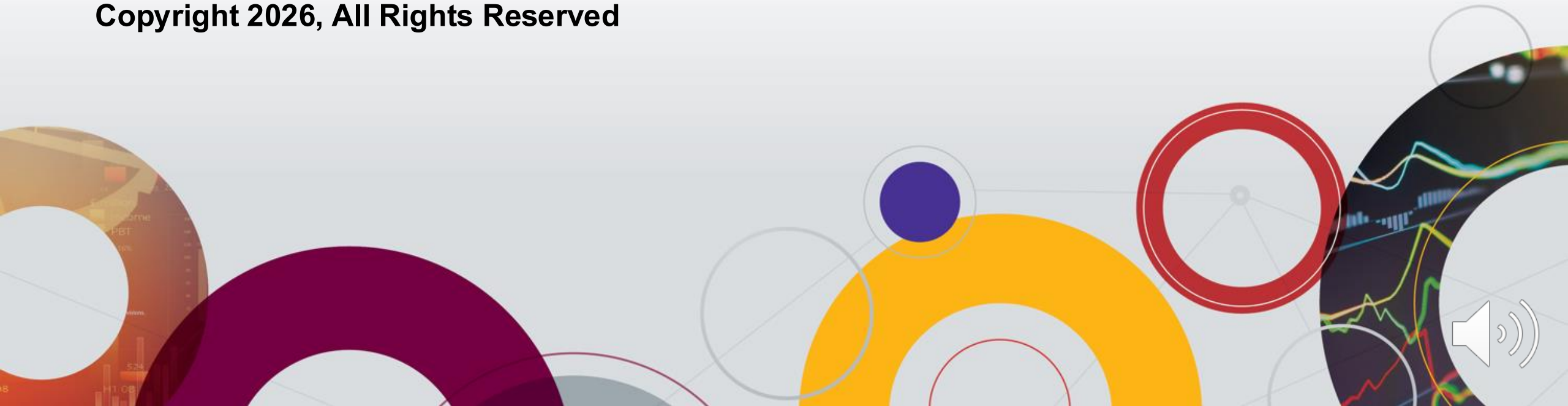


5.4b Generative Artificial Intelligence

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Answering clinical questions

Over 80% correct on general questions (Rydzewski, 2024) and about guidelines for cancer (Ferber, 2024)

Answering of questions developed by physicians – 5.5/6 for accuracy and 3/6 completeness (Goodman, 2023)

Answering questions using RAG – mixed results

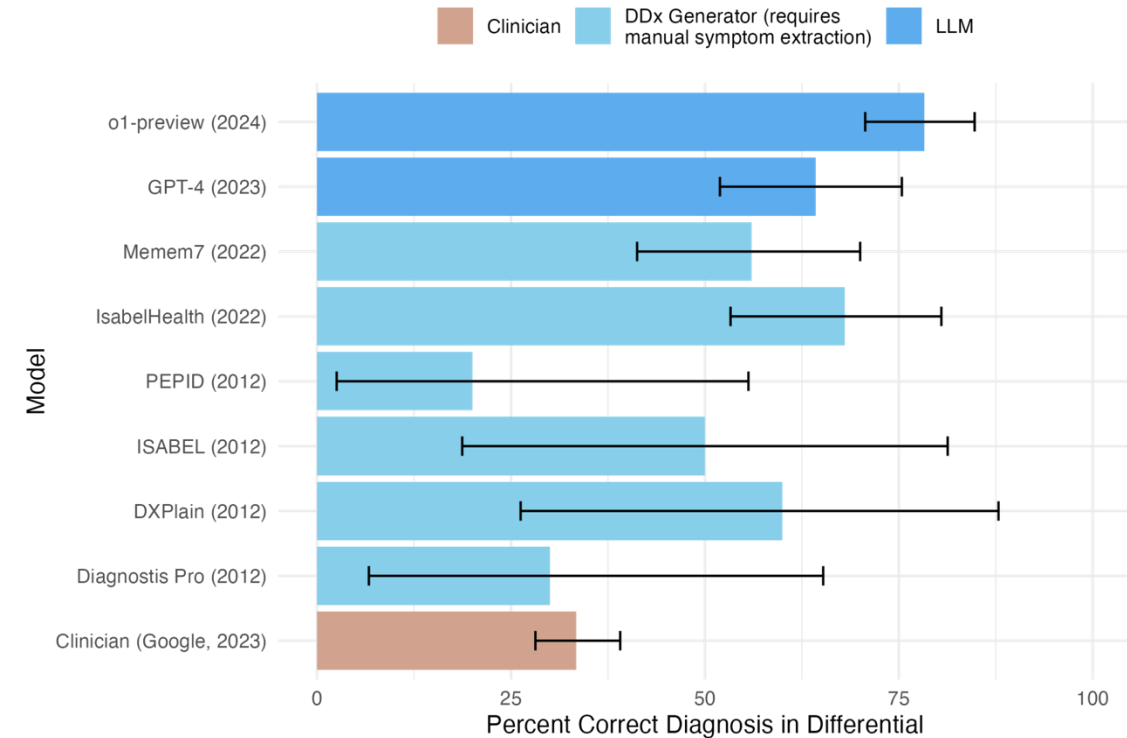
- LLM framework augmented with retrieval capabilities from curated medical resources showed significant improvement in performance compared with standard LLMs in factuality, completeness, user preference, and adversarial safety (Zakka, 2024)
- On questions seeking additional clinical evidence, ChatRWD using retrieval-augmented generation outperformed OpenEvidence and several general LLMs (Low, 2024)
- RAG often degraded performance, although could be mitigated by evidence filtering and query reformulation (Kim, 2025)

Solving clinical cases

Early work found comparable performance to human clinicians for generating differential diagnoses from well-known resources developed for other purposes

- New England Journal of Medicine (NEJM) clinicopathological cases (CPCs) (Kanjee, 2023; Eriksen, 2023; Buckley, 2025)
- Merck Manual vignettes (Rao, 2023)
- Mayo symptom checker (Chen, 2023)
- PMC-Patients (Bhasuran, 2025)

“Superhuman performance” (?) with OpenAI o1-preview LLM scoring better than previous results with (Brodeur, 2024)



Clinical vignettes

Diagnostic reasoning (Goh, 2024)

- Based on 6 unpublished diagnostic cases developed to assess clinical decision support (Berner, 1994)
- 50 physicians randomized to LLM or conventional information resources and assessed with diagnostic performance rubric
- No statistical difference between physicians using LLM (76%) vs. conventional (74%) resources; LLM alone scored better than either (92%)

Management reasoning (Goh, 2025)

- Based on 5 cases adapted from *Grey Matters* podcast from American College of Physicians
- 92 physicians randomized to LLM or conventional information resources and assessed with management performance rubric
- Physicians using LLM scored better than physicians using conventional resources and no different from LLM alone

Solving clinical cases (cont.)

Comparing generalist (ChatGPT and Gemini) vs. specialist (OpenEvidence, UpToDate) models found better performance for generalist models (Vishwanath, 2025) on MedQA and HealthBench benchmarks (Arora, 2025)

From foundation model built with Epic Cosmos database of 300 million unique patient records from 310 health systems performed better than specific models on 78 real-world tasks, including diagnosis prediction, disease prognosis, and healthcare operations (Waxler, 2025)

LLM-based consult system made fewer errors than previous rule-based system in Kenyan primary care clinics (Korom, 2025)

Clinical vignettes – LLMs not always better

Family medicine physicians scored better than LLMs on cases from Swedish family medicine specialist examination (Arvidsson, 2024)

Perform worse in

- “Conversational” than “examination-based” assessments (Johri, 2025)
- Script-based concordance testing (McCoy, 2025)

May cause harm

- For clinical scenarios with safety or ethical dilemmas, LLMs may generate clinically harmful decisions up to 10% of time (Omar, 2025)
- Across many specialties, LLMs may cause harm up to 22% of time (Wu, 2025)
 - Numerous Options Harm Assessment for Risk in Medicine (NOHARM) dataset measures completeness, safety, escalation, and restraint from LLM-generated medical recommendations

Using different drug names (generic vs. trade) led to differing performance on standard datasets (Gallifant, 2024)

DxPlain performed better than LLMs at differential diagnoses (Feldman, 2025)

Use of imaging models for imaging cases

Combining LLM with radiology reports and an image encoder enabled zero-shot detection of five CXR findings – atelectasis, cardiomegaly, consolidation, pleural effusion, and pulmonary edema (Xu, 2023)

Imaging case vignettes with MCQs from *JAMA Clinical Challenges* and *NEJM Image Challenges*

- GPT-4V without fine-tuning outperformed Gemini Pro, ChatGPT, and others (Han, 2024)
- GPT-4V performed comparable to humans but could present flawed rationales (Jin, 2024)

In live clinical care, LLM associated with 15.5% documentation efficiency improvement and no change in radiologist-evaluated clinical accuracy or textual quality of reports (Huang, 2025)

On annotated cases, LLM achieved accuracy comparable to clinicians in ordering imaging studies based on American College of Radiology Appropriateness Criteria (Yao, 2025)

Predictive tasks

Cardiovascular disease (Han, 2024)

ED acuity (Williams, 2024) and predicting admissions (Glicksberg, 2024)

Rare disease diagnosis (Zelin, 2024)

Diagnose specific infections, autoimmune disorders, vaccine responses, and disease severity differences based on T and B cell receptor sequences (Zaslavsky, 2025)

Transformer model predicted cardiovascular risk based on breast arterial calcification on screening mammography (Dapamede, 2025)

Clinical note generation

Summarization

Assisting documentation

Summarization of EHR data

Extracting details from discharge summaries with 81% accuracy (Ellershaw, 2024)

For radiology reports, patient questions, progress notes, and doctor-patient dialogue, LLM summaries found preferable to human summaries (Van Veen, 2024)

LLM-generated emergency medicine (EM)-to-inpatient physician (IP) handoff notes determined superior compared with physician-written summaries but marginally inferior in usefulness and safety (Hartman, 2024)

LLM answering questions from clinical notes in 3 languages in high agreement with humans (Menezes, 2025)

LLM-generated patient summaries slightly reduced encounter time, with most benefit for those spending most time in reviews (Silberlust, 2025)

Assisting documentation

De-identification

- Radiology reports (Chambon, 2023)
- Discharge summaries (Altalla', 2025)

Identifying social determinants of health (Guevara, 2024; Consoli, 2025)

Extraction of Severe Sepsis and Septic Shock Management Bundle (SEP-1) quality measure (Boussina, 2024)

Mixed performance in some tasks in ICD-10-CM and CPT-4 coding (Soroush, 2024; Klang, 2025)

Patient communication and education

Answering patient questions

Summarizing information

Drafting replies to communications

Other patient-related tasks

Answering clinical questions – patients

ChatGPT-3.5 answered 21 of 25 questions about cardiovascular disease prevention deemed acceptable by cardiology clinicians for patient-facing information platform and as AI-generated draft responses to questions sent by patients (Sarraju, 2023)

ChatGPT-3.5 provided evidence-based answers to public health questions (Ayers, 2023)

ChatGPT-4 responses to patient questions posted to public social media forum rated higher quality and more empathetic (Ayers, 2023)

For 150 consumer health-related questions, search engines answered 50-70% correctly, LLMs about 80%, and RAG up to 30% higher (Fernández-Pichel, 2025)

Summarization – for patients

ChatGPT-3.5 asked to generate simplified radiology reports found to be factually correct, complete, and not potentially harmful to patient but with instances of incorrect statements, missed relevant medical information, and potentially harmful passages (Jeblick, 2023)

Lay language summaries for research studies found more accessible and transparent (Shyr, 2024)

Transforming clinical notes for patients rated patient-friendly but 44% not entirely complete and 18% found safety concerns for incomplete or inaccurate information (Zaretsky, 2024)

Generating patient-friendly summaries of radiology reports understandable to patients (Park, 2024)

Drafting replies to patients

ChatGPT-3.5 wrote patient clinic letters with high level of correctness and measure of “humanness” (Ali, 2023)

Fine-tuned model generated patient portal messages deemed positive for responsiveness, empathy, and accuracy and neutral for usefulness (Liu, 2024)

Pilot study of clinical usage of draft letters found about 20% utilization for task, with significant reductions in burden and burnout score derivatives but no change in time taken (Garcia, 2024)

For drafting replies, LLM not associated with reduced time on writing a reply but was associated with longer read time, longer replies, and perceived value in making a more compassionate reply (Tai-Seale, 2024)

Help patients draft messages to health system (Liu, 2024)

Drafting replies (cont.)

GPT-4-generated patient portal message responses achieve comparable levels of empathy, relevance, and readability to those found in typical responses according to providers (Kaur, 2024)

Patient Advice Message Chatbot drafted reply to incoming messages from patient portal in 9 clinics for nurses, medical assistants (MAs), and clinicians (physicians and advance practice clinicians [APCs]) (English, 2024)

Satisfaction with AI-generated responses to medical questions in EHR (Kim, 2024; Cavalier, 2025) but reduced when patient notified response was AI-generated (Cavalier, 2025)

Errors in messages reviewed by primary care physicians were missed by 13-15 of 20 physicians and 35-45% submitted unedited, despite high enthusiasm for process (Biro, 2025)

Medical research assistance – researchers

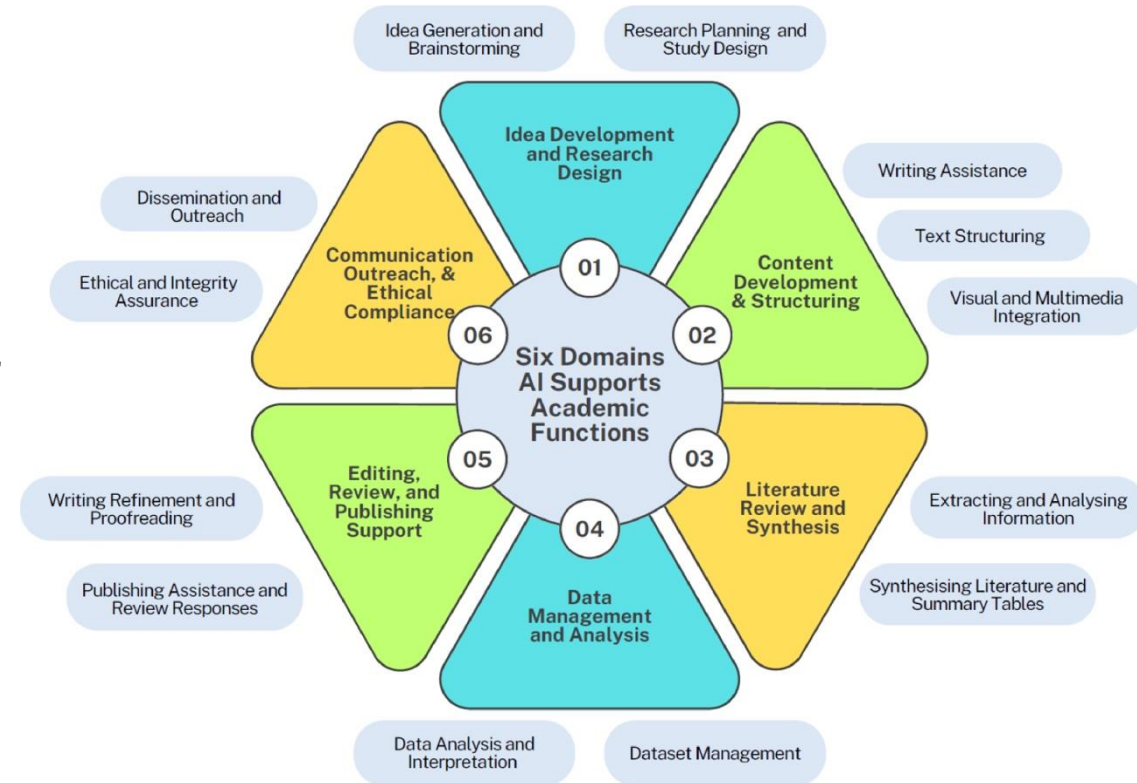
Potentially helpful at many steps of research process

- Systematic review of uses (Khalifa, 2024)
- Practices and perceptions (Andersen, 2025)

Concerns about non-disclosure to outright fraud not only in writing but also conduct of science, peer review, etc. (Kwon, 2025)

Recommendations from ICJME (2025)

- Disclosure
- Responsibility
- Non-authorship



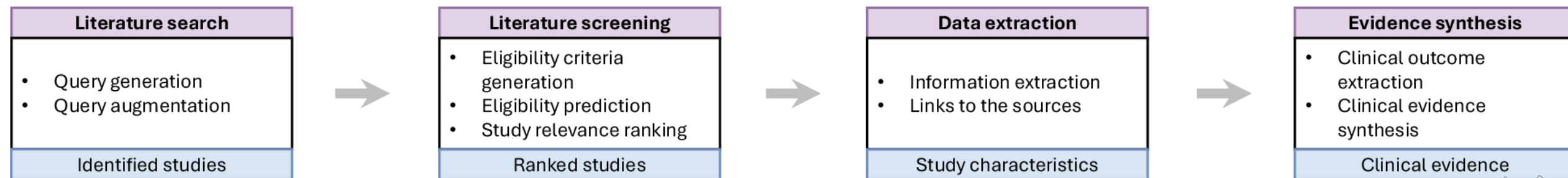
Facilitating systematic reviews

LLM potential in all steps of conducting systematic reviews (Wang, 2025)

Enhancing citation screening with ensemble models (Zhang, 2025) or RAG and prompt engineering (Trad, 2025)

Reduced work for 12 Cochrane reviews from 12 work-years to two days, with discovery of new findings (Cao, 2025)

Systematic reviews of usage – promising but not ready for prime time (Lieberum, 2025; Scherbakov, 2025)



Medical research assistance – readers

GPT-4 feedback on scientific papers (Liang, 2024)

- For PDFs of papers, found to have overlap comparable to between humans; higher for poorer-quality papers
- Over half (57.4%) of authors found generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers

Summaries of 140 evidence-based journal abstracts generated by ChatGPT 70% shorter than mean abstract length and found to have high quality, high accuracy, and low bias (Hake, 2024)

PaperQA2 summarized topics comparable to Wikipedia and identified contradictions in papers (Skarlinski, 2024)

Administration and workflow

Systematic review of LLMs in clinical workflows (Artsi, 2025)

- Applications – outpatient communication, mental health support, inbox message drafting, and clinical data extraction
- Associated with improvements in operational efficiency, user satisfaction, and reduced workload
- Challenges included performance variability across data types, limitations in generalizability, regulatory delays, and lack of post-deployment monitoring

Use of agents to manage tasks

- Dataset for agents used in interaction with EHR tasks (Jiang, 2025)
- ChatEHR to navigate within Epic EHR (McElfresh, 2025)
- Personal health agents (Heydari, 2025)

Administration and workflow (cont.)

Verifying facts in EHR (Chung, 2025)

Recognizing nursing events in standard nursing terminologies from spoken dialogue in home-care visits (Beerepoot, 2025; Zhang, 2025)

LLM-enabled classification of patient self-reported symptoms and needs (Naved, 2025)

Pediatric medication dosage errors (Levin, 2025)

Systematic and scoping reviews of generative AI

Healthcare generally (Bedi, 2024; Zhang, 2025; Shool, 2025)

Conversational LLMs (chatbots) in healthcare (Wang, 2024), mental health (Hua, 2025), and their empathy (Howcroft, 2025)

Primary care (Iannone, 2025; Katonai, 2025)

Clinical documentation (Woo, 2025)

Agents (Gorenshtein, 2025), including multimodal agents (Yu, 2025)

Radiology (Al Zaabi, 2025)

Bias and equity (Liao, 2025); clinical fairness (Liu, 2025)

Economic aspects (von Gerich, 2026)

Most conclude much promise for LLMs but studies narrow in scope and many gaps in literature

Beyond healthcare – work productivity

Assignment of occupation-specific, incentivized writing tasks to 453 college-educated professionals found 40% decreased time and 18% improved quality for half using ChatGPT (Noy, 2023)

At global management consulting firm, consultants randomized to using ChatGPT-4 were (Dell'Acqua, 2023)

- Significantly more productive – completed 12.2% more tasks on average, and completed task 25.1% more quickly)
- Produced significantly higher quality results – more than 40% higher quality compared to control group
- Noted to be part of “jagged technological frontier” where some tasks easily done by AI and others not, such as combining qualitative and quantitative data

Predicting company earnings (Kim, 2024; Shaffer, 2024)

Software engineering – mixed results

- Improved productivity (Cui, 2024)
- Developers took 19% more time though believed they were more productive (Becker, 2025)

Toward artificial general intelligence?

GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology, and more, without needing any special prompting (Bubeck, 2023)

- Shows “sparks of artificial general intelligence”
- “Strikingly close to human-level performance” that often vastly surpasses prior models such as ChatGPT-3.5

GPT-4 performed worse than humans in abstraction and analogy (Moskvichev, 2023) and on abstraction and reasoning corpus (ARC; Chollet, 2019) (Mitchell, 2023) but newer o3 has “[solved](#)” problem (Mitchell, 2024)