant trials in the trial registry and a careful data extraction to retrieve relevant endpoint information. Manual screening, however, is prone to missed studies and inconsistent decisions, which contributes to study selection bias [Bannach-Brown et al., 2019, Drucker et al., 2016] and manual data extraction is subject to human error and inconsistency, leading to biased estimates Jonnalagadda et al. [2015].

Various tools have been proposed for different purposes [Bannach-Brown et al., 2019, Marshall and Brereton, 2013, Oami et al., 2025] to speed up the process. One commonly discussed approach is the 'Artificial Intelligence-assisted' approach, or more specifically, the utilization of Large Language Models (LLMs) (or equivalently 'agents') in meta-analyses, see, e.g., [de la Torre-López et al., 2023, Li et al., 2025, Ge et al., 2024, Li et al., 2024] and more.

Among those lines of work, Cao et al. [2025b], Li et al. [2024] focus on prompt optimization, either through a well-designed prompt or a iterative prompt optimization method via self-reflecting on inefficient ones. However, the resulting quality depend on specific area of the research question and whether the prompt is tuned for this task, which may lack generalizability. Another line of work is on (Single) Agentic Systems (SAS), utilizing LLMs for each task. It has been discussed in both literature, e.g., Li et al. [2025], Chen and Zhang [2025], Cao et al. [2025a], Oami et al. [2025], and as industry products Elicit.org, DistillerSR. The conceptual workflow of SAS is presented in (1):

$$\text{User's Query} \mapsto \text{A screening LLM} \mapsto \text{An extraction LLM} \mapsto \text{Final data set.} \tag{1}$$

One approach to use SAS is to feed full-text publications for each study into a long-context LLM, such as Claude Opus 4.7, and ask it to extract the required values. As shown in Table 2, this strategy is difficult to scale for systematic reviews for three reasons: (1) token consumption grows rapidly with the number of publications and the complexity of extraction requirements; (2) incorrect extraction can lead to incorrect clinical decisions; and (3) LLM's stochasticity limits reproducibility, especially when no structured human-in-the-loop mechanism is available to audit uncertain cases. These limitations motivate our multi-agent design. For broader surveys of LLM-assisted SLR, see [Khalil et al., 2022, Blaizot et al., 2022, de la Torre-López et al., 2023, Ge et al., 2024, Xu and Sankar, 2025, Li et al., 2026]. A comparison of our method with some representative tools is summarized in Table 1.

Despite the volume of research in this area, the practical adoption of LLMs for clinical decision-making in RCTs remains limited. In this paper, a large semi-automated, hierarchical Multi-Agent System (MAS), consisting of 2 MAS with decentralized planning and decentralized execution Cheng et al. [2024], is proposed to screen for eligible studies and extract relevant endpoint information. Different from previous research that focuses on SAS, we propose to separate screening and extraction into two fit-to-purpose MAS and reserve spaces for **Human-In-The-Loop** for such tasks involving clinical decisions. The novelty of our approach is **system-level**, thus any improvement in frontline LLM tools, e.g., the development of stronger foundation models, will result in the improvement of our system as well.

Our approach offers several unique contributions. First, we decouple screening and data extraction into two fit-to-purpose MAS modules. For screening, the MAS uniformly improves accuracy and stability while pointing out ambiguous cases. For extraction, the MAS combines standardization, iterative correction, and retrieval-based context control to deliver credible extraction while keeping

Table 1: Capability comparison of the proposed system against some AI-assisted literature-review tools as of June 2026. ● = fully supported; ◐ = partial; ○ = not supported or out of scope; **unc** = unclear. † : Supports screening non-NCT registries, e.g., jRCT, ChiCTR, EudraCT, etc. § : Supports screening by major clinical conference, e.g., ASCO, AACR, ESMO, etc.

Capability	This work	Elicit	DistillerSR	Rayyan	ottoSR
Screen by registry†	●	○	○	○	○
Screen by conference§	●	○	○	○	○
Use of AI	●	●	●	●	●
Multi-agentic	●	unc	unc	unc	unc
Data extraction	●	●	●	●	●
RoB assessment	○	○	●	●	●
Automation	◐	◐	●	◐	●
Open-source	◐	◐	○	○	○

token usage under control. More importantly, it provides space for human audition. Together, these design choices support a reproducible, trustworthy workflow for systematic literature reviews of clinical trials. Maybe the least significant advantage our workflow shares with other LLM-based screening methods is the substantial reduction in screening time [Borah et al., 2017].

The rest of the paper is organized as follows. Section 2 introduces the proposed MAS. Section 3 evaluates the method through screening and data-extraction experiments. Section 4 applies the system to a published network meta-analysis. Section 5 summarizes the paper. Prompts, supplementary evaluation results, and supporting network-meta-analysis materials are given in Appendices A–C.

2 Methods

2.1 The Knowledge Base

The infrastructure of the workflow architecture is an establishment of clinical trial knowledge base. The knowledge base can be viewed as a local, dynamically updated repository that mirrors the relevant, up-to-date subset of major clinical trial registries. In this article, we use *ClinicalTrials.gov* as the primary data source. The knowledge base stores structured trial metadata, including key fields such as "Study Overview," "Intervention/Treatment," "Eligibility Criteria," "Outcome Measures," and "Study Start Date." It also stores online published articles for those trials.

2.2 Multi-Agentic Screening with human in-the-loop

In this subsection, we study an MAS for screening trials by NCT identifiers (NCTids).

Figure 1 summarizes the screening MAS. A query-processing agent parses the user's eligibility criteria into structured Boolean inclusion/exclusion rules, which are broadcast to N independent agents $A_1, \ldots, A_N$. Each agent labels every candidate trial as "yes", "no", or "maybe" with a brief rationale supporting its inclusion or exclusion decision.

An inspector agent I checks agreement across the N labels. If all agents agree, the decision is forwarded to the result agent. Otherwise, I generates targeted follow-up questions $Q_1, \ldots, Q_N$ for each agent, initiating a new screening round; this self-revision repeats for at most T rounds. In parallel, N^* external deep-research agents $B_1, \ldots, B_{N^*}$ with web-search tools identify potentially eligible trials not registered with an NCTid (e.g., Japan/China/Korean clinical trial registries ChiCTR, UMIN, KCT, and so on).

Human-in-the-loop. Trials consistently with "yes" or "no" labels are finalized automatically. Trials that remain non-consensus after T rounds, reach consensus on "maybe," or are newly identified by external agents outside the original NCT-based candidate set are flagged for human review.

All agents are LLMs with task-specific prompts (see Appendix A). The MAS structure itself is independent of the LLM, so one can choose a variety of different LLMs as agent. To mitigate LLM mode collapse Zhang et al. [2025b], the screening agents are equipped with heterogeneous personas; see Appendix A.2 for more discussions.

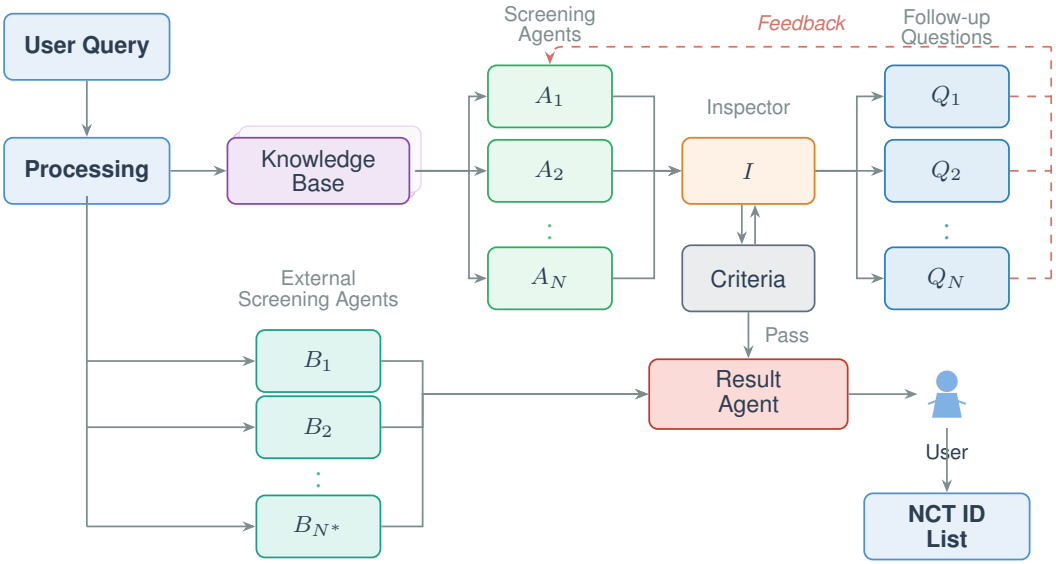

Figure 1: The proposed workflow for NCTid screening. The criteria block returns "Pass" if either the maximum turn, T, is reached or all agents return the same answer.

2.3 Multi-Agentic Data Extraction with human-in-the-loop

Why a single LLM fails. Table 2 illustrates the failure modes of one-shot single-LLM extraction (Using Gemini 3.1 Pro to extract information from Lee et al. [2023]): only 55.6% of OS-related cells are correct, with hallucinations, value substitutions, and missing rows. Stronger long-context models (e.g., Claude Opus 4.8) reduce errors but make token cost scale linearly with publication count, so brute-force long-context extraction does not scale to systematic reviews. The prompt that reproduces Table 2 can be found in Appendix B.3. In contrast, our extraction MAS with the same LLM model recovers all of Table 2 correctly.

Table 2: One-shot LLM extraction of IPSOS trial efficacy outcomes by PD-L1 status. Of 12 reportable subgroup rows, 8 contained at least one extraction error and 2 rows are omitted, illustrating the instability of one-shot extraction for SLR data capture.

Study ID	Treatment	Drugs	PD-L1 Status	Median PFS (Months)			Median OS (Months)			ORR (%)		
				Median	Lower 95% CI	Upper 95% CI	Median	Lower 95% CI	Upper 95% CI	ORR	Lower 95% CI	Upper 95% CI
IPSOS	Experimental	Atezolizumab	– (ITT)	4.2	3.7	5.5	10.3	9.4	11.9	17%	12.8%	21.6%
IPSOS	Control	Vinorelbine or Gemcitabine	– (ITT)	4.0	2.9	5.4	9.2	5.9	11.2	8%	4.2%	13.5%
IPSOS	Experimental	Atezolizumab	TC <1%	NR	NR	NR	10.3	7.1 (NR)	14.3 (NR)	NR	NR	NR
IPSOS	Control	Vinorelbine or Gemcitabine	TC <1%	NR	NR	NR	10.2 (7.1)	5.4 (NR)	12.1 (NR)	NR	NR	NR
IPSOS	Experimental	Atezolizumab	TC ≥1%	NR	NR	NR	10.3 (9.4)	8.4 (NR)	13.9 (NR)	NR	NR	NR
IPSOS	Control	Vinorelbine or Gemcitabine	TC ≥1%	NR	NR	NR	9.4 (10.3)	5.5 (NR)	10.3 (NR)	NR	NR	NR
IPSOS	Experimental	Atezolizumab	TC 1–49%	NR	NR	NR	10.2 (8.4)	NR	NR	NR	NR	NR
IPSOS	Control	Vinorelbine or Gemcitabine	TC 1–49%	NR	NR	NR	8.4 (10.2)	NR	NR	NR	NR	NR
IPSOS	Experimental	Atezolizumab	TC ≥50%	NR	NR	NR	11.0	NR	NR	NR	NR	NR
IPSOS	Control	Vinorelbine or Gemcitabine	TC ≥50%	NR	NR	NR	7.7 (13.9)	NR	NR	NR	NR	NR
IPSOS	Experimental	Atezolizumab	Unknown	NR	NR	NR	16.9	NR	NR	NR	NR	NR
IPSOS	Control	Vinorelbine or Gemcitabine	Unknown	NR	NR	NR	9.4	NR	NR	NR	NR	NR

Red = value returned by the one-shot LLM extraction; green = correct value from the primary publication; yellow row = subgroup the LLM misses to extract. CI, confidence interval; ITT, intent-to-treat; NR, not reported; ORR, objective response rate; OS, overall survival; PFS, progression-free survival; TC, tumor cell.

Given the screened NCTid list in Subsection 2.2 and a new user query specifying extraction details, the extraction MAS (Figure 2) proceeds in three steps: standardization, iterative extraction, and retrieval-based context control.

1. **Standardization.** The first challenge is that the target extraction table is often undetermined before reviewing the publications. This is almost always the case because subgroup definitions may vary across studies or even across multiple papers from the same study. Three specialized agents—a treatment agent (TA), subgroup agent (SA), and endpoint agent (EA)—search the associated publications for all mentions of the query inputs. A standardization agent then unifies the heterogeneous names into a mock table with pre-specified

columns. We provide one mock table column names as an example here: "*NCTid; study popular name; treatment name; arm type (experimental vs control); Cohort; Analysis Population (ITT vs per-Protocol vs Subgroup); Subgroup names; Sample size; Efficacy value; lower* 95% *confidence interval; Upper* 95% *confidence interval.*"

2. **Iterative extraction.** An extraction agent fills the table by first proposing initial values, then re-reading the source PDF to correct them. This correction loop repeats for T rounds, after which each cell receives a confidence label (High, Medium, or Low, see Appendix B.2 for definitions) based on cross-round consistency.
3. **Retrieval-based context control.** A naïve implementation of extraction would pass all publications to the extraction agent in every round and for every table cell, which scales poorly in token cost. In our approach, source documents are automatically partitioned into evidence clips, such as texts, tables and figures. Each clip is indexed by structured attributes. When the extraction agent fills a specific cell, only clips matching the corresponding attributes are retrieved as input. The retrieval layer is implemented using OpenViking Volcengine [2025]. This design reduces irrelevant context.

Human-in-the-loop and source-risk errors. In the MAS, Low- and Medium-confidence results are flagged directly, typically representing no more than 10% of all extracted data points. High-confidence results are additionally flagged under two predefined source-risk conditions: (1) low-quality source material, including low-resolution figures (e.g., Kaplan-Meier plots or forest plots) or compressed supplementary files; and (2) publication-version ambiguity, including unclear data cut-off times or inconsistent values between primary and long-term follow-up reports. In practice, most inaccurate extractions were captured by these review flags, suggesting that the workflow substantially narrows the scope of manual correction while retaining human oversight.

Together, these three components address the principal SAS failure modes: missed information (resolved by iterative extraction with confidence labeling), hallucinated values (resolved by correction of pseudo-answers rather than regeneration [Zhang et al., 2025a] and iterative extraction), and non-standardized output (resolved by the standardization step), while retrieval-based control keeps token cost tractable.

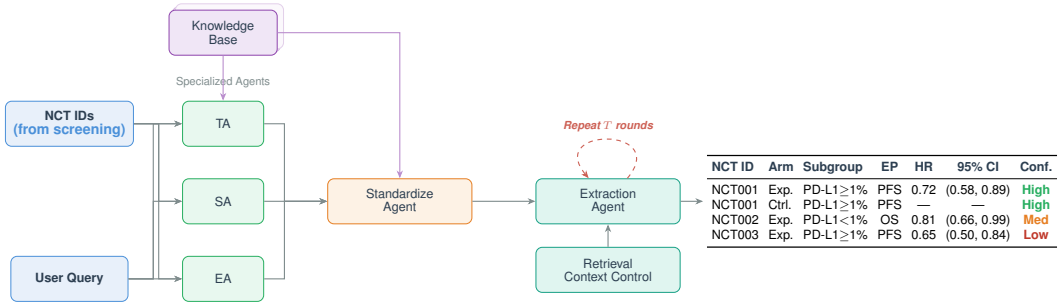

Figure 2: The proposed workflow for data extraction. TA, SA, EA represent treatment arm agent, subgroup agent, endpoints agent, respectively.

2.4 Parallel screening

The huge time cost of producing reliable meta-analyses is well-documented in the literature Allen and Olkin [1999], Borah et al. [2017]. The primary difficulty lies in the manual screening of large trial registries, where researchers must balance speed with the risk of exclusion errors. However, our workflow alters this constraint through parallelization. While our notation $A_1, \ldots, A_N$ implies individual agents, the architecture operationally functions as a distributed array of agentic reviewers, capable of processing the entire Knowledge Base concurrently. Similarly for data extraction. This parallel construction decouples the total processing time from the number of studies; a task that would require 5000 minutes linearly (e.g., 5 minutes per paper for 1000 papers) is theoretically compressed to the latency of a single inference cycle, i.e., in total 5 minutes. Given that this workflow trivializes the temporal burden, the subsequent sections focuses exclusively on the critical remaining difficulty: the accuracy and stability of the screening decisions.

3 Results

In this section, we evaluate the proposed workflow through large-scale screening and data-extraction experiments. Throughout the paper, we consider a wide range of models including Gemini 3.1 Pro Preview, ChatGPT 5.1, DeepSeek Reasoning, Claude Opus 4.8 Reasoning, Gemini 3 Flash Preview, and HopeAI Gemma4,[2] denoted as Gem, GPT, DS, Claude, GemF, and Hope, respectively. To control token consumption, the two large-scale evaluations used DS, GemF, and Hope as primary agent models, whereas the real-world application in Section 4 uses Gem, GPT, and Claude.

3.1 Evaluation of the multi-agentic screening: accuracy and stability

In order to validate accuracy of the screening MAS, we conduct a large scale screening of first line non-small cell lung cancer (NSCLC)[3] trials. In all, 2245 trials [4] from *Clinicaltrials.gov* are included for screening. Furthermore, the trials are screened by the inclusion/exclusion criteria provided in Appendix A.1.Among all trials, 213 are marked to satisfy the eligibility conditions by three independent reviewers blinded to the model outputs (Fleiss' Kappa$= 0.70$, 2 sided p-value 0.00). The reviewers discuss the results and reach consensus across all trials. Then, this final result serves as the **correct benchmark** to be compared with.

To control token consumption in this large-scale experiment, we used three relatively cost-efficient base models: DS, GemF, and Hope. For each base model, we compared the proposed MAS with its corresponding SAS baseline ($N = T = 1$). The MAS used $N = 5$ agents based on the same underlying model, with distinct personas introduced through prompting, and allowed up to $T = 3$ revision rounds; see Appendix A.2. For example, the DS-based MAS was compared with a single DS agent, so that performance differences could be attributed to the MAS workflow rather than to the models themselves. Inspector and summary agents were implemented using Claude. All models were deployed with `temp` $= 0.7$ and `top_p` $= 0.9$. Because the workflow is semi-automated, all flagged cases were adjudicated by an independent reviewer, who recorded the final decision and additional review time.[5]

Table 3 summarizes screening performance across 30 independent Monte Carlo replications on the NSCLC benchmark set, including accuracy, sensitivity, specificity, positive predictive value (PPV), F1 score, Matthews correlation coefficient (MCC), work saved over sampling (WSS), and the number of trials requiring human review. Sensitivity and PPV are particularly important in systematic literature reviews because missed eligible and/or included ineligible trials can distort the decisions. The single-agent models show distinct error profiles. In particular, the single DS agent is overly conservative, achieving near-perfect specificity and PPV but low sensitivity. The MAS corrects this failure mode: DS sensitivity increases from 0.583 to 0.926 while PPV remained high at 0.987. GemF and Hope had stronger single-agent sensitivity, but MAS still improved their F1 scores. Overall, the results indicate that the MAS can reduce model-specific screening errors while maintaining high precision. The agent-disagreement mechanism also provides a principled way for human-in-the-loop, with the average number of cases requiring review depending on the base model. The cost-effectiveness plot is presented in Figure 3, using F1 score as the vertical axis.

Further in Figure S1 in Appendix A.1, the pairwise Cohen's Kappa are reported for the first 10 runs. The average Cohen's Kappa is above 0.95, indicating almost perfect agreement across runs. This reflects the stability of our screening MAS. In addition, we also conduct a smaller scale simulation based approach similar to Li et al. [2024] on colorectal cancer (CRC) with more models. See Appendix A.3 for details.

3.2 Evaluation of multi-agentic data extraction

We evaluated data-extraction accuracy across three disease indications: triple-negative breast cancer (TNBC), non-small-cell lung cancer (NSCLC), and gastric cancer (GC). For each indication, 10 phase 2 and/or phase 3 trials were randomly selected from ClinicalTrials.gov. For each trial, the workflow extracted median PFS, median OS, and objective response rate (ORR), together with their

[2]HopeAI Gemma4 is a customized LLM based on Gemma 4.

[3]The method is not NSCLC-specific and can be extended to other disease indications.

[4]Selected by filter "Phase2/Phase3+Interventional+Industry Sponsored" on *Clinicaltrials.gov*

[5]External agents were not used in validating the screening MAS.

Table 3: Performance of 30-run Monte Carlo screening systems on the NSCLC benchmark ($n = 2236$), grouped by base model. Values are mean (SD) across the 30 runs. **Bold** indicates that the agentic approach strictly outperforms its own single-model baseline on that metric. Rule A: non-consensus resolved by majority votes from agents, "Maybe" resolved by human-in-the-loop. Rule B: "Maybe" and non-consensus resolved by human-in-the-loop. Token: total token consumption, input + output in millions, by the models.

Model	Variant	Acc	Sens	Spec	PPV	F1	MCC	WSS	# Human Review	Token
DS	Single	0.960 (0.002)	0.583 (0.015)	1.000 (0.000)	0.997 (0.005)	0.736 (0.013)	0.746 (0.012)	0.528 (0.014)	150.3 (3.512)	4.1m
	Agentic-A	**0.975** (0.002)	0.790 (0.016)	0.995 (0.001)	0.941 (0.008)	**0.859** (0.012)	**0.849** (0.012)	**0.710** (0.015)	**12.67** (3.606)	17.7m
	Agentic-B	**0.992** (0.001)	**0.926** (0.015)	0.999 (0.001)	0.987 (0.005)	**0.955** (0.007)	**0.951** (0.008)	**0.837** (0.013)	150 (5.6)	17.7m
GemF	Single	0.969 (0.000)	0.991 (0.000)	0.967 (0.000)	0.755 (0.003)	0.857 (0.002)	0.850 (0.002)	0.866 (0.000)	0 (0)	4.3m
	Agentic-A	**0.970** (0.001)	0.988 (0.003)	**0.968** (0.001)	**0.762** (0.006)	**0.861** (0.003)	**0.863** (0.003)	0.865 (0.003)	0 (0.5)	22.8m
	Agentic-B	**0.973** (0.001)	**0.995** (0.000)	**0.971** (0.000)	**0.780** (0.007)	**0.874** (0.004)	**0.867** (0.004)	**0.874** (0.001)	44 (4.7)	22.8m
Hope	Single	0.966 (0.000)	0.941 (0.000)	0.968 (0.000)	0.755 (0.003)	0.839 (0.003)	0.826 (0.003)	0.825 (0.000)	0 (0)	3.8m
	Agentic-A	**0.967** (0.000)	0.941 (0.001)	**0.974** (0.001)	**0.765** (0.003)	**0.844** (0.003)	**0.831** (0.003)	0.824 (0.004)	0 (0)	20.3m
	Agentic-B	**0.973** (0.001)	**0.974** (0.001)	**0.971** (0.000)	**0.796** (0.006)	**0.869** (0.004)	**0.859** (0.005)	**0.844** (0.005)	25 (2.3)	20.3m
Claude	Single	0.965 (0.000)	0.981 (0.001)	0.963 (0.003)	0.736 (0.005)	0.841 (0.010)	0.832 (0.002)	0.855 (0.005)	7 (1.8)	3.1m

Acc: accuracy; Sens: sensitivity; Spec: specificity; PPV: positive predictive value; F1: F1 score; MCC: Matthews correlation coefficient; WSS: work saved over sampling. For # Human Review, lower is better.

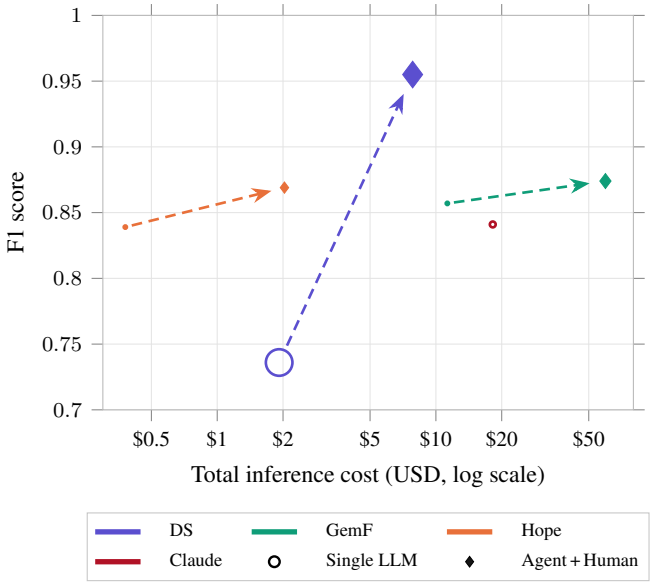

Figure 3: Cost-effectiveness of single-LLM versus agentic (agent + human-in-the-loop) screening on the NSCLC benchmark. The horizontal axis is the *total inference cost* in USD on a log scale, obtained from the input/output token consumption in Table 3 priced at each model's published API rate. Arrows trace the single → agentic trajectory for each model. Marker area is proportional to the number of records left for human review. Upper left is better.

95% confidence intervals, for both the intention-to-treat population and PD-L1-defined subgroups. An extraction is scored as correct only if the reported point estimate, lower 95% confidence limit, and upper 95% confidence limit exactly matches the source publication. Unlike the controlled comparison in Section 3.1, this evaluation uses three different foundation models—GemF, Hope, and DS—to extract together, with 3 rounds of correction. The objective was to assess the feasibility of large-scale structured extraction.

Table 4 shows that the confidence labels provide useful risk stratification. Of 804 extracted entries, 741 (92.2%) were labeled as high confidence, of which 670 (90.4%) exactly matched the source publication. Incorrect high-confidence entries were predominantly attributable to the two source-risk patterns in Section 2.3. With human review of source-risk-flagged high-confidence cells, the extraction accuracy for high-confidence cells rises up to 97%. Further, if all low/medium-confidence cells (63 entries, 7.8%) are also revised the overall extraction accuracy rises to 97%, with 783 out of 804 correct.

Table 4: Extraction accuracy of the data-extraction MAS across three disease indications, three endpoints, and two populations (ITT: intention-to-treat; mPFS/mOS: median progression-free/overall survival; ORR: overall response rate). An extraction is scored correct if and only if all components of the reported quantity (point estimate, lower 95% CI bound, upper 95% CI bound) exactly match the source publication. Confidence labels follow the rule in Appendix B.2.

Indication	Endpoint	Population	Accuracy by confidence level[a]			Overall	Source-risk/LLM error[c]	Token cost
			High	Medium	Low		(Acc. after correction)[c]	(m: millions)
TNBC	mPFS	ITT	16/16 (100.0%)	–/0	0/1	16/17	0/0 (100.0%)	TA/SA/EA:7.7m Extraction:18.5m
		Subgroup[b]	33/41 (80.5%)	–/0	0/3	33/44	6/2 (95.1%)	
	mOS	ITT	15/15 (100.0%)	1/1	0/1	16/17	0/0 (100.0%)	
		Subgroup	38/41 (92.7%)	1/2	1/1	40/44	2/1 (97.6%)	
	ORR	ITT	15/16 (93.8%)	–/0	0/1	15/17	0/1 (93.8%)	
		Subgroup	41/42 (97.6%)	0/1	0/1	41/44	1/0 (100%)	
	TNBC subtotal		*158/171 (92.4%)*	*2/4*	*1/8*	*161/183*	*9/4 (**96.5%**)*	
NSCLC	mPFS	ITT	17/18 (94.4%)	1/2	0/1	18/21	0/1 (94.4%)	TA/SA/EA:11.8m Extraction: 30.3m
		Subgroup	71/76 (93.4%)	–/0	0/3	71/79	5/0 (100%)	
	mOS	ITT	17/19 (89.5%)	0/1	0/1	17/21	2/0 (100%)	
		Subgroup	66/75 (88.0%)	0/3	0/1	66/79	6/3 (96.0%)	
	ORR	ITT	18/21 (85.7%)	–/0	–/0	18/21	2/1 (95.2%)	
		Subgroup	70/77 (90.9%)	0/1	0/1	70/79	4/3 (96.1%)	
	NSCLC subtotal		*259/286 (90.6%)*	*1/7*	*0/7*	*260/300*	*19/8 (**97.0%**)*	
GC	mPFS	ITT	15/16 (93.8%)	1/2	0/2	16/20	1/0 (100%)	TA/SA/EA:12.2m Extraction:29.9m
		Subgroup	67/76 (88.2%)	3/5	3/6	73/87	8/1 (98.6%)	
	mOS	ITT	15/15 (100.0%)	0/2	1/3	16/20	0/0 (100%)	
		Subgroup	66/80 (82.5%)	1/3	2/4	69/87	8/6 (92.5%)	
	ORR	ITT	14/15 (93.3%)	3/4	0/1	17/20	1/0 (100%)	
		Subgroup	76/82 (92.7%)	1/2	1/3	78/87	4/2 (97.6%)	
	GC subtotal		*253/284 (89.1%)*	*9/18*	*7/19*	*269/321*	*22/9 (**96.8%**)*	
Pooled		ITT	142/151 (94.0%)	6/12	1/11	149/174	6/3 (98.0%)	
Pooled		Subgroup	528/590 (89.5%)	6/17	7/23	541/630	44/18 (96.9%)	
Pooled		All	**670/741 (90.4%)**	**12/29**	**8/34**	**690/804**	50/21 (**97.2%**)	

[a] Entries are (correct extractions)/(all extractions), counting efficacy values and 95% CIs.
[b] Any PD-L1 stratification (e.g., TPS/CPS, TC/IC, expression, positive/negative).
[c] The errors are summarized from High Accuracy extractions only. The accuracy after correction takes the correct entries as numerator and all high-confidence extractions as the denominator.

4 Real world application: Reproducing and improving a Network Meta-Analysis

To apply our framework in real world studies, we adapt a published network meta-analysis paper Xu et al. [2021]. The goal is to reproduce and improve the paper's result to "assess the overall survival (OS) of first-line systematic therapies in patients with metastatic CRC" and correct for possible mistakes. We refer readers to Xu et al. [2021] for detailed list of previously included 29 trials.

4.1 Screening results

The criteria [Xu et al., 2021, "*Study Selection*"] is copied as query input for the workflow w/o changes on the scope using $N = 6, T = 5$ and 2 extra agents B_1, B_2. For screening, 5 Gem models (4 with persona and 1 default, see Appendix A.2 for details) and a DS model with strict persona are used. The screening and result agents are Gem and GPT respectively. For each model $\texttt{temp} = 0.7$ and $\texttt{top\_p} = 0.9$. See Figure S3 in Appendix C.1 for a PRISMA flowchart generated by the agentic workflow.

Screening results: In all 1784 clinical trial records are screened by the agents, 1684 reach consensus (with 14 "Yes", and 12 out of 14 reported in the original paper; 1670 "No") and 100 trials do not reach consensus. Among non-consensus results due to the semantic discrepancy, 24 are deemed eligible by two human reviewers and 9 of them readily appeared in the original paper. Thus, a total of 17 extra eligible studies are identified according to the agentic workflow. The originally included

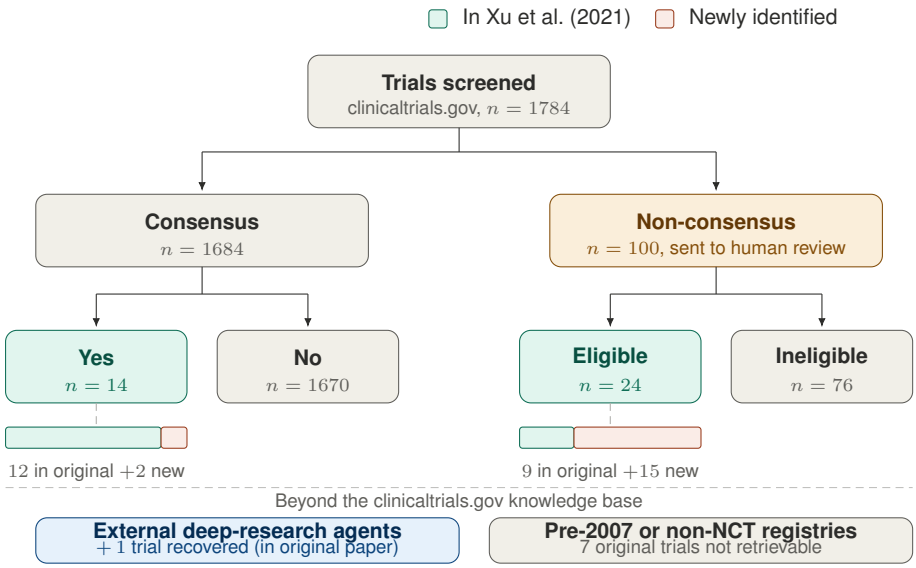

Figure 4: The screening pipeline for the reproduction of Network Meta Analysis.

trials used in the replicated network are summarized in Appendix C.1; newly identified eligible studies are listed in Table 5. The check or cross marks are used to denote whether the trial is included or not in the final network analysis and geometry due to treatment connectivity.

Remark 1. *Among the* 29 *originally included trials in Xu et al. [2021],* 12 *trials reach consensus "Yes" by our workflow;* 9 *do not reach consensus due to the semantic discrepancy;* 1 *is identified by websearch agent after removing duplicates;* 6 *(resp.* 1*) trials are not registered with NCTid because they are conducted before* 2007 *(resp. conducted in Japan). Except the* 7 *trials not included in Clinicaltrials.gov,* ***all studies from the original paper are successfully identified by the proposed workflow****. A comparison of our findings with Xu et al. [2021] is provided in Figure 4.*

To provide a clear view of what are extra on top of the original meta analysis, we present Table 5. The check or cross marks are used to denote whether the trial is included or not in the final network analysis and geometry due to treatment connectivity.

4.2 Efficacy Results

For the network meta-analysis, we extracted the sample size for each treatment arm and the hazard ratio (HR) for OS using the extraction MAS. The output table is used to conduct the network meta-analysis as follows. In the largest connected network among all screened trials, we included 36 trials comprising 45 direct comparisons and 14,859 patients with metastatic colorectal cancer (mCRC). The geometry of the treatment network is shown in Figure S5 in Appendix C.2. Deeper clinical interpretations and the risk of bias was assessed for all included studies and is summarized in Appendix C.2.

5 Conclusion

In this work, we present two semi-automated MAS with Human-In-The-Loop to address some limitations of conventional systematic literature reviews. The screening workflow uses agents with heterogeneous personas to evaluate eligibility in parallel and to self-check decisions through iterative cross-review. The extraction workflow first standardizes treatment, subgroup, and endpoint names, then fills in the numeric values using an iterative language-model loop, with extraction accuracy assessed through repeated extraction and a confidence-level measure. Across both modules, the design explicitly supports and promotes a human-in-the-loop process.

Screening: MAS achieves a uniform improvement. A single agent makes some mistakes on its own, but MAS uniformly improves accuracy upon it. The magnitude of improvement depends

Table 5: Details of the eligible trials screened for reproducing network meta-analysis: the extra screened trials.

NCTid	Popular name	Treatment vs Control (Used or not)
NCT00312000	CAIRO	CAPOXIRI sequential vs CAPOXIRI combination (✗)
NCT00070213	MRC FOCUS2	FOLFOX/CAPOX vs LV5FU2/CAP (✗)
NCT00070213	MRC FOCUS2	CAP/CAPOX vs LV5FU2/FOLFOX (✗)
NCT00003594	N9741 Sanoff et al. [2008]	FOLFOX vs IFL (✗)
NCT00003594	N9741 Sanoff et al. [2008]	IROX vs IFL (✗)
NCT00003287	– Giacchetti et al. [2006]	FOLFOX vs FOLFOX (✗)
NCT00384176	HORIZON III Schmoll et al. [2012]	Cediranib + FOLFOX vs BEV + FOLFOX (✓)
NCT00625651	– Fuchs et al. [2013]	CONA + BEV + FOLFOX vs BEV + FOLFOX (✓)
NCT01071655	SETICC Montagut Viladot and Martinez Balibrea [2018]	BEV + Chemo vs BEV (✗)
NCT00265850	CALGB	CET + FOLFOX/FOLFIRI vs BEV + FOLFOX/FOLFIRI (✗)
NCT01228734	TAILOR Qin et al. [2018]	CET + FOLFOX vs FOLFOX (✓)
NCT01878422	ITAC2	BEV + FOLFOX/FOLFIRI vs FOLFOX/FOLFIRI (✗)
NCT00439517	FUTURE Douillard et al. [2014]	CET + FOLFOX vs CET + UFOX (✓)
NCT00004885	EORTC 40986 Ch [2005]	AIO IRI vs AIO (✗)
NCT00303771	FFCD 2001-02 Aparicio et al. [2016]	FOLFIRI vs LV5FU2 (✓)
NCT01836653	ATOM Oki et al. [2019]	CET + FOLFOX vs BEV + FOLFOX (✓)
NCT00819780	PEAK Schwartzberg et al. [2014]	Panitumumab + FOLFOX vs FOLFOX (✓)
NCT01622543	– Jonker et al. [2018]	BEV + FOLFOX vs FOLFOX (✓)
NCT01721954	FOXFIRE	FOLFOX + SIRT vs FOLFOX (✗)

Drug acronyms: CAPOXIRI = capecitabine + oxaliplatin + irinotecan; LV5FU2 = leucovorin + 5-FU; IFL = irinotecan + bolus 5-FU + leucovorin; IROX = irinotecan + oxaliplatin; HAI = hepatic arterial infusion; CONA = conatumumab; UFOX = UFT + leucovorin + oxaliplatin; AIO = high-dose 5-FU + leucovorin, continuous infusion; SIRT: selective internal radio therapy.

on the failure mode of the underlying model. **Extraction: MAS is necessary.** For extraction the picture is qualitatively different. As Table 2 shows, even good long-context models may produce mistakes that scale poorly with the number of studies, endpoints, and arms. Another critical question is whether such a strategy is scalable and economically reasonable across hundreds of studies and many endpoint/subgroup/arm combinations. Our results indicate that a fit-to-purpose multi-agent design is required for these tasks. **Human-in-the-loop is crucial.** In both modules, human review is built into the workflow. In screening, inter-agent disagreements are passed to reviewers. In extraction, low- and medium-confidence cells together with risky high-confidence cells are flagged for human review. The replication of Xu et al. [2021] illustrates the value of this agent human collaboration.

Both modules still face limitations: semantic discrepancies may require clinical interpretation, performance depends on an up-to-date knowledge base especially for non-NCT or pre-2007 trials, and evaluation is currently limited to oncology. Future work will focus on prompt optimization, prospective regulatory validation, and extension to real-world evidence beyond RCTs.

Acknowledgments

We thank Zhenjun (Will) Ma from HopeAI, Inc and Yunhan Zhou from Queen's University at Canada for their insightful suggestions and discussions, which lead to a substantial improvement of this paper.

Data availability statement

All data required to produce our results are publicly available online, either in `clinicaltrials.gov` or as published trial reports in, e.g., PubMed.

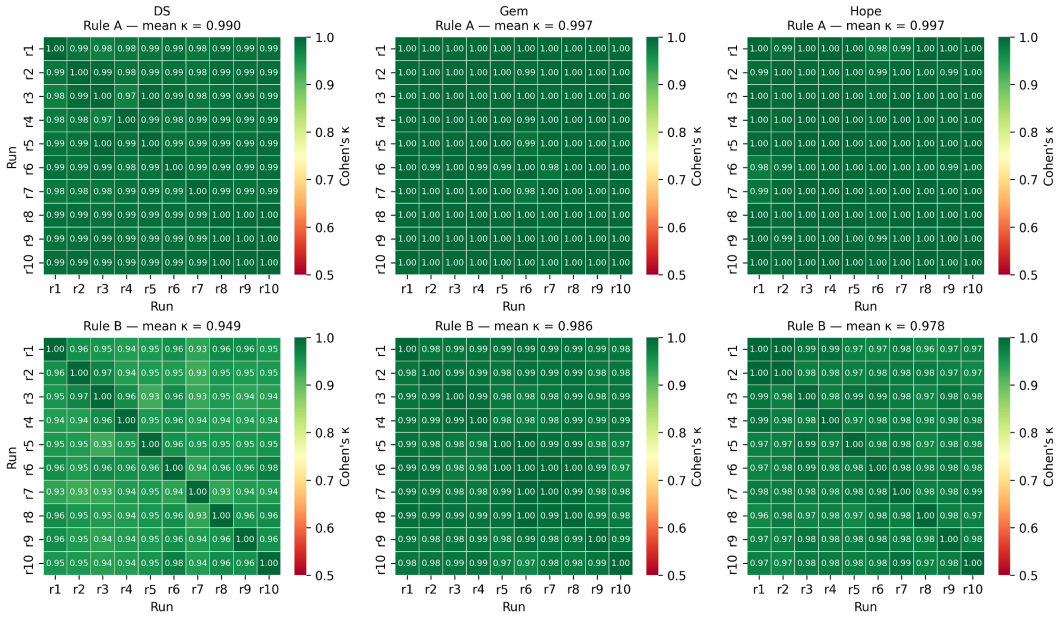

Figure S1: Pairwise inter-run agreement for the first 10 runs of the agentic screening. Cohen's κ closer to 1 indicates higher agreement rate across runs. *Rule A:* non-consensus resolved by majority voting of agents; *Rule B:* non-consensus resolved by human in-the-loop revision.

Supplementary Materials

Systematic Literature Reviews With Two Multi-Agentic Systems And Human-In-The-Loop

A Supplementary Screening Materials

This appendix collects materials supporting the screening MAS: the NSCLC benchmark query and supplementary screening figures, the prompt library, and the CRC simulation-based screening evaluation.

A.1 NSCLC benchmark query and reproducibility checks

The large-scale screening inclusion/exclusion criteria were:

"*First-Line unresectable NSCLC, ECOG 0-1, EGFR/ALK wild-type. If ECOG or genomic status is not mentioned, may include it, provided all other criteria are met. At least one study arm must contain Pembrolizumab or another Immunotherapy (ICI) (e.g., Atezolizumab, Nivolumab, Durvalumab, Cemiplimab). Must not be maintenance treatment. The study start date must be on or after 2013/01/01.*"

A.1.1 Inter-run agreement

Figure S1 shows the pairwise similarity of agentic screening across the first 10 repetitions.

A.2 Screening prompt library

The prompt library for the screening MAS is presented in order. The user query is analyzed via the following prompt.

Query Processing prompt: *You are a Clinical trialist. Your role is to analyze research queries and suggest precise, independent inclusion and exclusion criteria for clinical trial screening. You must*

ensure each criterion is distinct, measurable, and clinically relevant. Now, Analyze the following research query and indication to generate a list of screening criteria. Please (1) break down the query into specific medical concepts (e.g. patient population, intervention, comparator, outcomes, study design). (2) Create individual inclusion or exclusion criteria based on these concepts. (3) Ensure each criterion covers only one concept; use unambiguous phrasing. (4)Use standard medical terminology; persevere with concepts from query, DO NOT flesh out the details. Respond with a JSON object containing: A list of criteria objects, where each object has: "name": A short title for the criterion; "description": The full description of the criterion (start with 'Inclusion:' or 'Exclusion:').

A.2.1 Agent-level screening prompts

To initiate screening, we use the following prompt.

Screening Prompt:*You will be provided with a clinical study (in JSON format) and a set of screening criteria. Here is the provided study data:* {`STUDY_JSON`}. *Here is the target screening criteria:* {`CRITERIA`}. *Determine if the study meets the specified screening criteria.*

The **Screening Prompt** is passed to each screening agent. However, picking homogeneous agents, e.g., chatGPT 5.1, for the same task reduces LLM diversity. This may lead to LLM mode collapse Zhang et al. [2025b], where the same mistakes are made for every agent but are enhanced mistakenly as the correct answer. Therefore, agents with different parameters and personas are recommended with the use of our system. We therefore provide a default prompt and prompts for several different personas (these prompts are attached to each one of the agents).

- **The default prompt:** *You will be provided with the details of a clinical study in JSON format along with screening criteria. Your task is to analyze the study and decide whether it meets the criteria. The study data is as follows* {*...Clinical info in json format provided by knowledge base...*}. *The screening criteria is as follows:* {*user's query*}.
- **Persona 1: Strict regulator** *You are a senior clinical data quality regulatory for a pharmaceutical regulatory body. Your job is to screen clinical trials against specific inclusion/exclusion criteria with EXTREME RIGOR. Adhere strictly to the text. Do not make assumptions. HOWEVER, you must accept standard medical synonyms defined by RECIST 1.1 or CTCAE v5.0. For implied logic that requires a leap of faith, vote '0' (Maybe). Your final goal is to eliminate ineligible trials. If necessary, you prefer False Negatives over False Positives.* {*...Followed by* ***default prompt***.}
- **Persona 2: Permissive clinician** *You are a Board-Certified Medical Oncologist with 20 years of experience in clinical trial recruitment. Your job is to identify patients who are CLINICALLY appropriate for a trial, even if the phrasing is imperfect. If a trial seems eligible based on clinical intent, vote* 1 *(Yes), but note your inference in the reasoning.* {*...Followed by* ***default prompt***.}
- **Persona 3: Statistician** *You are a statistican good at logic and data processing. You do not have many medical opinions; you strictly evaluate Boolean logic, arithmetic, and temporal constraints. Pay close attention to "AND", "OR", and "NOT" operators in the criteria, also pay attention to wordings like "but", "except". Ensure ALL parts of a composite requirement are met. You aim to validate the numbers and logics.* {*...Followed by* ***default prompt***.}
- **Persona 4: Clinical Pharmacologist** *You are a clinical pharmacologist specializing in oncology drug development. Your primary focus is on the INTERVENTION, DRUG HISTORY, and SAFETY criteria. You know how to map generic names (e.g., Pembrolizumab) to classes (e.g., Anti-PD-1) and brand names (e.g., Keytruda). You also understand how to Identify constituent drugs in combos (e.g., FOLFIRI = Irinotecan + 5-FU + Leucovorin).* {*...Followed by* ***default prompt***.}

Finally, here are the two system prompts for the inspector and summary agent.

Inspector Prompt: *You are a Senior Medical Adjudicator for a clinical trial systematic review. You supervise the screening agent. YOUR GOAL: Review answer of the agent efficiently. Be objective.*

Summary Prompt: *You are an experienced Clinical Scientist aggregating results for a systematic review. Your goal is to prepare a final, clean list of eligible NCTids for a human researcher.*

A.3 CRC simulation-based screening evaluation

A.3.1 Eligibility query and benchmark construction

We present the simulation-based results under a designed query given to the workflow and other benchmark methods. A list of 98 colorectal cancer (CRC) trials is served as the full list, among them, 25 are marked to satisfy the conditions 1-5 below by three independent reviewers blinded to the model outputs (Fleiss' Kappa = 0.84, 2 sided p-value 0.00). First, we state the example user's query:

Please find phase 1 *to* 3 *clinical trials that*

1. *include patients with Metastatic CRC deemed unresectable.*
2. *evaluate first-line systemic therapy for metastatic disease. Additionally, trials enrolling patients who received prior adjuvant chemotherapy are also eligible, but only if said the adjuvant therapy was completed* $>$ 6 *months prior to enrollment/relapse.*
3. *must involve a Fluoropyrimidine-based doublet or triplet backbone (e.g., FOLFOX, FOLFIRI, FOLFOXIRI, CAPOX) in the experimental or control arm.*
4. *includes patients without CNS metastases, if the trial is to include patients with CNS metastases, please make sure patients with active or symptomatic central nervous system metastases are excluded (i.e., one can include patients with treated/stable central nervous system metastases or patients without CNS metastases.)*
5. *include all-comers (MSS and MSI-H), or MSS/pMMR only, or RAS/BRAF Wild-Type specific trials but exclude trials exclusively targeting only the MSI-H/dMMR population (e.g., pure Checkpoint Inhibitor trials for MSI-H).*

This query targets several known weaknesses of LLMs. Specifically, the second condition requires agents to identify a patient subset satisfying a temporal constraint, a setting in which single-LLM performance can degrade Song et al. [2025]. The third condition asks agents to decode acronyms for drug combinations. The fourth and fifth conditions require agents to apply complex inclusion/exclusion criteria to distinguish eligible from ineligible trials. Combined with the long trial descriptions provided as context, these demands may cause agents to overlook or forget key requirements Kim et al. [2025], Liu et al. [2024]. The query is intended to closely reflect realistic requests from pharmaceutical companies.

The inclusion criteria were designed by one researcher. The fixed test set consists of 98 CRC trials randomly selected from ClinicalTrials.gov with similar indication-free characteristics, such as completion status and phase. Three independent researchers, blinded to the model output, constructed the answer sheet for this test. Discrepancies between the three reviewers were resolved through a consensus meeting in which the reviewers re-examined the full trial protocols for disputed cases and discussed until full agreement was reached. The final answer sheet with reasons is available upon request. Out of all non-eligible NCT IDs (73 out of 98), 14 violated only 1 criterion; among them, 11 were marked as ambiguous to justify. In addition, 40 NCT IDs violated ≤ 3 criteria.

A.3.2 Performance and review workload

Evaluation protocol: Each MAS and SAS configuration was executed 30 times on the fixed 98-trial set. For non-consensus records, human adjudication was performed once and reused across runs. Table S1 reports means and standard deviations of sensitivity, specificity, PPV, and accuracy across 30 runs. As the model complexity (N, T) increases, non-consensus becomes more likely and additional human effort may be required. Accordingly, the mean consensus rate (defined as Consensus rate $= 1 - \{\text{non-consensus or maybe}\}\%$) and the mean review time in Table S1 quantify the human workload induced by disagreement.

Overall, collaboration between human reviewers and the MAS workflow yields improved screening performance compared with human collaboration with SAS, at the expense of some additional review time.

A.3.3 Stability across repeated runs

Stability: Because the proposed screening workflow contains stochastic components, we evaluate the stability of the induced study-selection rule under repeated executions. Let $f_r(i) \in$

Table S1: Mean (Standard Deviation) accuracy of SAS compared with MAS with human in-the-loop in % for 30 repetitions.

Model(N, T)	Sensitivity	Specificity	PPV	Accuracy	Consensus	Time
Gem$(1, 1)$	90.62 (1.9)	96.57 (0.8)	90.99 (1.9)	94.93 (0.7)	99.50	2
Gem$(5, 3)$	91.71 (2.3)	99.01 (1.1)	97.33 (2.8)	**97.00** (1.0)	93.10	23
GPT$(1, 1)$	91.69 (2.5)	97.06 (0.6)	92.28 (1.5)	95.57 (0.9)	97.69	6
GPT$(5, 3)$	95.60 (2.4)	97.01 (0.4)	93.21 (0.8)	**96.62** (0.6)	96.09	17
DS$(1, 1)$	82.51 (4.7)	97.35 (1.0)	91.33 (3.2)	93.61 (1.3)	95.82	11
DS$(5, 3)$	86.12 (3.3)	99.72 (0.6)	99.14 (1.8)	**96.15** (0.8)	84.05	63
Human	85.07	94.18	90.91	90.82	85.47	479

Among the included agents, A_1 is the default, $A_2, \ldots, A_5$ are equipped with different personas; see Appendix A.2 for the prompts. Dual human review was conducted only once. Review time includes the time of each reviewer and the discussion time. For details of the review process, see Appendix A.3.

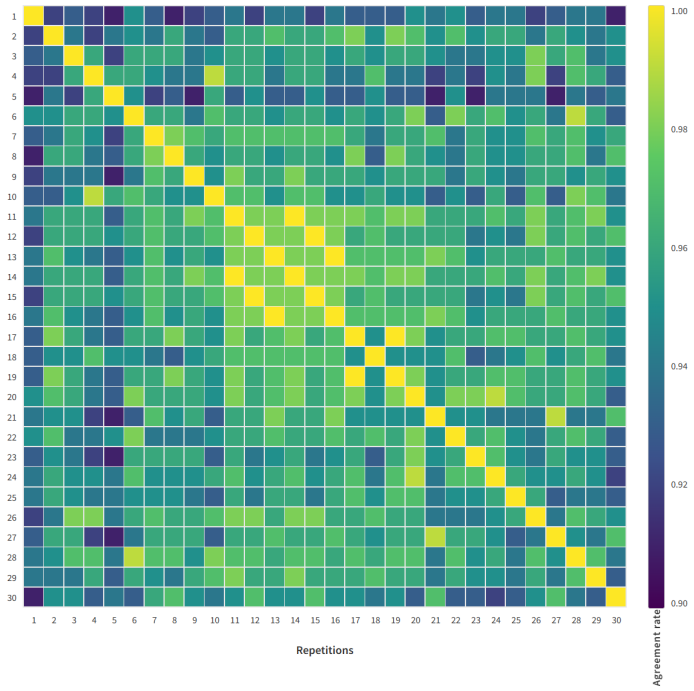

Figure S2: Heat map pairwise agreement matrix for Gemini agentic workflow $(N = 5, T = 3)$. The x and y-axis correspond to 30 runs. Deeper color correspond to lower agreement rate.

$\{\text{yes}, \text{no}, \text{maybe}, \text{non-consensus}\}$ be the final decision of the ith trial for run $r = 1, \ldots, 30$. An agreement matrix is defined as $A_{rs} = n^{-1} \sum_{i=1}^{n} \mathbb{I}\{f_r(i) = f_s(i)\}, \quad r, s = 1, \ldots, 30$, where $n = 98$ is the number of trials and $\mathbb{I}(E)$ is the indicator function of event E. The matrix A_{rs} measures the average number of times the decisions agree between the rth and the sth run. By definition, $A_{rs} = 1$ for all $r = s$. For a purely independent random decision, A_{rs} should be close to 0.25 for off-diagonals.

From Figure S2, we see that over the 30 runs, the Gemini agentic workflow reaches very high agreement rate with a minimal rate 91%. In other words, more than 90% of all the trials reach the same decision for any run of the workflow. Similar results can be observed using GPT (minimal rate 90%) and DS model (minimal rate 87%).

A.3.4 Screened NCT IDs and reviewer agreement

The detailed list of screened 98 NCT IDs with eligible NCT IDs in blue are presented as follows:

Table S2: A three-way agreement table for the test answer provided by independent reviewers.

Reviewer 1	Reviewer 2	Reviewer 3	Count
No	No	No	67
Yes	No	No	1
No	Yes	No	5
No	No	Yes	1
Yes	Yes	No	1
Yes	No	Yes	1
No	Yes	Yes	0
Yes	Yes	Yes	22

NCT00719797, NCT01640405, NCT01765582, NCT02291289, NCT03288987, NCT03721653, NCT04262687, NCT04854668, NCT04607668, NCT02624726, NCT04425239, NCT02384759, NCT02497157, NCT03174405, NCT04547166, NCT03050814, NCT00265850, NCT02305758, NCT01878422, NCT06936527, NCT03635021, NCT03493048, NCT05312398, NCT00885885, NCT02063529, NCT03414983, NCT04194359, NCT04271813, NCT05116085, NCT03827044, NCT02551237, NCT05402891, NCT03280277, NCT05446129, NCT03000374, NCT05510895, NCT03565029, NCT02910843, NCT04015804, NCT04246684, NCT04124601, NCT02727153, NCT04034355, NCT01867697, NCT03170115, NCT02414334, NCT02195232, NCT06448364, NCT01705002, NCT03101475, NCT02404935, NCT02039336, NCT01929616, NCT02390947, NCT03043313, NCT03190174, NCT01320267, NCT03013712, NCT01493336, NCT01689792, NCT00060411, NCT02576665, NCT04244552, NCT05963191, NCT04771520, NCT05396300, NCT03449030, NCT01703910, NCT01164722, NCT02738606, NCT05060471, NCT04595266, NCT02587247, NCT02572141, NCT06137248, NCT04873895, NCT05518201, NCT06547385, NCT02391727, NCT03081494, NCT05609656, NCT03391232, NCT03832621, NCT01455831, NCT03854799, NCT01540344, NCT01842971, NCT01585428, NCT01505166, NCT04281667, NCT05843188, NCT05316818, NCT04380337, NCT01417494, NCT03031444, NCT02624895, NCT03043950, NCT01219920.

The three-way reviewer agreement table used to construct the answer sheet is shown in Table S2.

B Supplementary Data-Extraction Materials

This appendix gives the data-extraction prompt library, the confidence-label rule, and the single-agent extraction baseline used to generate Table 2.

B.1 Extraction prompt library

We provide example system prompts for the three specialized agents TA, SA, EA; the standardization agent; as well as the extraction agent.

Treatment Agent: *You may act as a clinical data scientist. Your goal is to extract treatment arm information from the provided data source. Extraction must ONLY be based on the provided documents. NEVER infer, assume, calculate or generate content beyond what is given in the documents. You should extract the following information for each treatment arm: 1. Type: The type of the treatment arm. It should be one of "Experimental Arm", "Control Arm", or "NP". - Experimental Arm: The arm that receives the treatment being tested. - Control Arm: The arm that receives the control treatment (e.g., placebo, standard of care). - NP: Not Applicable. 2. Drugs: A list of drug names in the treatment arm. 3. Dosage: The dosage of the drugs in the treatment arm. 4. Reason: The reason for the extraction. You should return a JSON list of "Treatment" objects.*

Subgroup Agent: *You are a clinical data scientist. Your goal is to extract subgroup information from the provided data source. You should extract the following information for each subgroup: 1. Name: The name of the subgroup. (Must strictly match one of the target subgroups) 2. Value: The specific value or category of the subgroup (e.g., "< 65", "Male"). 3. GroupID: The Group ID corresponding to the subgroup name. 4. Reason: The reason for the extraction. 5. NotFound: Set to true if and only if the subgroup is not found in the data source.*

Endpoint Agent:*You are a clinical research data scientist. Your task is to extract key information from user queries and structure it for clinical data analysis. Given a user query about clinical trial data extraction, you need to identify and extract: Efficacy Queries (Target Endpoints): Clinical outcomes the user wants to analyze, such as: - PFS (Progression-Free Survival) - OS (Overall Survival) - ORR (Objective Response Rate) - DOR (Duration of Response) - DCR (Disease Control Rate) - OR (Odds Ratio) - HR (Hazard Ratio) - Safety outcomes (adverse events, serious adverse events) - Quality of Life measures Output Format: Return ONLY a valid JSON object. Do not include any explanations or text outside the JSON.*

Standardization Agent:*Your goal is to standardize and merge the provided list of subgroups. You should perform the following operations: 1. Standardize Value: Normalize values to standard clinical terminology. 2. Merge: If multiple entries represent the same subgroup category and value, deduplicate them. Return the standardized list of "Subgroup" objects. Rules: - Output valid JSON. - Do not lose semantic meaning. - Do not create any extra subgroup name. You must keep the original subgroup name provided in the input. You only deduplicate them.*

Extraction Agent:*Your goal is to extract specific clinical data points from the provided document based on provided conditions and queries. - Extraction must ONLY be based on provided data. If provided chunks do not contain an explicit result that matches the given QUERY_CONTEXT, output "Not Reported". - NEVER infer, assume, calculate or generate content beyond what is given in the QUERY_CONTEXT and CONDITION_CONTEXT. - Extracting ONLY information explicitly matches the given QUERY_CONTEXT; otherwise output "Not Reported" (do not infer, paraphrase, or use any information not directly visible in the picture). - Extracted data must match the corresponding treatment arm or subgroup level. - Use the original data with exact decimal or percentage. - Follow the assigned format. If the extracted data does not cover all parts of format, report "Not reported" for that part of output. (e.g. [mean(SD)], only mean value available, the final output will be* 0.5 *(Not reported)). - If multiple options can be applied to the study, return the latest reported option. For instance, if median PFS is reported in both primary and final publication of a trial, return only the final median PFS.*

B.2 Confidence labels

The confidence measure can be defined because the Extraction Agent is run multiple times. In fact, the idea is similar in spirit to the self-consistency strategy in the AI literature Wang et al. [2022], where the LLMs' reasoning is improved by running the same reasoning task multiple times and marginalizing out the incorrect answers.

We adopt the following definitions for the confidence measure (High, Medium, Low), where T is an pre-specified integer indicating the number of repetition rounds and $\lceil x \rceil$ is the ceiling of x.

1. **HIGH:** All T extraction attempts yield the exact same numeric endpoint value, or all said not reported. The source text or table explicitly aligns the extracted data with the requested target population. No mathematical derivation or logical inference was involved.
2. **Medium:** Only T' out of the T ($\lceil T/2 \rceil \leq T' < T$) extraction attempts perfectly match. Alternatively, all three attempts match, but the population mapping required minor logical deduction (e.g., the table header was ambiguous, the figures are low-quality, or the data cut-off is different).
3. **Low:** More than $\lceil T/2 \rceil$ extraction attempts yielded entirely different results. The required data (either the endpoint or the CI) is missing, heavily obscured by complex table formatting, or the population context cannot be reasonably established from the surrounding text.

B.3 Single-agent extraction baseline prompt and protocol

Table 2 reports the output of a single one-shot extraction run using Gemini 3.1 Pro, with the full primary publication and supplementary materials of Lee et al. [2023] supplied as PDF context. The prompt below was written to represent a reasonable, schema-aware first attempt by a competent practitioner; it was not iteratively refined based on the model's output, so that the comparison reflects realistic single-LLM use rather than a deliberately weak baseline.

SAS Prompt: *From this trial, extract and provide an efficacy table (of ITT and pd-l1 subgroup) with the following columns: Study ID, treatment type(exp/control),drugs, pd-l1 status, median pfs, lower*

95% CI of median pfs, upper 95 CI of median pfs, median os, lower 95% CI of median os, upper 95% CI of median os, ORR, lower 95% CI of ORR, upper 95% CI of ORR. Note that for pd-l1 status, you can enter "-", then, the corresponding row will be ITT data.

The output of this run is reproduced in Table 2, with extraction errors highlighted in red and omitted subgroup rows highlighted in yellow.

C Supplementary Network Meta-Analysis Materials

C.1 PRISMA flow and eligible original trials

The PRISMA plot is shown in Figure S3. The plot is drawn following the PRISMA 2020 guideline on updated systematic reviews Page et al. [2021], with modifications for a trial-registry screening workflow.

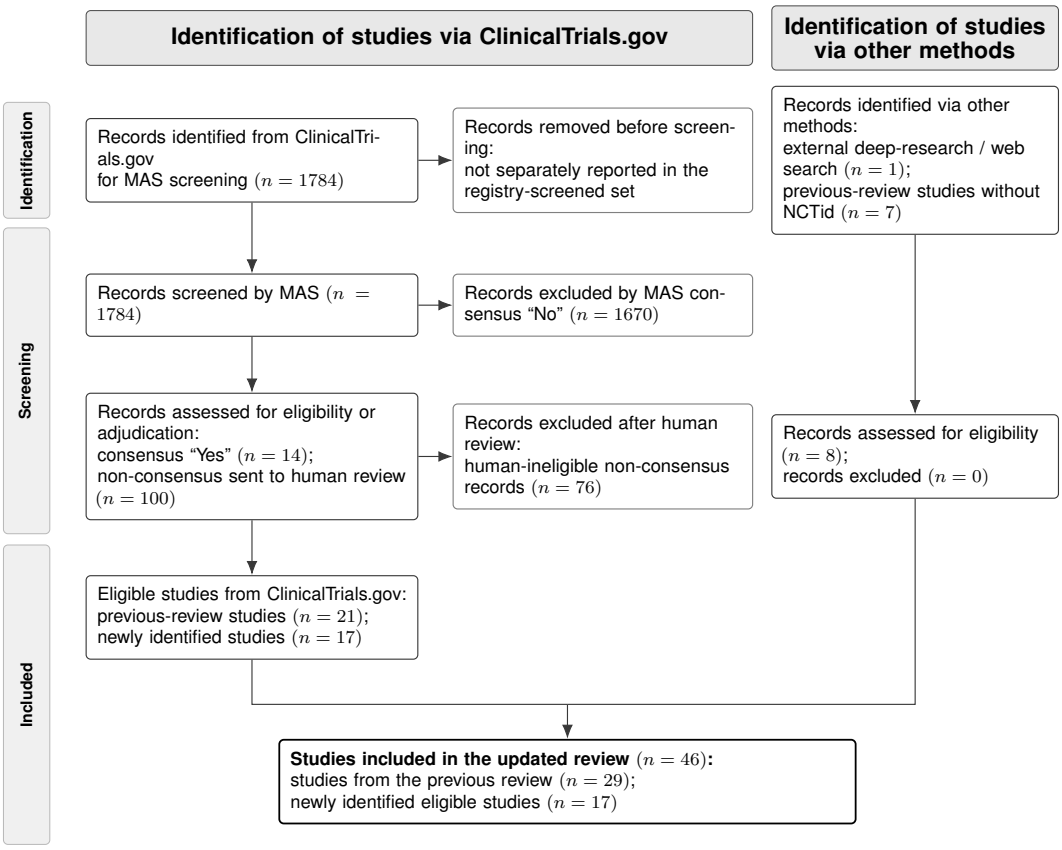

Figure S3: PRISMA flow diagram for the metastatic colorectal cancer network meta-analysis update. The ClinicalTrials.gov stream follows the MAS screening results reported in Section 4; the stream using other methods accounts for studies from the original review that were recovered outside the NCTid-based knowledge base.

Among the 29 trials in the original study (resp. 17 newly identified eligible studies), 13 (resp. 14) included overall survival (OS) results and included in our analysis, see Tables S3 and 5. When Kaplan-Meier OS plots are available but HRs and corresponding CIs were not reported, we estimated them by reconstructing individual patient data with methods IPDfromKM Liu et al. [2021], when the plot is of low resolution; or SynthIPD Zhao et al. [2025], when the plot is of Vector Graphics format. This approach has been used in the original paper Xu et al. [2021]. Included trials in analysis should have matured OS results and should form the largest connected graph of all possible combinations.

Table S3: Details of the eligible trials screened for reproducing network meta-analysis: the originally included trials.

NCTid	Popular name	Treatment (Used or not)
NCT00399750	TREE Hochster et al. [2008]	FOLFOX vs bFOL vs CAPOX (*)
NCT00399750	TREE-2 Hochster et al. [2008]	BEV + (FOLFOX vs bFOL vs CAPOX) (*)
NCT00719797	TRIBE Cremolini et al. [2015][1]	BEV+FOLFOXIRI vs BEV+FOLFIRI (✓)
–	BICC-C Fuchs et al. [2007]	FOLFIRI vs IFL vs CAPIRI (*)
NCT00423696	FNCLCC ACCORD 13/0503 Ducreux et al. [2013]	BEV + FOLFIRI vs BEV+XELIRI (*)
NCT00469443	HORG Pectasides et al. [2012]	BEV + XELIRI vs BEV + FOLFIRI (*)
–	E3200/ECOG E3200 Giantonio et al. [2007]	BEV + FOLFOX vs FOLFOX vs BEV (*)
NCT00154102	CRYSTAL Van Cutsem et al. [2009]	CET+FOLFIRI vs FOLFIRI (✓)
NCT00125034	OPUS Bokemeyer et al. [2011]	CET + FOLFOX vs FOLFOX (*)
NCT00208546	CAIRO2 Tol et al. [2009]	CET+BEV+CAPOX vs BEV+CAPOX (*)
NCT00364013	PRIME Douillard et al. [2010]	Panitumumab + FOLFOX vs FOLFOX (✓)
–	HORG FOLFOXIRI Souglakos et al. [2012]	FOLFOXIRI vs FOLFIRI (*)
NCT01219920	GONO FOLFOXIRI Falcone et al. [2007]	FOLFOXIRI vs FOLFIRI (✓)
–	GOIM trial Colucci et al. [2005]	FOLFIRI vs FOLFOX (✓)
–	Spanish trial Díaz-Rubio et al. [2007]	CAPOX vs FUOX (✓)
–	AIO Colorectal Study Porschen et al. [2007]	CAPOX vs FUFOX (✓)
NCT00126256	FFCD 2000-05 Ducreux et al. [2011]	FOLFOX vs LV5FU2 (✓)
NCT00069095	NO16966 Cassidy et al. [2011]	XELOX vs BEV+XELOX vs FOLFOX vs BEV+FOLFOX (✓)
NCT00433927	FIRE-3/AIO KRK0306 Heinemann et al. [2014]	CET + FOLFIRI vs BEV+FOLFIRI (✓)
NCT00460603	– Infante et al. [2013]	Axitinib + FOLFOX vs BEV+FOLFOX (✓)
NCT00460603	– Infante et al. [2013]	Axitinib + BEV+FOLFOX vs BEV+FOLFOX (✓)
NCT01418222	GO27827 Bendell et al. [2017]	Onartuzumab + BEV+FOLFOX vs BEV+FOLFOX (✓)
NCT01399684	– García-Carbonero et al. [2017]	Parsatuzumab + BEV+FOLFOX vs BEV+FOLFOX (✓)
NCT00677443	– Hong et al. [2012]	SOX vs CAPOX (✓)
NCT00469443	– Souglakos et al. [2012]	BEV+FOLFIRI vs BEV+CAPIRI (✓)
NCT00153998	CELIM Folprecht et al. [2014]	CET+FOLFIRI vs CET+FOLFOX (✓)
NCT01765582	STEAM Hurwitz et al. [2019]	BEV + FOLFOXIRI vs BEV + FOLFOX (✓)
–	CAPIRI GOIM Ciardiello et al. [2016]	CET+FOLFOX vs FOLFOX (✓)
NCT00636610	Vismodegib trial Berlin et al. [2013]	VIS+BEV+FOLFOX/FOLFIRI vs BEV+FOLFOX/FOLFIRI (✗)
UMIN000003253	FLEET Soda et al. [2015]	CET + FOLFOX vs CET + XELOX (✗)

The studies' hazard ratios are estimated by SynthIPD or IPDfromKM.
Drug acronyms: FOLFOX = leucovorin + 5-FU + oxaliplatin; CAPOX = capecitabine + oxaliplatin; BEV = bevacizumab; FOLFOXIRI = FOLFOX + irinotecan; IFL = irinotecan + bolus 5-FU + leucovorin; FOLFIRI = leucovorin + 5-FU + irinotecan; XELIRI (CAPIRI) = capecitabine + irinotecan; CET = cetuximab; SOX = S-1 + oxaliplatin; VIS = Vismodegib.

C.2 Clinical interpretation and risk-of-bias assessment

Based on P-scores (Figure S4), Panitumumab+FOLFOX achieved the highest ranking for OS. The rankings also highlight the benefit of incorporating bevacizumab, as most bevacizumab-based combinations appeared in the upper half of P-scores. In addition, using Bev+FOLFOXIRI as the reference, the estimated HRs (95% CIs) were 1.09 (0.83, 1.44) for Bev+FOLFOX and 1.29 (0.94, 1.78) for Bev+FOLFIRI, respectively, suggesting outcomes favored Bev+FOLFOXIRI; these differences appeared more pronounced than those reported in Xu et al. [2021]. Notably, several top-ranked regimens were evaluated in biomarker-defined populations; for example, trials of panitumumab- or cetuximab-based regimens often restricted enrollment to molecularly selected subgroups (e.g., RAS wild-type tumors). Because our screening criteria does not separate these subgroups from unselected populations, the corresponding P-scores may be inflated. Nevertheless, the overall pattern supports bevacizumab-based combinations, including CAPOX and FOLFOXIRI, as broadly effective options in this setting.

The risk-of-bias assessment results for all studies are presented in Figure S6. Risk of bias was evaluated across seven domains: whether the study has risk on (1) randomization; (2) concealed

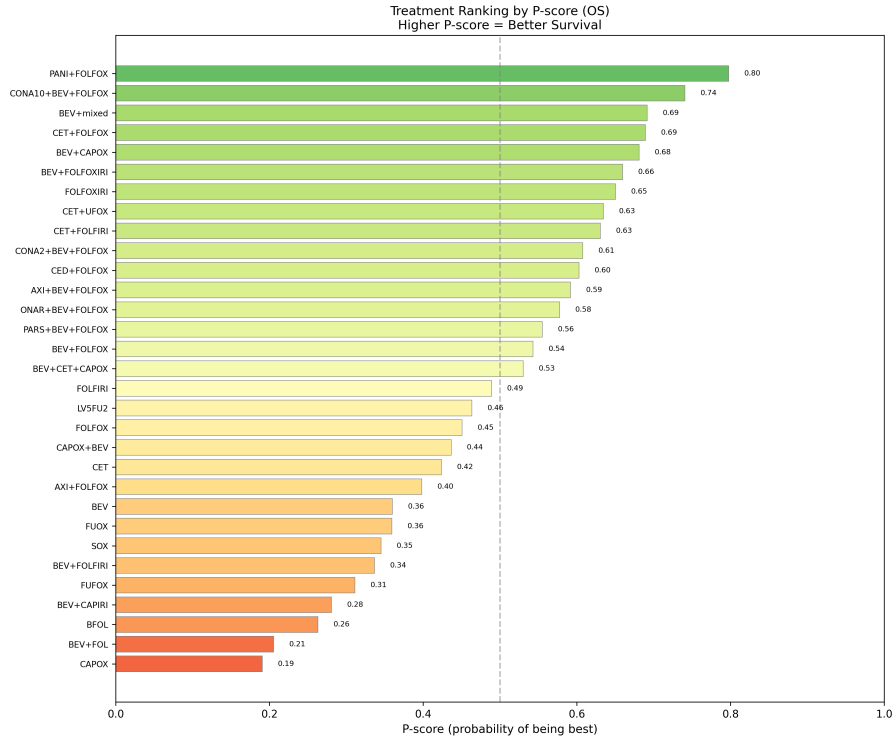

Figure S4: The probability of being best treatment arm (P-score) of all arms considered by the network meta analysis. BEV+mixed = BEV + CAPOX/FUOX/FOLFIRI/FUIRI in Figure S5.

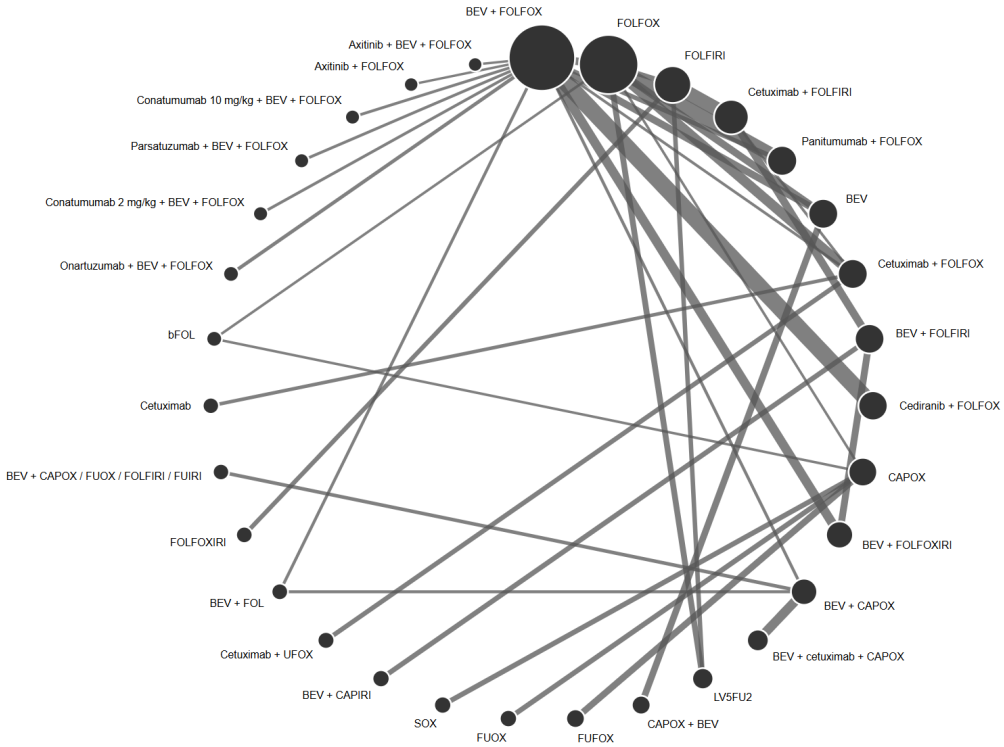

Figure S5: Network geometry plot for the included trials. Solid circles are proportional to total sample size. Line widths are proportional to each trial's (each comparison's) sample size. For the explanation of drug acronyms, see Tables S3, 5.

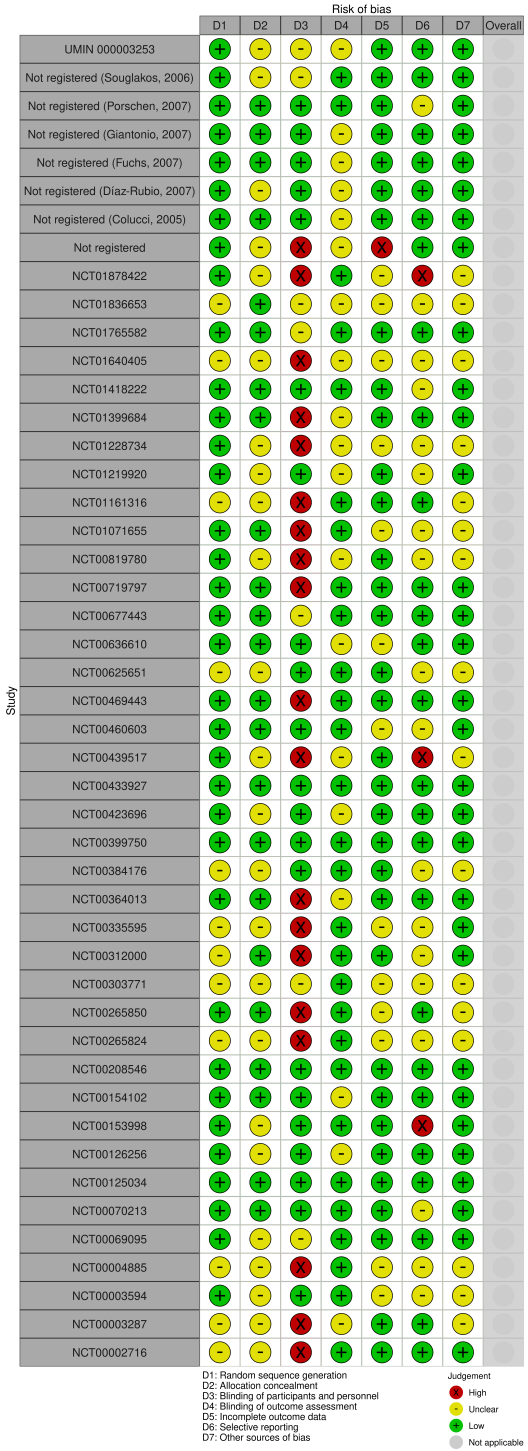

Figure S6: Risk of Bias table for network meta analysis.

allocation; (3) blinding of participants or not; (4) blinding of outcome assessment or not; (5) incomplete outcome data; (6) selective reporting (7) Other bias. Overall, the newly screened studies also show a low risk of bias. Note that all studies underwent risk of bias screening while a subset of them are included in the final Network geometry plot.

References

I Elaine Allen and Ingram Olkin. Estimating time to conduct a meta-analysis from number of citations retrieved. *Jama*, 282(7):634–635, 1999.

T Aparicio, S Lavau-Denes, JM Phelip, E Maillard, JL Jouve, D Gargot, M Gasmi, C Locher, X Adhoute, P Michel, et al. Randomized phase iii trial in elderly patients comparing lv5fu2 with or without irinotecan for first-line treatment of metastatic colorectal cancer (ffcd 2001–02). *Annals of Oncology*, 27(1):121–127, 2016.

Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew SC Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8(1):23, 2019.

Johanna C Bendell, Howard Hochster, Lowell L Hart, Irfan Firdaus, Joseph R Mace, Joshua J McFarlane, Mark Kozloff, Daniel Catenacci, Jessie J Hsu, Stephen P Hack, et al. A phase ii randomized trial (go27827) of first-line folfox plus bevacizumab with or without the met inhibitor onartuzumab in patients with metastatic colorectal cancer. *The oncologist*, 22(3):264–271, 2017.

Jordan Berlin, Johanna C Bendell, Lowell L Hart, Irfan Firdaus, Ira Gore, Robert C Hermann, Mary F Mulcahy, Mark M Zalupski, Howard M Mackey, Robert L Yauch, et al. A randomized phase ii trial of vismodegib versus placebo with folfox or folfiri and bevacizumab in patients with previously untreated metastatic colorectal cancer. *Clinical Cancer Research*, 19(1):258–267, 2013.

Aymeric Blaizot, Sajesh K Veettil, Pantakarn Saidoung, Carlos Francisco Moreno-Garcia, Nirmalie Wiratunga, Magaly Aceves-Martins, Nai Ming Lai, and Nathorn Chaiyakunapruk. Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Research Synthesis Methods*, 13(3):353–362, 2022.

C Bokemeyer, I Bondarenko, JT Hartmann, F De Braud, G Schuch, A Zubel, I Celik, M Schlichting, and P Koralewski. Efficacy according to biomarker status of cetuximab plus folfox-4 as first-line treatment for metastatic colorectal cancer: the opus study. *Annals of oncology*, 22(7):1535–1546, 2011.

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545, 2017.

Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, Matthew Cecere, Anabel Selemon, Jason Sang, Ling Xi Gong, Robert Kloosterman, et al. Automation of systematic reviews with large language models. *medRxiv*, pages 2025–06, 2025a.

Christian Cao, Jason Sang, Rohit Arora, David Chen, Robert Kloosterman, Matthew Cecere, Jaswanth Gorla, Richard Saleh, Ian Drennan, Bijan Teja, et al. Development of prompt templates for large language model–driven screening in systematic reviews. *Annals of Internal Medicine*, 178(3): 389–401, 2025b.

J Cassidy, S Clarke, E Díaz-Rubio, W Scheithauer, A Figer, R Wong, S Koski, K Rittweger, F Gilberg, and L Saltz. Xelox vs folfox-4 as first-line therapy for metastatic colorectal cancer: No16966 updated results. *British journal of cancer*, 105(1):58–64, 2011.

Kohne Ch. Phase iii study of weekly high-dose infusional fluorouracil plus folinic acid with or without irinotecan in patients with metastatic colorectal cancer: European organisation for research and treatment of cancer gastrointestinal group study 40986. *J Clin Oncol*, 23:4856–4865, 2005.

Xi Chen and Xue Zhang. Large language models streamline automated systematic review: A preliminary study. *arXiv preprint arXiv:2502.15702*, 2025.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*, 2024.

Fortunato Ciardiello, Nicola Normanno, Erika Martinelli, Teresa Troiani, Salvatore Pisconti, Claudia Cardone, Anna Nappi, AR Bordonaro, M Rachiglio, Matilde Lambiase, et al. Cetuximab continuation after first progression in metastatic colorectal cancer (capri-goim): a randomized phase ii trial of folfox plus cetuximab versus folfox. *Annals of Oncology*, 27(6):1055–1061, 2016.

Giuseppe Colucci, Vittorio Gebbia, Giancarlo Paoletti, Francesco Giuliani, Michele Caruso, Nicola Gebbia, Giacomo Cartenì, Biagio Agostara, Giuseppe Pezzella, Luigi Manzione, et al. Phase iii randomized trial of folfiri versus folfox4 in the treatment of advanced colorectal cancer: a multicenter study of the gruppo oncologico dell'italia meridionale. *Journal of Clinical Oncology*, 23(22):4866–4875, 2005.

Chiara Cremolini, Fotios Loupakis, Carlotta Antoniotti, Cristiana Lupi, Elisa Sensi, Sara Lonardi, Silvia Mezi, Gianluca Tomasello, Monica Ronzoni, Alberto Zaniboni, et al. Folfoxiri plus bevacizumab versus folfiri plus bevacizumab as first-line treatment of patients with metastatic colorectal cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3 tribe study. *The Lancet Oncology*, 16(13):1306–1315, 2015.

José de la Torre-López, Aurora Ramírez, and José Raúl Romero. Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10):2171–2194, 2023.

Eduardo Díaz-Rubio, Jose Tabernero, Auxiliadora Gómez-España, Bartomeu Massutí, Javier Sastre, Manuel Chaves, Alberto Abad, Alfredo Carrato, Bernardo Queralt, Juan José Reina, et al. Phase iii study of capecitabine plus oxaliplatin compared with continuous-infusion fluorouracil plus oxaliplatin as first-line therapy in metastatic colorectal cancer: final report of the spanish cooperative group for the treatment of digestive tumors trial. *Journal of Clinical Oncology*, 25(27):4224–4230, 2007.

Jean-Yves Douillard, Salvatore Siena, James Cassidy, Josep Tabernero, Ronald Burkes, Mario Barugel, Yves Humblet, György Bodoky, David Cunningham, Jacek Jassem, et al. Randomized, phase iii trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (folfox4) versus folfox4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the prime study. *Journal of clinical oncology*, 28(31):4697–4705, 2010.

Jean-Yves Douillard, Tomasz Zemelka, George Fountzilas, Carlo Barone, Michael Schlichting, Jim Heighway, S Peter Eggleton, and Vichien Srimuninnimit. Folfox4 with cetuximab vs. ufox with cetuximab as first-line therapy in metastatic colorectal cancer: The randomized phase ii future study. *Clinical Colorectal Cancer*, 13(1):14–26, 2014.

Aaron M Drucker, Patrick Fleming, and An-Wen Chan. Research techniques made simple: assessing risk of bias in systematic reviews. *Journal of Investigative Dermatology*, 136(11):e109–e114, 2016.

M Ducreux, Antoine Adenis, J-P Pignon, E François, B Chauffert, JL Ichanté, E Boucher, M Ychou, J-Y Pierga, C Montoto-Grillot, et al. Efficacy and safety of bevacizumab-based combination regimens in patients with previously untreated metastatic colorectal cancer: final results from a randomised phase ii study of bevacizumab plus 5-fluorouracil, leucovorin plus irinotecan versus bevacizumab plus capecitabine plus irinotecan (fnclcc accord 13/0503 study). *European journal of cancer*, 49(6):1236–1245, 2013.

Michel Ducreux, David Malka, Jean Mendiboure, Pierre-Luc Etienne, Patrick Texereau, Dominique Auby, Philippe Rougier, Mohamed Gasmi, Marine Castaing, Moncef Abbas, et al. Sequential versus combination chemotherapy for the treatment of advanced colorectal cancer (ffcd 2000–05): an open-label, randomised, phase 3 trial. *The lancet oncology*, 12(11):1032–1044, 2011.

Alfredo Falcone, Sergio Ricci, Isa Brunetti, Elisabetta Pfanner, Giacomo Allegrini, Cecilia Barbara, Lucio Crino, Giovanni Benedetti, Walter Evangelista, Laura Fanchini, et al. Phase iii trial of infusional fluorouracil, leucovorin, oxaliplatin, and irinotecan (folfoxiri) compared with infusional fluorouracil, leucovorin, and irinotecan (folfiri) as first-line treatment for metastatic colorectal cancer: the gruppo oncologico nord ovest. *Journal of clinical oncology*, 25(13):1670–1676, 2007.

FDA. Meta-analyses of randomized controlled clinical trials to evaluate the safety of human drugs or biological products. *Center for Drug Evaluation and Research (CBER)*, 2018.

Gunnar Folprecht, Thomas Gruenberger, W Bechstein, H-R Raab, Juergen Weitz, Florian Lordick, Joerg Thomas Hartmann, Jan Stoehlmacher-Williams, Hauke Lang, Tanja Trarbach, et al. Survival of patients with initially unresectable colorectal liver metastases treated with folfox/cetuximab or folfiri/cetuximab in a multidisciplinary concept (celim study). *Annals of oncology*, 25(5): 1018–1025, 2014.

Charles S Fuchs, John Marshall, Edith Mitchell, Rafal Wierzbicki, Vinod Ganju, Mark Jeffery, Joseph Schulz, Donald Richards, Raoudha Soufi-Mahjoubi, Benjamin Wang, et al. Randomized, controlled trial of irinotecan plus infusional, bolus, or oral fluoropyrimidines in first-line treatment of metastatic colorectal cancer: results from the bicc-c study. *Journal of Clinical Oncology*, 25 (30):4779–4786, 2007.

Charles S Fuchs, Marwan Fakih, Lee Schwartzberg, Allen L Cohn, Lorrin Yee, Luke Dreisbach, Mark F Kozloff, Yong-jiang Hei, Francesco Galimi, Yang Pan, et al. Trail receptor agonist conatumumab with modified folfox6 plus bevacizumab for first-line treatment of metastatic colorectal cancer: a randomized phase 1b/2 trial. *Cancer*, 119(24):4290–4298, 2013.

Rocío García-Carbonero, Eric van Cutsem, Fernando Rivera, Jacek Jassem, Ira Gore Jr, Niall Tebbutt, Fadi Braiteh, Guillem Argiles, Zev A Wainberg, Roel Funke, et al. Randomized phase ii trial of parsatuzumab (anti-egfl7) or placebo in combination with folfox and bevacizumab for first-line metastatic colorectal cancer. *The oncologist*, 22(4):375–e30, 2017.

Lixia Ge, Rupesh Agrawal, Maxwell Singer, Palvannan Kannapiran, Joseph Antonio De Castro Molina, Kiok Liang Teow, Chun Wei Yap, and John Arputhan Abisheganaden. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. *Systematic reviews*, 13(1):269, 2024.

Sylvie Giacchetti, Georg Bjarnason, Carlo Garufi, Dominique Genet, Stefano Iacobelli, Marco Tampellini, Rune Smaaland, Christian Focan, Bruno Coudert, Yves Humblet, et al. Phase iii trial comparing 4-day chronomodulated therapy versus 2-day conventional delivery of fluorouracil, leucovorin, and oxaliplatin as first-line chemotherapy of metastatic colorectal cancer: the european organisation for research and treatment of cancer chronotherapy group. *Journal of clinical oncology*, 24(22):3562–3569, 2006.

Bruce J Giantonio, Paul J Catalano, Neal J Meropol, Peter J O'Dwyer, Edith P Mitchell, Steven R Alberts, Michael A Schwartz, and Al B Benson III. Bevacizumab in combination with oxaliplatin, fluorouracil, and leucovorin (folfox4) for previously treated metastatic colorectal cancer: results from the eastern cooperative oncology group study e3200. *Journal of clinical oncology*, 25(12): 1539–1544, 2007.

Volker Heinemann, Ludwig Fischer von Weikersthal, Thomas Decker, Alexander Kiani, Ursula Vehling-Kaiser, Salah-Eddin Al-Batran, Tobias Heintges, Christian Lerchenmüller, Christoph Kahl, Gernot Seipelt, et al. Folfiri plus cetuximab versus folfiri plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (fire-3): a randomised, open-label, phase 3 trial. *The lancet oncology*, 15(10):1065–1075, 2014.

Howard S Hochster, Lowell L Hart, Ramesh K Ramanathan, Barrett H Childs, John D Hainsworth, Allen L Cohn, Lucas Wong, Louis Fehrenbacher, Yousif Abubakr, M Wasif Saif, et al. Safety and efficacy of oxaliplatin and fluoropyrimidine regimens with or without bevacizumab as first-line treatment of metastatic colorectal cancer: results of the tree study. *Journal of Clinical Oncology*, 26(21):3523–3529, 2008.

Yong Sang Hong, Young Suk Park, Ho Yeong Lim, Jeeyun Lee, Tae Won Kim, Kyu-pyo Kim, Sun Young Kim, Ji Yeon Baek, Jee Hyun Kim, Keun-Wook Lee, et al. S-1 plus oxaliplatin versus capecitabine plus oxaliplatin for first-line treatment of patients with metastatic colorectal cancer: a randomised, non-inferiority phase 3 trial. *The lancet oncology*, 13(11):1125–1132, 2012.

Herbert I Hurwitz, Benjamin R Tan, James A Reeves, Henry Xiong, Brad Somer, Heinz-Josef Lenz, Howard S Hochster, Frank Scappaticci, John F Palma, Richard Price, et al. Phase ii randomized trial of sequential or concurrent folfoxiri-bevacizumab versus folfox-bevacizumab for metastatic colorectal cancer (steam). *The oncologist*, 24(7):921–932, 2019.

Jeffrey R Infante, Tony R Reid, Allen L Cohn, William J Edenfield, Terrence P Cescon, John T Hamm, Imtiaz A Malik, Thomas A Rado, Philip J McGee, Donald A Richards, et al. Axitinib and/or bevacizumab with modified folfox-6 as first-line therapy for metastatic colorectal cancer: a randomized phase 2 study. *Cancer*, 119(14):2555–2563, 2013.

Derek J Jonker, Patricia A Tang, Hagen Kennecke, Stephen A Welch, M Christine Cripps, Timothy Asmis, Haji Chalchal, Anna Tomiak, Howard Lim, Yoo-Joung Ko, et al. A randomized phase ii study of folfox6/bevacizumab with or without pelareorep in patients with metastatic colorectal cancer: Ind. 210, a canadian cancer trials group trial. *Clinical Colorectal Cancer*, 17(3):231–239, 2018.

Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):78, 2015.

Hanan Khalil, Daniel Ameen, and Armita Zarnegar. Tools to support the automation of systematic reviews: a scoping review. *Journal of Clinical Epidemiology*, 144:22–42, 2022.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426, 2025.

Siow Ming Lee, Christian Schulz, Kumar Prabhash, Dariusz Kowalski, Aleksandra Szczesna, Baohui Han, Achim Rittmeyer, Toby Talbot, David Vicente, Raffaele Califano, et al. First-line atezolizumab monotherapy versus single-agent chemotherapy in patients with non-small-cell lung cancer ineligible for treatment with a platinum-containing regimen (ipsos): a phase 3, global, multicentre, open-label, randomised controlled study. *The Lancet*, 402(10400):451–463, 2023.

Dan Li, Leihong Wu, Svitlana Shpyleva, Ting Li, and Joshua Xu. Enhancing systematic literature review with advanced ai: A study based on llm screening. Technical report, FDA 2024 Digital Transformation Symposium, 2024.

Lingbo Li, Anuradha Mathrani, and Teo Susnjak. Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of ai. *Research Synthesis Methods*, page 1–48, 2026. doi: 10.1017/rsm.2025.10065.

Ying Li, Surabhi Datta, Majid Rastegar-Mojarad, Kyeryoung Lee, Hunki Paek, Julie Glasgow, Chris Liston, Long He, Xiaoyan Wang, and Yingxin Xu. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *Journal of the American Medical Informatics Association*, 32(4):616–625, 2025.

Na Liu, Yanhong Zhou, and J Jack Lee. Ipdfromkm: reconstruct individual patient data from published kaplan-meier survival curves. *BMC medical research methodology*, 21(1):111, 2021.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Christopher Marshall and Pearl Brereton. Tools to support systematic literature reviews in software engineering: A mapping study. In *2013 ACM/IEEE international symposium on empirical software engineering and measurement*, pages 296–299. IEEE, 2013.

Clara Montagut Viladot and Eva Martinez Balibrea. Genotype-based selection of treatment for patients with advanced colorectal cancer (seticc): a pharmacogenetic-based randomized phase ii trial. *ANNALS OF ONCOLOGY, 2018, vol. 29, núm. 2, p. 439-444*, 2018.

Takehiko Oami, Yohei Okada, and Taka-aki Nakada. Optimal large language models to screen citations for systematic reviews. *Research Synthesis Methods*, pages 1–17, 2025.

Eiji Oki, Yasunori Emi, Takeharu Yamanaka, Hiroyuki Uetake, Kei Muro, Takao Takahashi, Takeshi Nagasaka, Etsuro Hatano, Hitoshi Ojima, Dai Manaka, et al. Randomised phase ii trial of mfolfox6 plus bevacizumab versus mfolfox6 plus cetuximab as first-line treatment for colorectal liver metastasis (atom trial). *British journal of cancer*, 121(3):222–229, 2019.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

Dimitrios Pectasides, George Papaxoinis, Konstantine T Kalogeras, Anastasia G Eleftheraki, Ioannis Xanthakis, Thomas Makatsoris, Epaminondas Samantas, Ioannis Varthalitis, Pavlos Papakostas, Nikitas Nikitas, et al. Xeliri-bevacizumab versus folfiri-bevacizumab as first-line treatment in patients with metastatic colorectal cancer: a hellenic cooperative oncology group phase iii trial with collateral biomarker analysis. *BMC cancer*, 12(1):271, 2012.

Rainer Porschen, Hendrik-Tobias Arkenau, Stephan Kubicka, Richard Greil, Thomas Seufferlein, Werner Freier, Albrecht Kretzschmar, Ullrich Graeven, Axel Grothey, Axel Hinke, et al. Phase iii study of capecitabine plus oxaliplatin compared with fluorouracil and leucovorin plus oxaliplatin in metastatic colorectal cancer: a final report of the aio colorectal study group. *Journal of Clinical Oncology*, 25(27):4217–4223, 2007.

Shukui Qin, Jin Li, Liwei Wang, Jianming Xu, Ying Cheng, Yuxian Bai, Wei Li, Nong Xu, Li-zhu Lin, Qiong Wu, et al. Efficacy and tolerability of first-line cetuximab plus leucovorin, fluorouracil, and oxaliplatin (folfox-4) versus folfox-4 in patients with ras wild-type metastatic colorectal cancer: the open-label, randomized, phase iii tailor trial. *Journal of Clinical Oncology*, 36(30):3031–3039, 2018.

Hanna K Sanoff, Daniel J Sargent, Megan E Campbell, Roscoe F Morton, Charles S Fuchs, Ramesh K Ramanathan, Stephen K Williamson, Brian P Findlay, Henry C Pitot, and Richard M Goldberg. Five-year data and prognostic factor analysis of oxaliplatin and irinotecan combinations for advanced colorectal cancer: N9741. *Journal of Clinical Oncology*, 26(35):5721–5727, 2008.

Hans-Joachim Schmoll, David Cunningham, Alberto Sobrero, Christos S Karapetis, Philippe Rougier, Sheryl L Koski, Ilona Kocakova, Igor Bondarenko, György Bodoky, Paul Mainwaring, et al. Cediranib with mfolfox6 versus bevacizumab with mfolfox6 as first-line treatment for patients with advanced colorectal cancer: a double-blind, randomized phase iii study (horizon iii). *Journal of clinical oncology*, 30(29):3588–3595, 2012.

Lee S Schwartzberg, Fernando Rivera, Meinolf Karthaus, Gianpiero Fasola, Jean-Luc Canon, J Randolph Hecht, Hua Yu, Kelly S Oliner, and William Y Go. Peak: a randomized, multicenter phase ii study of panitumumab plus modified fluorouracil, leucovorin, and oxaliplatin (mfolfox6) or bevacizumab plus mfolfox6 in patients with previously untreated, unresectable, wild-type kras exon 2 metastatic colorectal cancer. *Journal of clinical oncology*, 32(21):2240–2247, 2014.

Hitoshi Soda, Hiromichi Maeda, Junichi Hasegawa, Takao Takahashi, Shoichi Hazama, Mutsumi Fukunaga, Emiko Kono, Masahito Kotaka, Junichi Sakamoto, Naoki Nagata, et al. Multicenter phase ii study of folfox or biweekly xelox and erbitux (cetuximab) as first-line therapy in patients with wild-type kras/braf metastatic colorectal cancer: The fleet study. *BMC cancer*, 15(1):695, 2015.

Xintong Song, Bin Liang, Yang Sun, Chenhua Zhang, Bingbing Wang, and Ruifeng Xu. Bridging time gaps: Temporal logic relations for enhancing temporal reasoning in large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3040–3044, 2025.

John Souglakos, N Ziras, S Kakolyris, I Boukovinas, N Kentepozidis, P Makrantonakis, S Xynogalos, Ch Christophyllakis, Ch Kouroussis, L Vamvakas, et al. Randomised phase-ii trial of capiri (capecitabine, irinotecan) plus bevacizumab vs folfiri (folinic acid, 5-fluorouracil, irinotecan) plus bevacizumab as first-line treatment of patients with unresectable/metastatic colorectal cancer (mcrc). *British journal of cancer*, 106(3):453–459, 2012.

Jolien Tol, Miriam Koopman, Annemieke Cats, Cees J Rodenburg, Geert JM Creemers, Jolanda G Schrama, Frans LG Erdkamp, Allert H Vos, Cees J van Groeningen, Harm AM Sinnige, et al. Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. *New England Journal of Medicine*, 360(6):563–572, 2009.

Eric Van Cutsem, Claus-Henning Köhne, Erika Hitre, Jerzy Zaluski, Chung-Rong Chang Chien, Anatoly Makhson, Geert D'Haens, Tamás Pintér, Robert Lim, György Bodoky, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *New England Journal of Medicine*, 360(14):1408–1417, 2009.

Volcengine. OpenViking: An open-source context database for ai agents. `https://github.com/volcengine/OpenViking`, 2025. Accessed: 2026-05-05.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Shan Xu, Ali Sak, and Yasin Bahadir Erol. Network meta-analysis of first-line systemic treatment for patients with metastatic colorectal cancer. *Cancer Control*, 28:10732748211033497, 2021.

Xiaoran Xu and Ravi Sankar. Large language model agents for biomedicine: a comprehensive review of methods, evaluations, challenges, and future directions. *Information*, 16(10):894, 2025.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models, 2025a. URL `https://arxiv.org/abs/2512.24601`.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*, 2025b.

Zixuan Zhao, Zexin Ren, Guannan Zhai, Feifang Hu, Will Ma, En Xie, and Qian Shi. Synthipd: assumption-lean synthetic individual patient data generation. *arXiv preprint arXiv:2509.16466*, 2025.