PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Nay JJ, Karamardian D, Lawsky SB, Tao W, Bhat M, Jain R, Lee AT, Choi JH, Kasai J. 2024 Large language models as tax attorneys: a case study in legal capabilities emergence. *Phil. Trans. R. Soc. A* **382**: 20230159. https://doi.org/10.1098/rsta.2023.0159

Received: 1 June 2023
Accepted: 4 October 2023

One contribution of 15 to a theme issue 'A complexity science approach to law and governance'.

Subject Areas:
artificial intelligence

Keywords:
artificial intelligence, large language models, machine learning, computational law, law-informed AI, law informs code

Author for correspondence:
John J. Nay
e-mail: john.j.nay@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7031271.

THE ROYAL SOCIETY
PUBLISHING

Large language models as tax attorneys: a case study in legal capabilities emergence

John J. Nay[1], David Karamardian[2], Sarah B. Lawsky[3], Wenting Tao[2], Meghana Bhat[4], Raghav Jain[5], Aaron Travis Lee[6], Jonathan H. Choi[7] and Jungo Kasai[8]

[1]CodeX, Center for Legal Informatics, Stanford University, Stanford, CA, USA
[2]Stanford University, Stanford, CA, USA
[3]Northwestern Pritzker School of Law, Chicago, IL, USA
[4]Engineering, University of Michigan, Ann Arbor, MI, USA
[5]SimPPL, India
[6]Independent, Northern Ireland
[7]School of Law, University of Southern California, Los Angeles, CA, USA
[8]Department of Computer Science, University of Washington, Seattle, WA, USA

JJN, 0009-0006-5749-8518

Better understanding of Large Language Models' (LLMs) legal analysis abilities can contribute to improving the efficiency of legal services, governing artificial intelligence and leveraging LLMs to identify inconsistencies in law. This paper explores LLM capabilities in applying tax law. We choose this area of law because it has a structure that allows us to set up automated validation pipelines across thousands of examples, requires logical reasoning and maths skills, and enables us to test LLM capabilities in a manner relevant to real-world economic lives of citizens and companies. Our experiments demonstrate emerging legal understanding capabilities, with improved performance in each subsequent OpenAI model release. We experiment with retrieving and using the relevant legal authority to assess the impact of providing additional legal context to LLMs. Few-shot prompting, presenting examples of question–answer pairs, is also found to significantly enhance

© 2024 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

the performance of the most advanced model, GPT-4. The findings indicate that LLMs, particularly when combined with prompting enhancements and the correct legal texts, can perform at high levels of accuracy but not yet at expert tax lawyer levels. As LLMs continue to advance, their ability to reason about law autonomously could have significant implications for the legal profession and AI governance.

This article is part of the theme issue 'A complexity science approach to law and governance'.

1. Introduction

AI capabilities are marching forward [1–5]. Large Language Models (LLMs) [6] are the locus of the rapid advances. State-of-the-art LLMs are able to pass standardized tests [7] and plan, reason and leverage tools [8]. LLMs, though, are essentially black boxes, even to their developers. We have little insight into their inner workings and have no guarantees on how an LLM will behave on a new task [9–11]. Best practice is to measure LLM performance on a litany of benchmarks before models are deployed beyond the research environment, but these benchmarks are often not real-world tasks we care about, or may have been memorized by the LLM during its training [12]. This phenomenon typically arises when the datasets used for training LLMs, often sourced from the internet, contain the same data used for performance evaluation. The overlap can inflate the estimate of the model's performance, giving an illusion of understanding that could instead be basic recognition [13].

We focus evaluation effort specifically on legal analysis capabilities of LLMs for three reasons.

First, assessing the extent that LLMs grasp the law can contribute toward governing LLMs and automated systems more generally.[1] One policy-relevant approach seeks to leverage regulatory reasoning and legal reasoning within LLMs for 'Law-Informed AI' aligned with societal values as determined by democratic processes and law-making. This 'Law Informs Code' approach rests on the established effectiveness of the democratic process in creating adaptive legal standards such as fiduciary duties through iterative debate and litigation [14]. The premise is that learning the spirit of the law can guide AI systems in making reasonable choices in novel scenarios. For instance, LLMs exhibit an early ability to predict when fiduciary duties are violated [15], and this capability could power safer AI deployments where an LLM-powered system serves a human principal.

Second, LLMs can be used as tools for humans to more efficiently and effectively provide legal services, whether that be self-service or through a professional attorney. If the models better understand law, they can be more reliable and ultimately more useful. LLMs might potentially assist in tasks ranging from contract analysis to case prediction, potentially democratizing access to legal advice, reducing the cost and complexity for those who might otherwise struggle to navigate the legal system. Rigorous safeguards should be put in place as these models are deployed, given the sensitive nature of legal work. This includes increasing data privacy, minimizing bias, maintaining accountability for the decisions made with the help of these models, and evaluating the suitability of the LLMs for any given use case. Hence, the need for systematic evaluations.

Third, if LLMs understand the law well enough, they could be deployed by the government, citizens and researchers to identify inconsistencies in existing laws [16]. LLMs could increase the efficiency and transparency of governments more broadly. For instance, LLMs can oftentimes provide clear, understandable explanations of complex laws and regulations. Eventually, LLMs

---

[1] Some examples in the USA and UK include National Artificial Intelligence Initiative, About: National Artificial Intelligence Initiative, https://www.ai.gov/about/#naii-national-artificial-intelligence initiative. [Accessed 11 April 2023]; Advancing Trustworthy AI, National Artificial Intelligence Initiative, https://www.ai.gov/strategic pillars/advancing-trustworthy-ai. [Accessed 11 April 2023]; European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. (Accessed 11 April 2023).

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

may help predict likely impacts of new laws or policies. By scanning vast amounts of legal text and associated implementations, LLMs could flag potentially 'outdated' law, or areas where the law is silent when, in other similar circumstances, the legislature or regulators provide guidance.

In this paper, we study retrieval-augmented generation of LLMs leveraging the text of the US Code of Federal Regulations (CFR) and the US Code (a compilation of federal statutes). We test the emerging capabilities of a suite of LLMs in understanding tax law.

We chose tax law for four reasons. First, unlike some legal subjects where the doctrines are distilled from numerous precedents, the legal authority in tax law is principally concentrated in two sources: the Treasury Regulations under the CFR and Title 26 of the US Code (also called the Internal Revenue Code). This allows us to use a fixed universe of potentially relevant documents for the LLM's retrieval augmentation. Second, many tax rules allow for definitive answers to inquiries. This allows us to set up consistent and automated validation pipelines. Third, answering tax law questions for a given scenario usually requires logical reasoning skills and even maths skills beyond just reading the relevant legal authority, enabling us to test LLM capabilities in a manner relevant to real-world practice. Fourth, tax law is highly significant to the economic lives of nearly every citizen and company on a regular basis.

We assess the accuracy of responses generated by LLMs on thousands of tax law inquiries across experimental set-ups: the use of the LLM alone and the integration of the LLM with the underlying legal texts, along with various retrieval techniques (with comparisons made across different retrieval methods). We conduct these experiments across LLMs, from smaller and weaker models, up through the largest state-of-the-art model, OpenAI's GPT-4. Each LLM we tested was state-of-the-art when it was originally released. Through examining results across increasingly large models, we find evidence for emerging legal understanding capabilities of LLMs, improving with each model release. This suggests that we may see the advent of superhuman AI legal skills as the state of the art continues to rapidly advance.

2. Our approach to validating LLM legal understanding

We test LLMs' ability to 'understand' law. Giving relevant and correct legal advice for a specific situation is a task combining precise knowledge of legal sources as well as reasoning (and sometimes maths) capabilities to analyse situations.

We generate multiple-choice problems, each consisting of a question and a set of potential answers (only one of which is correct). The facts, names and numbers for each problem are randomly generated by Python code. As a result, our synthetic validation set consists solely of brand-new problems that do not exist on the internet and could not have been included in the training set for any LLM. This is an important distinction setting our validation apart from many other LLM benchmarking exercises. In many cases, the LLM being evaluated—which is trained on much of the internet—may have been trained on the validation data itself.

While some of the questions involve only qualitative understanding (e.g. 'checking-the-box' questions, see electronic supplementary material, appendix for more), others involve more arithmetic operations as well (e.g. calculation to determine basis amount). The solution to a given problem can refer to either the CFR or the US Code. To produce problems of a specific question type, we call our Python code to generate a bulk collection of multiple-choice problems. Each generated problem contains: (i) the legal question with answer options for the LLM to choose from; (ii) the correct option; (iii) the explanation for why that option is correct and (iv) the citation to the specific part of the law in which the answer to the question can be determined.

We generate two overarching multiple-choice exams for evaluation, one based on the CFR, and one based on the US Code. The CFR and US Code exams are composed of three and four sections, respectively, with each 100-question section pertaining to a specific tax law question type. See the electronic supplementary material, appendix for details on these seven types of questions across tax law categories.

For each question, we prompt an LLM to pick one of the multiple-choice answers, and we evaluate the LLM's performance based on whether it chooses the correct answer. Since manually

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

grading over 28 000 questions across all experiments by hand is not feasible, and since the models do not always produce outputs in a consistent format that we can parse consistently and directly compare to the real answer, we use the most powerful available LLM, GPT-4, to carry out the bulk of the final simple step of the evaluation. GPT-4 is instructed to grade the accuracy of a predicted answer choice by comparing it to the real answer choice for a given question.[2]

3. Our approach to retrieval-augmented generation and LLM prompting

We compare results across retrieval methods, each with its own prompt template that provides different supporting context to the LLM; see the electronic supplementary material, appendix for a full example of a prompt template from one of our experimental runs. When supplying supporting context to models with smaller context windows, we sometimes had to truncate the retrieved context to fit inside the window. The LLMs have the following context windows: davinci, 2049 tokens; text-davinci-002 and gpt-3.5-turbo, 4097 tokens; gpt-4, 8192 tokens.

Our first experimental setting for retrieval, 'bypass_retrieval,' creates a baseline for testing the impact of retrieval and LLM knowledge. In these cases, we simply provide the LLM with a multiple-choice question and the answer options with no additional explicitly provided legal context. This method assesses the ability of an LLM to answer a tax law question solely from its 'internal knowledge base'.

For the second retrieval experimental setting, 'similarity_search,' we inject potentially relevant legal text into the prompt. Offline, before running the evaluations, we extracted the statutes from Title 26 of the US Code and regulations from the CFR Treasury Regulations in the form of discrete documents, each corresponding to a subsection from the legal source. The discrete subsection documents are roughly 130 tokens on average for our CFR vector database, and roughly 250 tokens on average for our US Code vector database. We leverage the open-source 'Facebook AI Similarity Search' library to create a vector database that maps the discrete subsections to 768-dimensional embeddings, computed by a state-of-the-art [17] dense retrieval model, GTR-large [18]. GTR-large is trained on large amounts of retrieval data from various domains, including biomedical and science text, but not legal text; thus, our retrieval use-case is 'out-of-domain' for the embeddings model.[3] When a question is presented as input, our system retrieves the four most 'relevant' documents from the vector store, where relevance is estimated based on the cosine similarity between the documents and the question. These documents are then injected as context into the prompt together with the original question, and the LLM is also instructed to return the metadata for which legal source subsections it referenced in its answer.

The third experimental setting, 'gold_truth,' does not rely on a vector database or similarity search to provide the LLM with additional context; instead, we directly provide as context the correct source material, obtained by referencing each given question's citation to the specific part of law it pertains to. Incorporating this method in the experimental design helps estimate the impact of the theoretically best possible retrieval. Another way of looking at this design is that it isolates errors in the LLM's reasoning caused by inaccurate retrieval in the 'similarity search' method.

For the fourth retrieval method, 'lecture_notes,' we provide context to the LLM in the form of lecture notes (written by Sarah Lawsky, one of this paper's co-authors and a Professor of Law at Northwestern University Law School) relevant to the given question type [19].

Another experimental variable was whether we employ 'chain-of-thought' (CoT) prompting, which asks the LLM to think through its response step-by-step.

---

[2] Before deploying this evaluation method at scale to grade all of our experiments, we assessed its efficacy on a subset of questions across all the LLMs we were going to evaluate; our GPT-4 LLM evaluator perfectly graded nearly all of these questions, only erring less than 1% of times when grading the most ambiguous and difficult-to-grade outputs from the oldest and least capable model we evaluated, OpenAI's 'davinci' release. For example, davinci can at times generate outputs with multiple conflicting answer choices (even when instructed not to), occasionally fooling our LLM grader that davinci correctly picked the one true answer. But this did not occur frequently enough to significantly impact the overall trends we report, which we determined by manually reviewing much of the davinci grading.

[3] We are developing 'in-domain' embeddings currently for follow-up work.

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

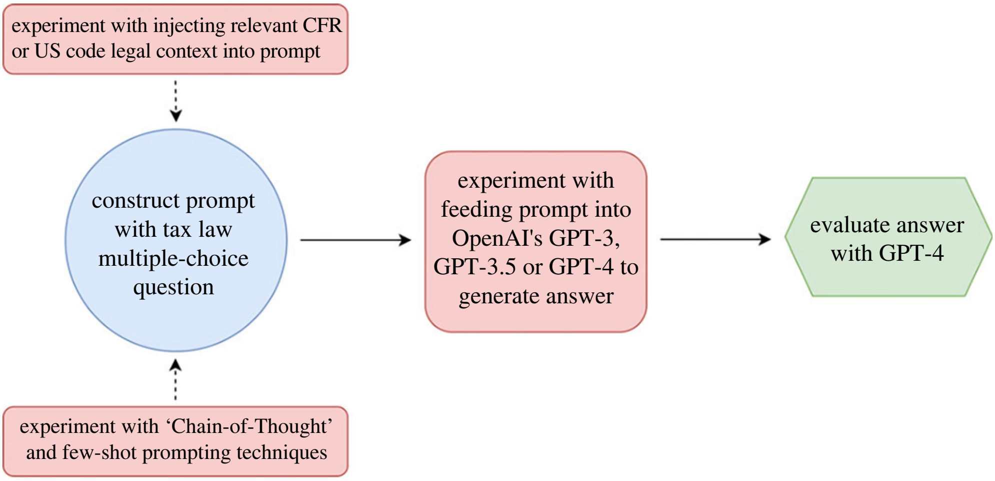
experiment with injecting relevant CFR or US code legal context into prompt
construct prompt with tax law multiple-choice question
experiment with feeding prompt into OpenAI's GPT-3, GPT-3.5 or GPT-4 to generate answer
evaluate answer with GPT-4
experiment with 'Chain-of-Thought' and few-shot prompting techniques

Figure 1. Our experimental pipeline compares performance on multiple-choice tax law exams across different LLMs, document retrieval techniques and prompting techniques.

Finally, we experimented with few-shot prompting. This is where we provide a set of three other question–answer pair examples to the LLM, along with the question being asked. We varied the pairs to match the question type for the given problem and ensured that the question–answer pairs were not any of the questions used for evaluation. The notion behind few-shot prompting is to guide the LLM toward how to answer the given question by observing examples of how to answer questions. We did this for all LLMs without providing contextual source documents or lecture notes.

4. The LLMs

A primary factor we vary in our experiments is the LLM itself. We run the experimental design across four increasingly advanced LLMs released by OpenAI over the past three years. The weakest model we employ, 'davinci,' is the 'most capable GPT-3 model.' 'text-davinci-002' is an earlier version of GPT-3.5 that is 'trained with supervised fine-tuning instead of reinforcement learning.' 'gpt-3.5-turbo' is the 'most capable GPT-3.5 model.' The most capable model, 'gpt-4,' is 'more capable than any GPT-3.5 model, able to do more complex tasks[4].'

For all models across all experiments, we set temperature equal to zero when generating responses to our prompts. Temperature is a parameter that controls the 'randomness' of the model's output. For these LLMs, lower temperatures make the outputs more deterministic.

Finally, after running the experimental design across factors, we then run a final setting, 'mega_run,' which combines the 'gold_truth' retrieval method, few-shot prompting, and CoT prompting of GPT-4. In other words, the most powerful combination of techniques and the most powerful model. This allows us to assess the upper bound on performance with these particular techniques. We discuss more advanced prompting in the Related Work and Next Steps sections that is likely to further boost performance.

Our total sample across the experiments contains 28 700 answers. Figure 1 visualizes the process and the experimental factors (in red).

5. Results

The first question we wanted to answer is whether CoT consistently improved the results for all (or most) models and all (or most) methods of retrieval. The answer is no, as evidenced by

---

[4] See https://platform.openai.com/docs/models/.

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

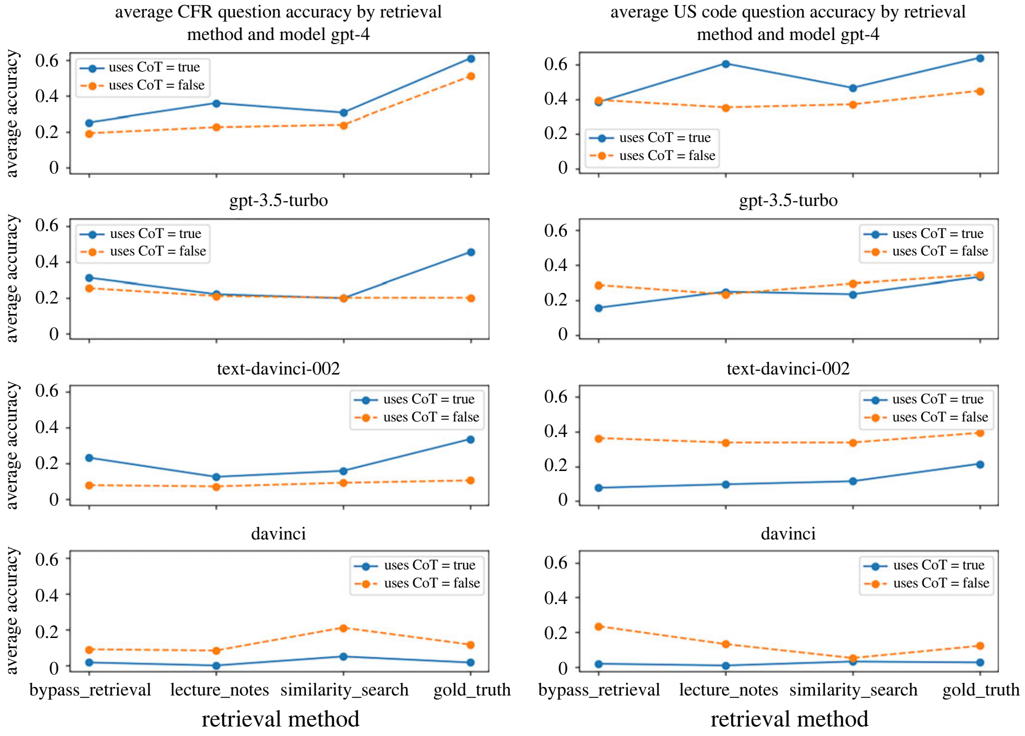

Figure 2. The *y*-axis is the accuracy of that experimental setting averaged across the different question sub-types. Spanning from left to right within each of the two columns, we generally see a slight improvement in overall accuracy as the LLM is provided with more relevant legal source material with each subsequent retrieval method.

the difference between the solid and dashed lines in the charts of figure 2. CoT does boost the performance of GPT-4, though. This suggests an LLM might need to have a certain capability level to be able to exhibit improved performance through additional reasoning. Two responses from GPT-4 prompted with CoT provide a sense of what our evaluation data looks like (Examples 1 and 2).

(a) Example 1

(i) Question

Aurora is hired by Grasshopper, Inc., on 17 September 2023. On her first day of work, when the stock trades at $18 per share, Aurora acquires 20 shares of Grasshopper stock for $15 per share, as contemplated by her employment agreement. The stock is a capital asset in Aurora's hands. As part of the terms of her employment, if she does not work for Grasshopper for 4 years, Aurora must return the stock. Anyone to whom Aurora transfers the stock is subject to the same restriction. Aurora makes no elections with respect to the stock. When the restriction expires and the stock vests in 2027, it is worth $23. Aurora sells the shares in 2028, when the stock is worth $25 per share. How much may Grasshopper deduct per share in 2027 due to the vesting of the shares?

(1) $0 (2) $2 (3) $3 (4) $7 (5) $8 (6) $10 (7) $15 (8) $18 (9) $23 (10) $25.

(ii) GPT-4 response

According to §83*(a), the amount to be included in the gross income of the person who performed the services is the excess of the fair market value of the property at the time the rights become transferable or not subject to a substantial risk of forfeiture, over the amount paid for the property. In this case, the fair market value of*

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

the stock when it vests in 2027 is $23 per share, and Aurora paid $15 per share. The excess is $23–$15 = $8 per share. §83(h) states that the employer, Grasshopper, can deduct an amount equal to the amount included in the gross income of the person who performed the services. In this case, that amount is $8 per share. So, Grasshopper may deduct (5) $8 per share in 2027 due to the vesting of the shares.

(iii) Correct answer

(5) $8

(b) Example 2

(i) Question

Zahra owns property that is worth $41 000, with a basis of $33 000. Zahra dies and leaves the property to Dylan. Several years later, Dylan sells the property for $37 000. How much gain or loss does Dylan recognize due to his sale of the property for $37 000?

(1) –$4000 (2) –$3000 (3) –$488 (4) $0 (5) $2000 (6) $4000 (7) $14 000

(ii) GPT-4 response

Zahra dies and leaves the property to Dylan. According to the context provided, the basis of the property in the hands of Dylan would be the fair market value of the property at the date of Zahra's death, which is $41 000. Now, Dylan sells the property for $37 000. To calculate the gain or loss, we subtract the basis from the selling price: $37 000–$41 000 = –$4000. So, the answer is (1) –$4000.

(iii) Correct answer

(1) –$4000

The second experimental question is whether giving the LLM more legal text and more relevant (to the specific question we are asking it) legal text increases accuracy for all (or most) models. The answer is weakly yes, as evidenced by the upward slope of some of the lines in the charts of figure 2. As you move from the left to the right in these charts, the experimental setup is providing more and/or more relevant legal text to the LLM, starting with the 'bypass_retrieval' setup, where we don't give any source material to the LLM, to 'gold_truth,' where we give the most relevant source material to the LLM.

The third experimental question is whether few-shot prompting helps. The answer is a strong yes for GPT-4 and seems to be less consistently useful for weaker models. In the 'few_shot' experimental setting, we do not give any source material to the LLM, but we input into the prompt examples of questions and answers from other questions that we are not testing it on, i.e. 'few_shot' is 'bypass_retrieval' plus few-shot prompting. The 'mega_run' experiment combines 'gold_truth' sources with few-shot and CoT prompting. As evidenced by figure 3, GPT-4 is able to leverage relevant legal text and examples of the question-and-answer task to 'reason' and come to a correct answer a large proportion of the time on difficult tax questions.

The primary experimental factor causing consistent increases in accuracy, when averaged across the other factors, is which underlying LLM is being used. This is consistent across the CFR and US Code focused questions; the same pattern holds of newer models outperforming older models, as shown in figure 4.

6. Implications

Our work represents a step toward adapting LLMs to autonomously and reliably reason about law. While our experiments are limited to US tax law, the capabilities the experiments investigate—finding relevant legal authorities and applying them to specific factual scenarios—are at the heart of legal work and could be generalized to other areas of legal practice. The

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

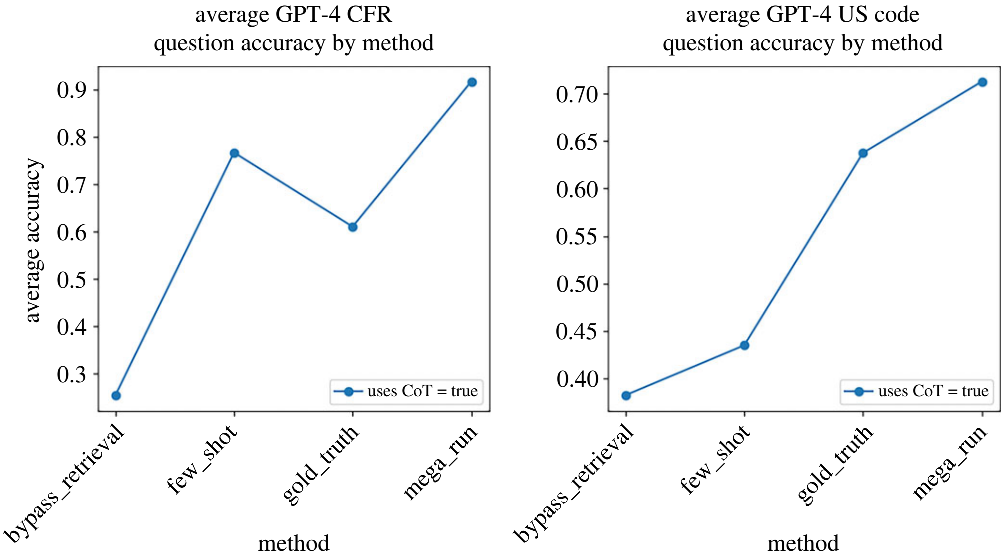

Figure 3. The *y*-axis is the accuracy of that experimental setting averaged across the different question sub-types. The 'mega_run' experimental set-up for GPT-4, which combines few-shot and CoT prompting, along with providing 'gold truth' legal sources, results in best overall accuracy for both the CFR and US Code exams. CoT boosts GPT-4 performance in the retrieval experimental settings of providing both no legal text ('bypass retrieval' and 'few shot') and the most relevant possible legal text ('gold truth' and 'mega run').

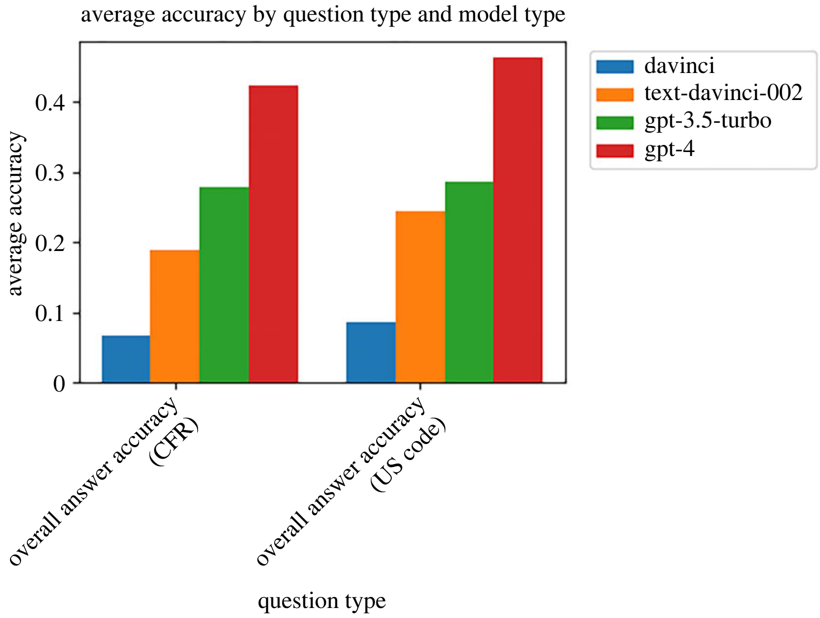

Figure 4. For both the CFR and US Code exams, we see a clear increase in overall answer accuracy with each subsequently released OpenAI model. The most capable model, GPT-4, with both prompting enhancements (CoT and few-shot) and the most relevant 'gold_truth' legal text input into the prompt, is able to perform extremely well, far better than any other set-up in the experiments (see 'mega_run' in figure 3).

increasing performance of LLMs on these tasks could have profound implications for the practice of law in areas similar to tax law, and the governance of AI.

First, lawyers are highly trained professionals, and LLMs could disrupt the legal services industry to the extent they are able to replicate much of the work of a skilled lawyer. We

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

do not wish to overstate this possibility, since even our best current models underperform a professional tax lawyer, who would be expected to answer these questions with near-perfect accuracy. Moreover, answering clear-cut legal questions is only a small part of the work of a practising lawyer. Clients rely on lawyers for contextual advice, ethical counsel, and nuanced judgement, which at present LLMs are not able to provide as consistently as most human lawyers. Nevertheless, there is no strong reason to believe that LLMs could not eventually accomplish a wide range of legal tasks with greater performance, and our work represents a benchmark to track the improvement of LLMs at legal reasoning.

Second, even if LLMs are not *replacing* trained lawyers, they can *assist* a lawyer or provide a first draft of work a lawyer could subsequently check. This could significantly increase the productivity of practising lawyers and decrease the cost of legal services, potentially improving access to legal counsel for many people who currently cannot afford it. In addition, LLMs could provide useful legal information to consumers who are not engaging a traditional lawyer. As LLMs become more capable of autonomously providing basic legal advice, policymakers might have to reconsider regulations on how legal advice is delivered, including regulations about the unauthorized practice of law.

Third, governance is a key component of aligning AI with humans. Methods that improve LLM legal analysis skills are relevant, either by helping AI models ‘self-police’ to ensure they are acting in accordance with law, or by designing separate models that can apply legal and ethical standards to confirm whether or not another AI is properly aligned with the law.

Our work also adds to the literature on emergent capabilities of LLMs by demonstrating the emergence of tax law understanding, which occurs once the LLM is of sufficient underlying general capability and is adequately prompted to elicit ‘reasoning’ behaviour. Extrapolating these capabilities forward, LLMs being able to ‘understand’ law would affect lawmaking [20] and necessitate changes to legal services regulation and emerging AI governance regimes.

7. Related work

LLM prompting involves designing text inputs to generate a response from an LLM. The goal of prompting is to steer the behaviour of the LLM in a way that elicits a desired outcome. Recent research has focused on developing effective prompting techniques that can expand LLMs’ capabilities when carrying out a variety of tasks. Examples include prompt patterns [21], in-context instruction learning [22], evolutionary prompt engineering [23] and domain-specific keywords with a trainable gated prompt to guide toward a target domain for general-domain LLMs [24]. Zhong *et al.* [25] experiment with prompting LLMs to do scientific tasks across fields like business, science, and health by providing the LLM with a research goal and two large corpora, asking the LLM for corpus-level difference. Reppert *et al.* [26] develop iterated decomposition, a human-in-the-loop workflow for developing and refining compositional LLM programs that improves performance on real-world science question and answer tasks.

More advanced techniques involve processes such as annotation, distillation, and model self-reflection. Diao *et al.* [27] developed Active-prompt, which finds the most uncertain questions for the LLM and annotates those from the pool, achieving the state of the art on complex reasoning tasks. Zhuo *et al.* [28] develop methods for automatically designing multiple prompts and integrating automatic verbalizers without sacrificing performance. LLMs can also improve through introspection. For example, Kim *et al.* develop a prompting scheme where an LLM agent recursively criticizes and improves its output (RCI), outperforming supervised learning and reinforcement learning approaches on the MiniWoB++ benchmark, a web-based simulation task suite with tasks ranging from simple clicking to complex maths problems [29]. Press *et al.* investigate LLMs’ abilities to engage in compositional reasoning tasks, finding that as model size increases, single-hop question-answering ability improves more rapidly than multi-hop question-answering ability, resulting in a ‘compositionality gap’. The authors propose ‘elicitive’ prompting methods, such as CoT and self-talk, to mitigate this gap [30]. Yao *et al.* [31] developed the popular ReAct approach where LLMs generate reasoning and actions in an interleaved manner,

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

outperforming state-of-the-art baselines at the time across various tasks. Jin *et al.* [32] develop 'Moral Chain-of-Thought' (MORALCoT) prompting, which draws from cognitive science theories of moral reasoning and excels in a novel challenge set centred on permissible rule-breaking.

A growing body of research examines the characteristics of prompting. For instance, Lu *et al.* [33] find that the performance of LLMs is not associated with the perceived difficulty of prompts estimated by human annotators, and that employing definitions, demonstrations and explanations can enhance performance. Halawi *et al.* [34] investigate model performance when confronted with misleading or false prompts and reveal that LLMs exhibit comparable performance, irrespective of few-shot prompt accuracy, while accuracy discrepancies due to deceptive prompts only emerge in later layers of the model. Focusing specifically on discrete prompts, Ishibashi *et al.* demonstrate that although these prompts exhibit a degree of robustness against certain perturbations, they remain vulnerable to others and fail to generalize effectively across natural language inference datasets. This underscores the necessity for further exploration into robust discrete prompting [35]. Focusing on the role of prompting in boosting LLMs' 'Theory-of-Mind' performance, Moghaddam and Honey show that in-context learning prompts boost Theory-of-Mind abilities in GPT-4 and GPT-3.5 models [36].

Prompting serves as a crucial element in using LLMs for real-world applications such as legal services and legal question-answering, as it connects model capabilities with targeted functionalities. In the context of our study, we examine LLMs' capacity to comprehend regulations and laws and experiment with the effects of very simple prompting techniques on accuracy. We leave the more advanced prompting discussed here for follow-up work in adapting these techniques to the legal domain.

Another burgeoning part of the LLM literature is dedicated to the capacity of LLMs to function as agents that perform tasks, make decisions, and interact with their environment. Andreas *et al.* [37] demonstrate that LLMs can serve as agent models when only trained on bodies of documents, by implicitly inferring fine-grained communicative intentions and using that for subsequent text generation. LLM-powered agents have demonstrated competence on some tasks that require reasoning, especially when combined with 'tools' and symbolic systems. For instance, an AI system, Cicero, achieved human-level performance in the strategy game Diplomacy by integrating an LLM with strategic reasoning [38]. Furthermore, Shinn *et al.* [39] explore LLM agents' ability for learning from mistakes with Reflexion, an approach that equips LLM-based agents with dynamic memory, 'self-reflection' capabilities and a method for detecting hallucinations.

Regarding agentic LLMs more generally, Yang *et al.* [40] investigate connections between LLMs and external entities, and their decision-making, using methods such as prompting, conditional generative modelling, planning, optimal control and reinforcement learning. A primary focus of autonomous agents lies in the interface between the LLM as an agent and the environment with which it interacts. Li *et al.* employ the 'Internet Explorer' approach, which enables LLMs to dynamically use the internet as a continuously updating, open-ended dataset. In this approach, smaller models explore the web through self-supervision, locating relevant data to quickly enhance task performance [41]. Carta *et al.* [42] examine a method to improve the alignment between the LLM's knowledge and its environment, while augmenting functional competence; the LLM is grounded in an interactive text world with online reinforcement learning, incrementally updating its knowledge based on observations. Agents need to plan, and there is substantial interest in LLMs' ability to act as planners. Valmeekam *et al.* investigate the planning capabilities of LLMs, which exhibit poor performance in fully autonomous mode during common-sense planning tasks. However, when 'heuristic guidance' and 'human-in-the-loop' modes are employed, performance improves, albeit marginally [43]. As an example of a direction toward autonomous planning, Wang *et al.* developed a 'Describe, Explain, Plan and Select' approach, which explores the use of LLMs as planning agents in open-ended planning scenarios with long-term, multi-step tasks. This approach significantly improved performance in over 70 Minecraft tasks [44]. Other research examines LLMs as a component in building AI agents. For example, Li *et al.* [45] explore the use of LLMs as probabilistic priors for generalized

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

decision-making, applicable to non-linguistic perception and control, as well as tasks such as semantic segmentation, household navigation and activity recognition. The explosion of research interest at the intersection of autonomous agents and LLMs is relevant to our work, since agents that better understand the law are more likely to be aligned with society. By benchmarking legal understanding of LLMs, we can contribute to assessing the safety of agentic LLM deployments.

As LLMs demonstrate significant potential in tackling diverse tasks, research has focused on methods of evaluating their performance. Increasingly specific benchmarks are being developed. Examples include G-Eval, a framework using LLMs to evaluate natural language generation output via a CoT paradigm [46], and AmbiEnt, where even advanced models like GPT-4 struggle with correctly disentangling ambiguous meanings [47].

Providing LLMs with domain-specific knowledge, updated data, and specialized reasoning and computation abilities can improve their performance on some tasks. Mialon *et al*. review the current advancements in augmentation, where LLMs are enhanced through reasoning capabilities, external modules and tools. The authors argue that augmentation could potentially ameliorate interpretability, consistency and scalability issues in LLMs [48]. Researchers have devised several methods for LLMs to employ external resources. For instance, Peng *et al*. [49] introduce a system that employs plug-and-play external modules to refine grounded responses using external knowledge and iterative revision based on utility function feedback, substantially reducing LLM hallucinations. Zhou *et al*. [50] develop Doc-Prompting, a natural-language-to-code technique that uses library documentation retrieval for code generation. External documentation can also facilitate LLM self-assessment: Wu *et al*. establish a 'Read and Reward' framework to enable an LLM to self-evaluate through manual learning. This framework employs a Question and Answer (QA) extraction module that condenses manual information and a reasoning module to assess interactions based on this information [51].

QA has served as the testing ground for most of the LLM augmentation research thus far. Chen *et al*. review open-domain QA research [52]. Sil *et al*. introduced PRIMEQA, an open-source repository to democratize cutting-edge QA methodologies. This end-to-end QA toolkit allows for custom app creation with trainable retrievers and readers for deployment [53]. Sun *et al*. [54] propose recitation-augmented language models, enabling LLMs to retrieve pertinent information from their own memory through sampling to answer questions. Khattab *et al*. [55] present Demonstrate-Search-Predict (DSP) for retrieval-augmented in-context learning that decomposes problems into more manageable components for both the language and retrieval models. Ye *et al*. [56] develop Compositional Exemplars for In-context Learning to assist in selecting the most diverse yet useful examples for LLMs to learn from for in-context learning. Ram *et al*. [57] present a simpler alternative to Retrieval-Augmented Language Modeling (RALM): in-context RALM, where grounding documents are incorporated into the LLM's input without modifying its architecture. In our paper, we focus on simple forms of augmentation, and leave testing these more sophisticated methods for future work.

There are many tasks that larger LLMs can complete that smaller models cannot [58]. Larger models have more inherent resources (for example, GPT-2 has 1.5 billion parameters while GPT-3 has 175 billion), and for some tasks that require various complex types of reasoning, LLMs' capability to do such tasks 'emerges' in a nonlinear fashion after reaching a certain model size. Jason Wei has compiled a list of 137 emergent abilities of LLMs that have been uncovered by research, which includes things like 'hindu language,' 'causal judgement' and 'geometric shapes [59].' Our experiments suggest that legal understanding could be one such emergent ability.

8. Next steps

With clear evidence showing increases in capabilities from older to newer LLMs, attention can be shifted towards validating and improving the abilities of the newest, most powerful models available.

Regarding prompting, further analysis of our results could investigate the relationship between prompt length and accuracy. One possibility is that the LLMs do not perform as well as

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

they could because their performance degrades as the length of the input increases; just because newer models like GPT-4 have a wider context window may not necessarily mean filling it to the max is optimal.

Many of the more advanced prompting techniques discussed in the Related Work section are prime candidates for increasing performance; in particular, the self-reflection and self-refinement techniques. For example, the LLM can be prompted with its own answers, and the relevant context, and asked, ‘Are there any ambiguities in this question that make it difficult to answer or for you to doubt your current answer? If so, conduct additional legal research by generating a topic that we need to search legal sources for’. The response can then be used to conduct further retrieval-augmented generation.

Regarding document retrieval, we seek to close the gap between the ‘similarity search’ and ‘gold truth’ retrieval methods through better retrieval. Especially for GPT-4, we saw a clear performance boost when feeding in the ‘gold truth’ legal documents, rather than performing similarity search to extract the relevant documents from a vector database. This result indicates that our similarity search technique, and the various hyperparameter defaults we used, did not provide the most relevant ‘gold truth’ sources into the LLM a significant portion of the time. Ultimately, as LLMs are deployed in real-world settings where humans would not be providing the exact legal documents necessary, the ability to retrieve the relevant documents will be important. We need to experiment with factors such as the choice of model embeddings, retrieval technique, and the token length of vector database subsections and number of subsections retrieved and placed into the prompt.

Finally, future work could compare performance between generally pre-trained LLMs, such as the OpenAI models in our experiments, and language models specifically pre-trained and fine-tuned for legal reasoning. Developing best practices for fine-tuning models for legal reasoning tasks is an important step towards sufficiently boosting AI legal capabilities in real-world settings.

Data accessibility. The datasets supporting this article are available in GitHub: https://github.com/JohnNay/llm-tax-attorney/tree/main/data. The code that generated that data is available upon request for other researchers by emailing sarah.lawsky@law.northwestern.edu.

Supplementary material is available online [60].

Declaration of AI use. We have used AI-assisted technologies in creating this article.

Authors’ contributions. J.J.N.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, visualization, writing—original draft, writing—review and editing; D.K.: conceptualization, data curation, formal analysis, investigation, project administration, validation, writing—original draft, writing—review and editing; S.B.L.: conceptualization, data curation, investigation, methodology, resources, software, supervision, validation, writing—review and editing; W.T.: investigation, project administration, writing—original draft, writing—review and editing; M.B.: data curation, investigation, validation, writing—review and editing; R.J.: conceptualization, data curation, formal analysis, investigation, software, validation, visualization, writing—review and editing; A.T.L.: conceptualization, data curation, investigation, software, validation, writing—review and editing; J.H.C.: conceptualization, methodology, resources, writing—review and editing; J.K.: conceptualization, data curation, methodology, software, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. J.J.N. and D.K. are employees at Norm Ai, an AI technology company.

Funding. The Mercatus Center at George Mason University funded Meghana Bhat’s work related to this research, and some of the computing costs for running the experiments.

Acknowledgements. Pierre-Loic Doulcet, of Stanford University’s CodeX Center for Legal Informatics, contributed to the US Code vector embeddings work.

References

1. Benaich N, Hogarth I. 2022 State of AI Report 2022. STATEOF.AI. 11 Oct 2022. See https://www.stateof.ai/2022-report-launch.html.

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

2. Wei J. 2022 137 Emergent Abilities of Large Language Models. JASONWEI.NET. 14 November 2022. See https://www.jasonwei.net/blog/emergence.
3. Webb T *et al.* 2022 Emergent Analogical Reasoning in Large Language Models. 19 December 2022. See https://arxiv.org/abs/2212.09196.
4. Hafner D *et al.* 2023 Mastering Diverse Domains through World Models. 10 January 2023. See https://arxiv.org/abs/2301.04104.
5. Institute for Human-Centered AI. 2023 STANFORD UNIV. [cited 20 March 2023]. See https://hai.stanford.edu.
6. OpenAI. 2019 Better Language Models and Their Implications. OPENAI. 14 February 2019. See https://openai.com/research/better-language-models.
7. OpenAI. 2023 GPT-4. OPENAI. 14 March 2023. See https://openai.com/research/gpt-4.
8. Schick T *et al.* 2023 Toolformer: language models can teach themselves to use tools. 2023. See https://arxiv.org/abs/2302.04761.
9. World Economic Forum. 2022 The Promises and Perils of AI, with Prof Stuart Russell. WORLDECON. FORUM. 6 January 2022. See https://www.weforum.org/agenda/2022/01/artificial-intelligence-stuart-russell-radio-davos/.
10. Zhang D *et al.* 2022 The AI Index 2022 Annual Report. STANFORD INST. FOR HUMAN-CENTERED AI, STANFORD UNIV. [cited 20 March 2023]. See https://aiindex.stanford.edu/report/.
11. The New AI-Powered Bing is Threatening Users. 2023 That's no Laughing Matter. TIME. 17 February 2023. See time.com/6256529/bing-openai-chatgpt-danger-alignment/
12. Magar, I, and Schwartz R. 2022 Data contamination: from memorization to exploitation. (http://arxiv.org/abs/2203.08242)
13. Chang KK, Cramer M, Soni S, Bamman D. 2023 Speak, memory: an archaeology of books known to chatgpt/gpt-4. (http://arxiv.org/abs/2305.00118)
14. Nay J. 2023 Law informs code: a legal informatics approach to aligning artificial intelligence with humans. *Northwestern J. Tech. Intell. Prop.* **20**, 309.
15. Nay JJ. 2023 Large language models as fiduciaries: a case study toward robustly communicating with artificial intelligence through legal standards. SSRN. 30 Jan 2023. See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335945.
16. Jean R. 2023 'Wyoming lawmakers turn to experts to stay ahead of AI curve, even as ChatGPT4 Passes Bar Exam' 20 May 2023. See https://cowboystatedaily.com/2023/05/20/wyoming-lawmakers-turn-to-experts-to-stay-ahead-of-ai-curve-even-as-chatgpt4-passes-the-bar-exam/https://cowboystatedaily.com/2023/05/20/wyoming-lawmakers-turn-to-experts-to-stay-ahead-of-ai-curve-even-as-chatgpt4-passes-the-bar-exam/.
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017 Attention is all you need. In Advances in neural information processing systems.
18. Ni J *et al.* 2022 Large dual encoders are generalizable retrievers. In *Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, pp. 9844–9855. New York, NY: Association for Computational Linguistics.
19. Lawsky S. 2023 Lawsky practice problems. [cited 2023 April 25]. See www.lawskypracticeproblems.org.
20. Nay JJ. 2023 Large language models as corporate lobbyists. [eprint]. 1 January 2023. See https://arxiv.org/abs/2301.01181.
21. White J *et al.* 2023 A prompt pattern catalog to enhance prompt engineering with ChatGPT. [cs.SE]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.11382.
22. Ye S, Hwang H, Yang S, Yun H, Kim Y, Seo M. 2023 In-context instruction learning. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.14691.
23. Chen A, Dohan DM, So DR. 2023 EvoPrompting: Language Models for Code-Level Neural Architecture Search. [cs.NE]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.14838.
24. Goswami K, Lange L, Araki J, Adel H. 2023 SwitchPrompt: learning domain-specific gated soft prompts for classification in low-resource domains. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.06868.
25. Zhong R, Zhang P, Li S, Ahn J, Klein D, Steinhardt J. 2023 Goal driven discovery of distributional differences via language descriptions. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.14233.

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025

26. Reppert J, Rachbach B, George C, Stebbing L, Byun J, Appleton M, Stuhlmüller A. 2023 Iterated decomposition: improving science Q&A by supervising reasoning processes. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2301.01751.
27. Diao S, Wang P, Lin Y, Zhang T. 2023 Active prompting with chain-of-thought for large language models. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.12246.
28. Zhou Y, Maharjan S, Liu B. 2023 Scalable prompt generation for semi-supervised learning with language Models. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.09236.
29. Kim G, Baldi P, McAleer S. 2023 Language models can solve computer tasks. (http://arxiv.org/abs/2303.17491). 31 March 2023.
30. Press O, Zhang M, Min S, Schmidt L, Smith NA, Lewis M. 2022 Measuring and narrowing the compositionality gap in language models. (http://arxiv.org/abs/2210.03350). 7 October 2022.
31. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. 2023 ReAct: synergizing reasoning and acting in language models. (http://arxiv.org/abs/2210.03629). 10 October 2023.
32. Jin Z, Levine S, Gonzalez F, Kamal O, Sap M, Sachan M, Mihalcea R, Tenenbaum J, Schölkopf B. 2022 When to make exceptions: exploring language models as accounts of human moral judgment. *Adv. Neural Inform. Processing Systems* **35**, 28 458–28 473.
33. Lu A, Zhang H, Zhang Y, Wang X, Yang D. 2023 Bounding the capabilities of large language models in open text generation with prompt constraints. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.09185.
34. Halawi D, Denain JS, Steinhardt J. 2023 Overthinking the truth: understanding how language models process false demonstrations. 2023 [cited 31 March 2023]. See https://openreview.net/forum?id=em4xg1Gvxa.
35. Ishibashi Y, Bollegala D, Sudoh K, Nakamura S. 2023 Evaluating the robustness of discrete prompts. [cs.CL]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.05619.
36. Moghaddam SR, Honey CJ. 2023 Boosting theory-of-mind performance in large language models via prompting. (http://arxiv.org/abs/2304.11490)
37. Andreas J. 2022 Language models as agent models. (http://arxiv.org/abs/2212.01681)
38. FAIR Bakhtin A *et al.* 2022 Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* **378**, 1067–1074. (doi:10.1126/science.ade9097)
39. Shinn N, Labash B, Gopinath A. 2023 Reflexion: an autonomous agent with dynamic memory and self-reflection. (http://arxiv.org/abs/2303.11366). 28 March 2023.
40. Yang S, Nachum O, Du Y, Wei J, Abbeel P, Schuurmans D. 2023 Foundation models for decision making: problems, methods, and opportunities. [cs.AI]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2303.04129.
41. Li AC, Brown E, Efros AA, Pathak D. 2023 Internet explorer: targeted representation learning on the open web. [cs.LG]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.14051.
42. Carta T, Romac C, Wolf T, Lamprier S, Sigaud O, Oudeyer PY. 2023 Grounding large language models in interactive environments with online reinforcement learning. (http://arxiv.org/abs/2302.02662). February 2023.
43. Valmeekam K, Sreedharan S, Marquez M, Olmo A, Kambhampati S. 2023 On the planning abilities of large language models (a critical investigation with a proposed benchmark). [cs.AI]. 2023 [cited 31 March 2023]. See https://arxiv.org/abs/2302.06706.
44. Wang Z, Cai S, Liu A, Ma X, Liang Y. 2023 Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. (http://arxiv.org/abs/2302.01560). 3 February 2023.
45. Li BZ, Chen W, Sharma P, Andreas J. 2023 LaMPP: language models as probabilistic priors for perception and action. (http://arxiv.org/abs/2302.02801). 7 February 2023.
46. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. 2023 G-Eval: NLG evaluation using GPT-4 with better human alignment. (http://arxiv.org/abs/2303.16634)
47. Liu A, Wu Z, Michael J, Suhr A, West P, Koller A, Swayamdipta S, Smith NA, Choi Y. 2023 We're afraid language models aren't modeling ambiguity. (http://arxiv.org/abs/2304.14399)
48. Mialon G *et al.* 2023 Augmented language models: a survey. 21 February 2023. (http://arxiv.org/abs/2302.07842)

49. Peng B *et al.* 2023 Check your facts and try again: improving large language models with external knowledge and automated feedback. 28 February 2023. (http://arxiv.org/abs/2302.12813)
50. Zhou S, Alon U, Xu FF, Wang Z, Jiang Z, Neubig G. 2023 DocPrompting: generating code by retrieving the docs. (http://arxiv.org/abs/2207.05987)
51. Wu Y, Fan Y, Liang PP, Azaria A, Li Y, Mitchell TM. 2023 Read and reap the rewards: learning to play atari with the help of instruction manuals. (http://arxiv.org/abs/2302.04449). 1 February 2023.
52. Chen D, Yih W. 2020 *Open-domain question answering*. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 34–37: Online: Association for Computational Linguistics.
53. Sil A *et al.* 2023 PrimeQA: the prime repository for state-of-the-art multilingual question answering research and development. (http://arxiv.org/abs/2301.09715). 26 January 2023.
54. Sun Z, Wang X, Tay Y, Yang Y, Zhou D. 2023 Recitation-augmented language models. (http://arxiv.org/abs/2210.01296)
55. Khattab O, Santhanam K, Li XL, Hall D, Liang P, Potts C, Zaharia M. 2023 Demonstrate-search-predict: composing retrieval and language models for knowledge-intensive NLP. (http://arxiv.org/abs/2212.14024)
56. Ye J, Wu Z, Feng J, Yu T, Kong L. 2023 Compositional exemplars for in-context learning. (http://arxiv.org/abs/2302.05698). 15 February 2023.
57. Ram O, Levine Y, Dalmedigos I, Muhlgay D, Shashua A, Leyton-Brown K, Shoham Y. 2023 In-context retrieval-augmented language models. (http://arxiv.org/abs/2302.00083). February 2023.
58. Ornes S. 2023 The unpredictable abilities emerging from large AI models. Quanta Magazine. 2023 Mar 16 - [cited 21 May 2023]. See https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316.
59. Wei J. 2023 137 Emergent abilities of large language models. Jason Wei Blog. [Date unknown] - [cited 21 May 2023]. See https://www.jasonwei.net/blog/emergence.
60. Nay JJ, Karamardian D, Lawsky SB, Tao W, Bhat M, Jain R, Lee AT, Choi JH, Kasai J. 2024 Large language models as tax attorneys: a case study in legal capabilities emergence. Figshare. (doi:10.6084/m9.figshare.c.7031271)

Downloaded from https://royalsocietypublishing.org/ on 02 January 2025