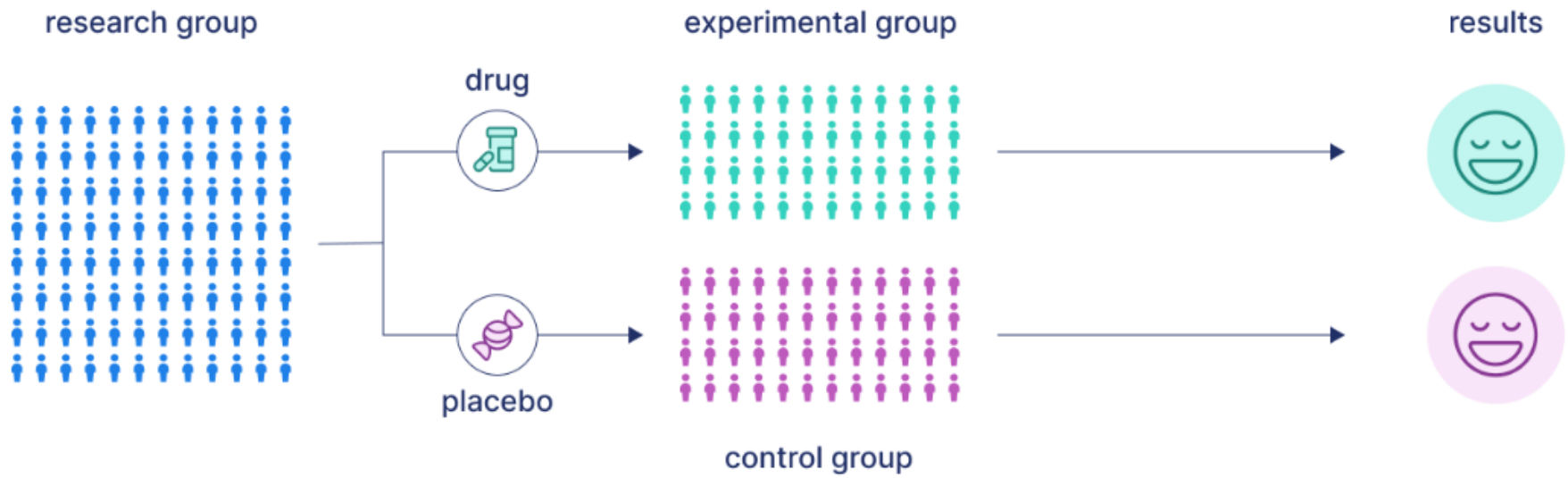
TrialGPT
Matching patients to clinical trials with large language models

Jin, Q., Wang, Z., Floudas, C.S. *et al.*
National Center for Biotechnology Information (NCBI)
University of Illinois at Urbana-Champaign
Published online: 18 November 2024

Clinical Trial

"**A research study** in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) **to evaluate the effects of** those **interventions** on health-related biomedical or behavioral outcomes." – National Institutes of Health (U.S.)



https://www.scribbr.com/research-bias/placebo-effect/

Clinical Trial

"title": "Evaluating the Safety and Efficacy Civamide in Osteoarthritis (OA) of the Knee(s)",

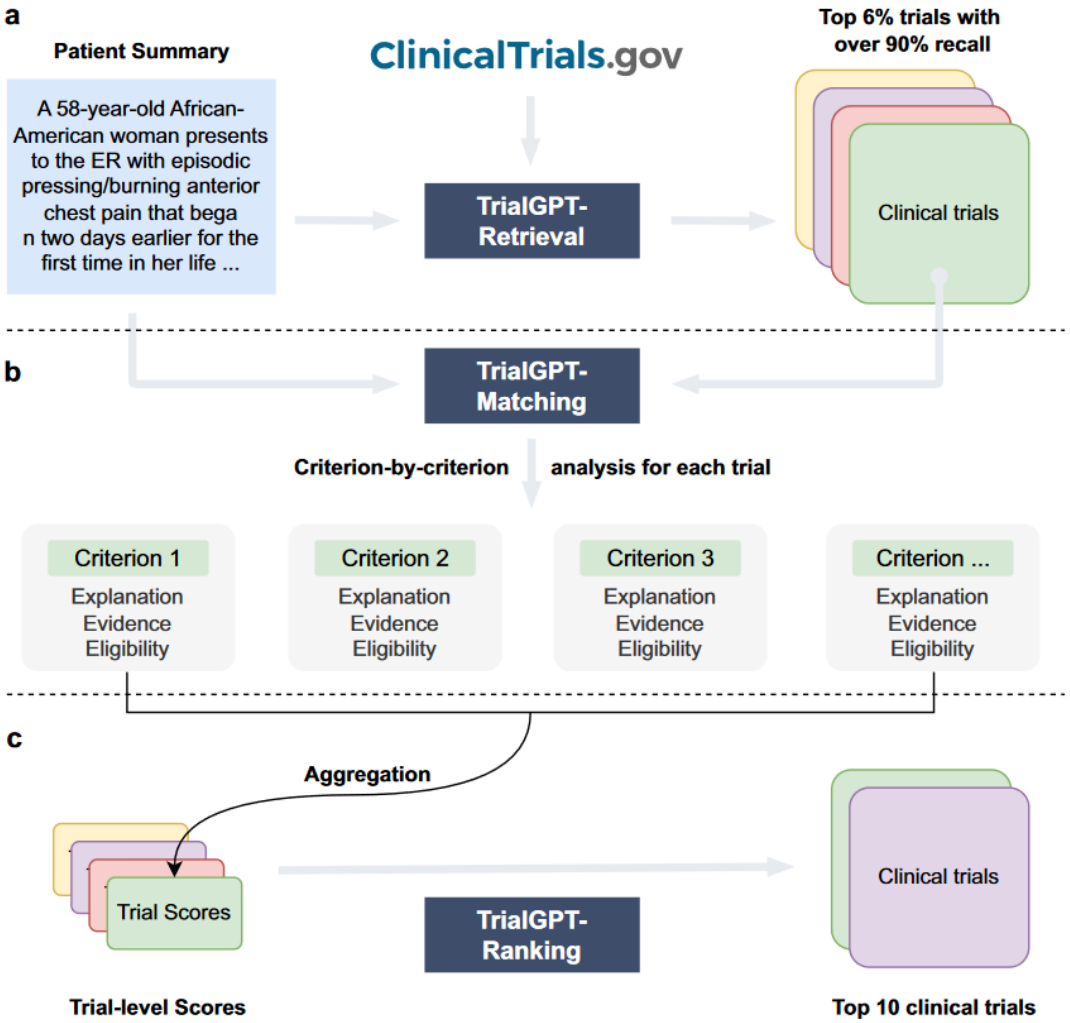
"inclusion_criteria": "inclusion criteria: \n\n Subject voluntarily agrees to participate in this study and signs an IRB-approved informed consent prior to entry into the Screening Period (Day -3). \n\n Subject has OA pain of the Target Knee with a WOMAC Pain Subscale baseline value of > 9 at the Baseline /Randomization Period (Maximum score is 20 for 5 questions with 0 = none; 4 = extreme. \n\n Subject must have a Functional Capacity Classification of I-III. \n\n Subject has taken a stable dose of NSAIDs or COX-2 inhibitor agents for OA pain for at least 22 of the previous 28 days and for each of the 2 days prior to the Screening Period (Day -3) and for at least each of the 2 days prior to the Baseline/Randomization Period (Day 1). Subject must also, agree and be expected to remain on this stable daily dose throughout the study. \n\n Subject is between 40-75 years of age. \n\n Diagnosis of OA is present for at least 6 months according to the ACR criteria for OA of the knee. \n\n Radiographic evidence of OA of the Target Knee (within the last 3 years) with a Kellgren -Lawrence scale of 2 or 3. \n\n Subject is generally in good health. \n\n Subject is expected to be compliant with study procedures. \n\n Females of child -bearing potential must have a negative urine pregnancy test at Screening. \n\n Female subjects of child-bearing potential agree to use an approved form of contraception and must be on the same contraceptive method and dosage schedule during the entire study. \n\n ",

"exclusion_criteria": ": \n\n Presence of tendonitis, bursitis, partial or complete joint replacement of Target Knee. \n\n Presence of active skin disease, erythema, infection, wound or irritation near the treatment area of the Target Knee. \n\n Subject has history of frequent headache or other painful conditions (other than OA) that is expected to require any use of systemic opiates or derivatives, or more than twice a week additional administration of different oral NSAIDs or COX-2 inhibitors (see Section 6.1, Table 2). \n\n Subject experiences regular significant pain due to osteoarthritis or other conditions in the non -target knee or other joints while on stable doses of their current analgesic therapy. \n\n Subject has an anticipated need for any surgical or other invasive procedure that will be performed on the Target Knee or other part of the body during the course of the study. \n\n OA secondary to local joint disorders (e.g ., mechanical disorder, internal derangement of the knee), systemic metabolic disease, endocrine disorders, bony dysplasia, calcium crystal deposition disease , neuropathic arthropathy, frostbite, congenital abnormalities. \n\n Subject has history and/or diagnosis of rheumatoid arthritis, fibromyalgia, connective tissue disease, psoriatic arthritis, erosive inflammatory OA, diffuse idiopathic skeletal hyperostosis, severe neurologic or vascular disease. \n\n Subject has active (redness, swelling, fever, etc.) gout/pseudogout within 6 months prior to screening. \n\n Subject has Type I or Type II diabetes with peripheral neuropathies. \n\n Subject is extremely obese with BMI ≥ 39. \n\n Subject has had trauma to or surgery on the Target Knee within 1 year of Screening/Baseline. \n\n Subject has an underlying clinical condition, including previous malignancies that in the Investigator's judgment, is unstable. \n\n Subject has known allergy or hypersensitivity to capsicum, civamide, or capsaicin-containing products or any constituent of the cream formulation. \n\n Subject has a history of substance abuse within the past 12 months. \n\n Subject has participated in previous clinical study with Civamide Cream. \n\n Use of restricted medications (See Medication/Treatment Table, Section 5.1.2).",

"brief_summary": "To evaluate the safety and efficacy of Civamide Cream 0.075% as a treatment of the signs and symptoms associated with osteoarthritis of the knee."

Image: https://raw.githubusercontent.com/ncbi-nlp/TrialGPT/refs/heads/main/dataset/sigir/corpus.jsonl

TrialGPT: Patient-to-Trial Matching



Data:

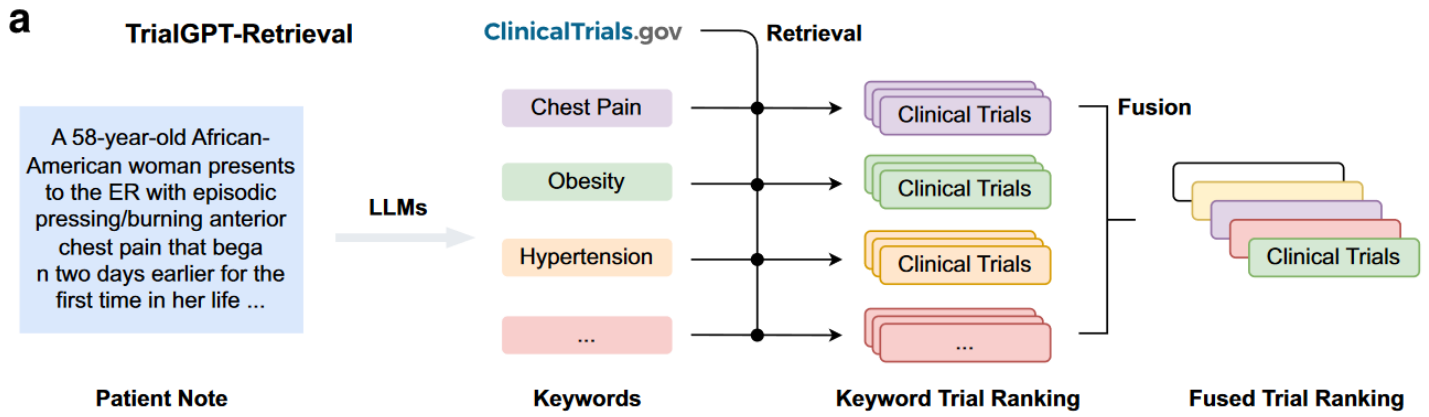
- 183 patient notes
- Over 75,000 trial annotations

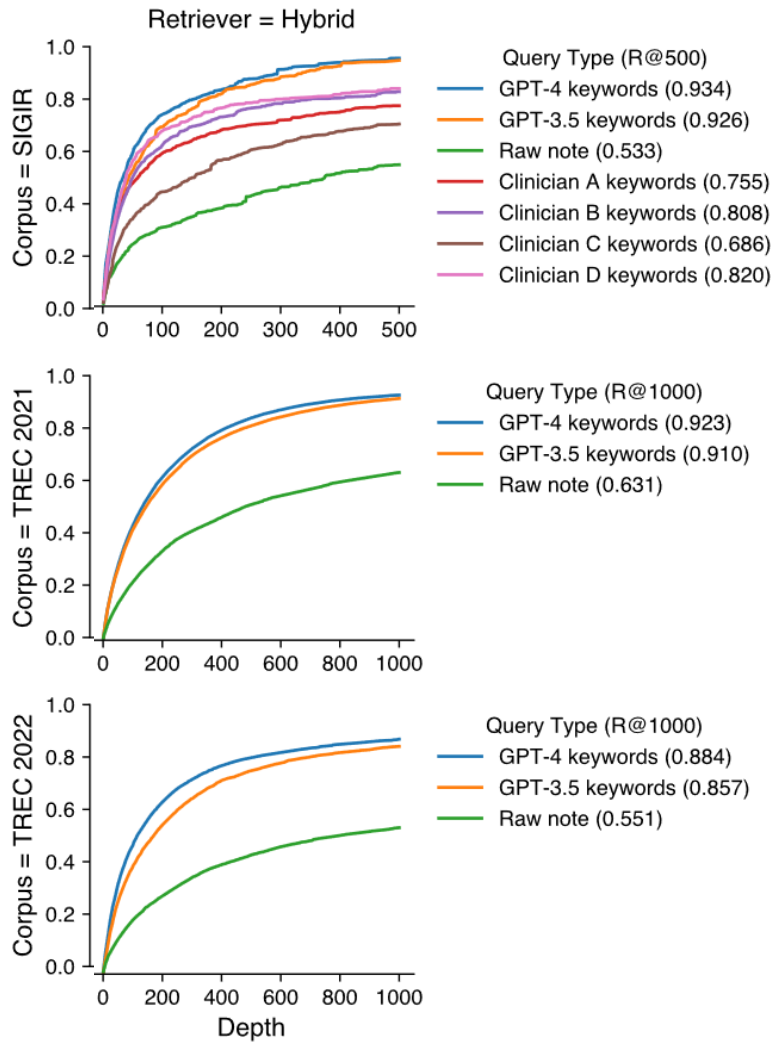
Modules:

- **TrialGPT-Retrieval**: large-scale filtering to retrieve candidate trials.
- **TrialGPT-Matching**:
  - Explanation of the eligibility
  - Locations of relevant sentences in the patient note that are relevant to the target criterion
  - Eligibility classification
- **TrialGPT-Ranking**: generates trial-level scores

Image: Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. "Matching Patients to Clinical Trials with Large Language Models." Nature Communications 15, no. 1 (November 18, 2024): 9074. https://doi.org/10.1038/s41467-024-53081-z.

TrialGPT-Retrieval



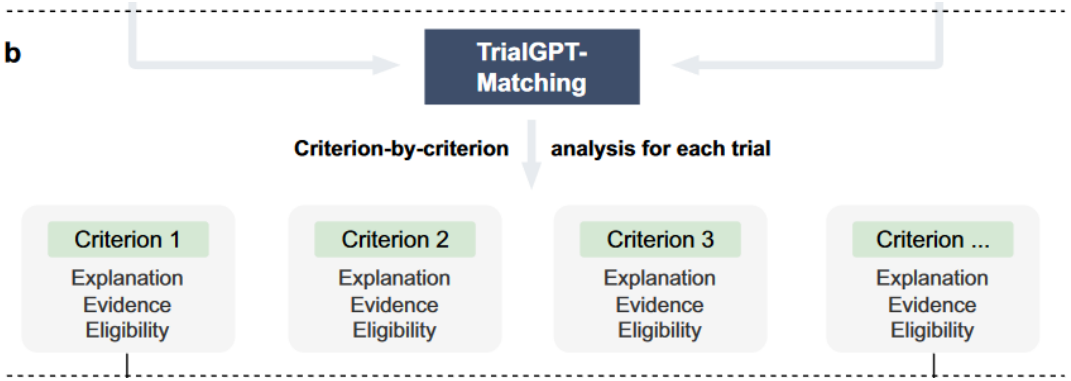


Evaluation Results:

- GPT-4 Retrieves 90% of the relevant clinical trials with only 5.5% of the data (500 Clinical Trials per Pacient).
- LLMs can already generate better keywords than human clinicians for clinical trial retrieval.

Image: Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. “Matching Patients to Clinical Trials with Large Language Models.” Nature Communications 15, no. 1 (November 18, 2024): 9074. https://doi.org/10.1038/s41467-024-53081-z.

TrialGPT-Matching



Evaluation Data:

- Manually annotated 1015 patient-criterion from 53 patients (SIGIR)
  - Correctness of explanations
  - Relevant sentence locations patients notes
  - Patient's eligibility prediction

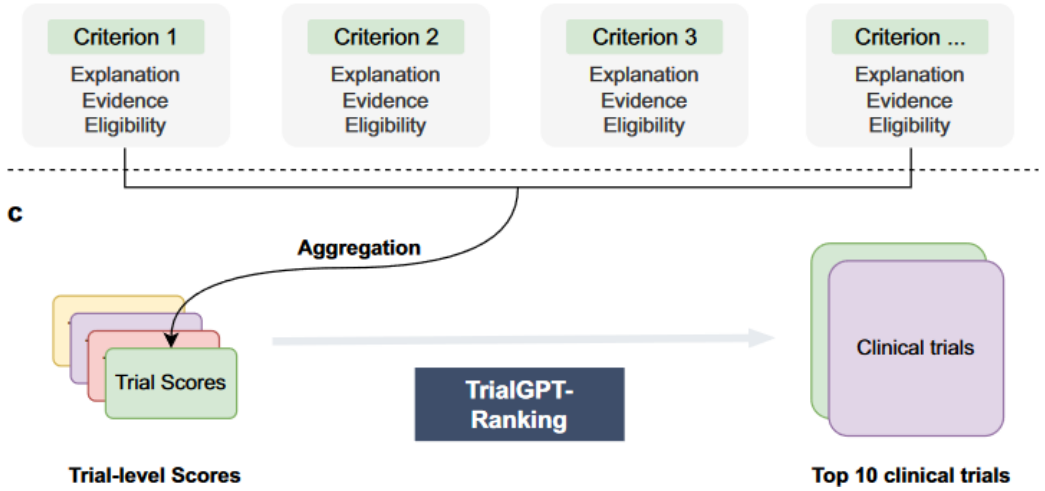
Relevance Explanation: TrialGPT-Matching provides **88% correct relevance explanations**, with errors mainly in criteria requiring implicit reasoning.

Sentence Location: It identifies relevant sentences with **89% F1, matching human expert performance.**

Eligibility Prediction: Achieves **87% accuracy, close to experts**, but struggles with ambiguous exclusion labels, highlighting areas for improvement.

Image: Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. "Matching Patients to Clinical Trials with Large Language Models." Nature Communications 15, no. 1 (November 18, 2024): 9074. https://doi.org/10.1038/s41467-024-53081-z.

TrialGPT-Ranking



Evaluation Data:

- Again, the previous 1015 patient-trial

Metrics:

- Trial level metric
- Linear Aggregation (%met inclusion/exclusion)
- LLM Aggregation
- Feature combination

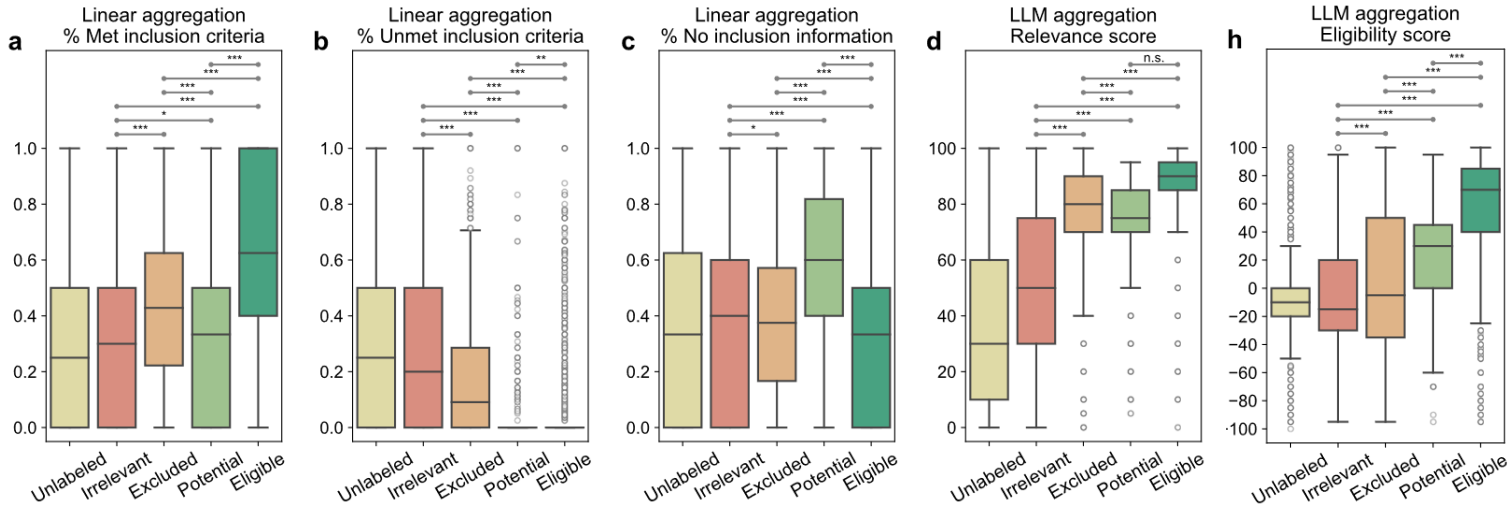


Image: Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. “Matching Patients to Clinical Trials with Large Language Models.” Nature Communications 15, no. 1 (November 18, 2024): 9074. https://doi.org/10.1038/s41467-024-53081-z.

Results

Table 2 | Performance of different methods for ranking and excluding clinical trials

Application			Ranking		Excluding	Overall
Method / Metric			NDCG@10	P@10	AUROC	Average
SciFive[46] (encoder-decoder)		Further trained on MedNLI[37]	0.4271	0.3787	0.5895	0.4652
BioBERT[47] (dual-encoder)		Further trained[48] on MNLI[49], SNLI[50], SciNLI[51], SciTail[52], MedNLI[37], and STSB[53]	0.4091	0.3746	0.5952	0.4596
PubMedBERT[54] (dual-encoder)			0.4327	0.3874	0.5976	0.4726
SapBERT[55] (dual-encoder)			0.4148	0.3738	0.5933	0.4606
BioLinkBERT[36] (cross-encoder)		Further trained on MedNLI[37]	0.4797	0.4281	0.6176	0.5085
TrialGPT-Ranking (GPT-3.5)		Feature combination	0.5395	0.5115	0.6582	0.5697
TrialGPT-Ranking (GPT-4)	Linear aggr.	Sign(% Included)	0.6097	0.5653	0.6765	0.6172
		Sign(% Not included)	0.4923	0.4556	0.6652	0.5377
		Sign(% Excluded)	0.3930	0.3715	0.6477	0.4707
		Sign(% Not excluded)	0.4130	0.3886	0.5820	0.4612
	LLM aggr.	Sign(Relevance)	0.7281	0.6700	0.7402	0.7128
		Sign(Eligibility)	0.7252	0.6724	0.7895	0.7290
	Feature combination		0.7275	0.6688	0.7979	0.7314

The Sign() function assigns suitable signs for the corresponding task, e.g., for "% Included", it will be "+" for ranking and "-" for excluding clinical trials. Aggr.: aggregation. NDCG@10: normalized discounted cumulative gain at 10. P@10: precision at 10.
AUROC the area under the receiver operating characteristic curve.

Conclusions

TrialGPT Framework Effectiveness:

- TrialGPT is a novel architecture using GPT-4 and GPT-3.5 for patient-to-trial matching, consisting of three components: Retrieval, Matching, and Ranking.
- It demonstrates high accuracy and efficiency
  - Recalls over 90% of relevant trials
  - Achieves 87.3% accuracy in criterion-level predictions
  - Outperforms baselines by 43.8% in trial ranking and exclusion tasks.

Explainability and User Assistance:

- Provides natural language explanations.
- Highlights relevant patient data to provide evidences.
- Supports human decision-making and improves trust and usability.

Future work

Drawbacks

- Performance difference between GPT-4 and GPT-3.5 and long input prompts which can mean high economic cost.
- Subjective evaluation system dependent on the clinicians.
- Multimodal input (patient note + radiography + lab results).

References

[1] Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. “Matching Patients to Clinical Trials with Large Language Models.” Nature Communications 15, no. 1 (November 18, 2024): 9074. https://doi.org/10.1038/s41467-024-53081-z.

[2] Cuconasu, Florin, Giovanni Trappolini, Nicola Tonellotto, and Fabrizio Silvestri. “A Tale of Trust and Accuracy: Base vs. Instruct LLMs in RAG Systems.” arXiv, June 21, 2024. https://doi.org/10.48550/arXiv.2406.14972.

[3] Khaliq, M. Abdul, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić. “RAGAR, Your Falsehood RADAR: RAG-Augmented Reasoning for Political Fact-Checking Using Multimodal Large Language Models.” arXiv, April 18, 2024. https://doi.org/10.48550/arXiv.2404.12065.

[4] Koopman, B., & Zuccon, G. (2016). A Test Collection for Matching Patients to Clinical Trials. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.

[5] Jin, Q., Tan, C., Zhao, Z., Yuan, Z. & Huang, S. Alibaba DAMO Academy at TREC Clinical Trials 2021: Exploring Embedding-based First-stage Retrieval with TrialMatcher. In Proc. Thirtieth Text REtrieval Conference (TREC 2021) (2021).

[6] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S. & Hersh, W. R. Overview of the TREC 2022 Clinical Trials Track. In Proc. Thirty-First Text REtrieval Conference (TREC 2022) (2022).