I'd really like there to be a lot more systematic scientific study of what systems built out of language models are and aren't capable of, ideally as closely tied to real-world impacts as possible (rather than sticking to well-scoped benchmark tasks such as multiple choice tests). For example, I'd like to see more work on all of the following:

- **Evaluations of language agents.** This is work that builds agents out of language models (e.g. natbot, AutoGPT, LangChain, ReAct, the forthcoming ACT-1, which I'll call *language agents*) and assesses these language agents' ability to autonomously achieve open-ended goals that require taking a sequence of actions. I am especially interested in evaluating language agents on tasks that require interacting with the real world over the internet. Recent work in this vein includes WebShop (Yao et al. 2022), Mind2Web (Deng et al. 2023), WebArena (Zhou et al. 2023), and AgentBench (Liu et al 2023).
    - **More difficult and open-ended tasks.** Currently, most tasks studied in these papers are relatively simple.[1] While language agents are not achieving high success rates at this moment (e.g. performance on WebArena is around 10%), I expect this to change very quickly given the pace of recent progress.[2] That means I'd really like to see task sets with a broader scale of difficulty, including some tasks that are nearly as difficult as fully automating certain professions (e.g. "Find and exploit a novel vulnerability in a real-world piece of software").
        - The forthcoming paper Kinniment et al. 2023 explores whether language agents can "acquire resources, create copies of themselves, and adapt to novel challenges they encounter in the wild." This paper evaluates agents on a suite of tasks spanning a much broader range of difficulty than seen in previous work. I'd love to see more work in this vein.
        - I am especially interested in studying potentially harmful capabilities, such as the ability to carry out targeted phishing or ransomware attacks, the ability to find and exploit software vulnerabilities, or the ability to design and manufacture weapons.[3]
    - **Human judgments of success.** In most cases it won't be practical to automatically evaluate success on these more difficult, more open-ended tasks; success criteria may have to be determined by human judgment.
        - For example, in Kinniment et al. 2023, success criteria for each task is pre-registered but must ultimately be determined by human judgment. Similarly, this game by Nicholas Carlini asks players to predict whether

[1] They are often variations of searching the web for some information, or else performing tasks that involve navigating only a few pages or buttons, such as "Set my gitlab status as Enjoying Life" or "Cancel order 307."

[2] For example, scores on MATH went from ~5% in 2021 to over 50% in mid-2022 (for the best models not specifically fine-tuned for performance on MATH). In mid-year 2021, Jacob Steinhardt commissioned forecasts on MATH performance from the superforecasting group Hypermind, and these forecasters estimated that it would take until 2025 to achieve that performance. A similar story happened for MMLU: forecasts made in 2021 substantially underpredicted progress in 2022. Broadly, I have the sense that the time from introducing a benchmark to that benchmark reaching high levels of performance has shrunk.

[3] Work that studies overtly dangerous capabilities obviously has more ethical and practical issues than other work, but I think it's possible to study effectively and ethically; if you're interested in doing a project along these lines, I'd be happy to chat more.

GPT-4 will succeed at a variety of tasks; for most of them, Carlini subjectively determines (based on pre-registered criteria) whether or not the model "succeeded."
        - Unfortunately, these human evaluations are highly disputable; it can be difficult to know what human evaluations really mean, or to have a replicable process for generating well-grounded evaluations.[4]
        - **Grading rubrics.** One possible way to generate somewhat more grounded human evaluations is to exploit tasks which come with pre-established grading rubrics, such as take-home essays and exams, or take-home job interviews.[5] For example, Maya Bodnick of SlowBoring asked her Harvard professors to grade essays written by ChatGPT in response to real take-home essay prompts from her classes (it received pretty good grades).
        - **Blinded preference comparisons.** While remote work tasks like Upwork, MTurk, or Fiverr don't come with pre-established standards of success, we could potentially directly compare which work product the client prefers (the human one or the one made by an LLM system). This is another possible way to generate somewhat more grounded and interpretable human evaluations.
        - There may be other creative means of generating semi-grounded human evaluations of language agent work. For example, can a language agent get its pull request to an open source repository accepted?
    - In some cases there may be tasks that are simultaneously *open-ended* (involving a long sequence of steps or interaction with the world) and *automatically gradable.* For example, can a language agent solve difficult Kaggle competitions or capture the flag competitions if it can browse the web and use a code interpreter as it works? Can a language agent prove a difficult math theorem, according to a formal proof verifier? Can a language agent replicate an economics or statistics paper and get results within some tolerance of the original results? I can imagine more creative automatically-gradable tasks like these.
    - **Analysis or measurement of partial success.** Particularly if we include very difficult tasks in the scale, it's likely that current language agents won't be able to

[4] For example, in Kinniment et al 2023 the success criteria for a phishing task involves human overseers making a judgment about whether a false login page created by the agent is convincing enough to fool a phishing target. In Carlini's game, one of the questions involves making a subjective judgment about whether a song is sufficiently close to "Happy Birthday." In both cases, there is a lot of room to dispute that the criteria were too harsh or too lax.

[5] In many of these tasks humans would be allowed to access the internet or other tools to perform these tasks, so we should compare their performance to language agents / systems that also have access to the relevant tools. For example, take-home programming interviews would be done with access to a code interpreter; the language agent should also be given the same access for the same interview. In some cases (e.g. take-home exams), humans would not be allowed access to the internet, so we could compare performance to a language agent that is similarly limited.

fully complete the task, but will be able to make some progress toward it.[6] (This is likely to apply whether ultimate success is judged by human evaluations or by automatic evaluations; e.g. an agent might be able to prove certain lemmas but not a full theorem.)
        - I'd be interested in qualitative analysis of the kinds of steps browsing agents tend to fail at and the kinds of errors they tend to make. I'd also be excited about trying to quantify partial success (this could be as simple as having a panel of experts subjectively rate "how far" the agent got in a certain task attempt).
- **Human assistance RCTs.** This is research that assesses how much language model systems / products help humans perform real-world tasks through randomized controlled trials. For example, GitHub did a study in mid-2022 which found that having access to GitHub CoPilot halved the time that programmers needed to write an HTTP server in JavaScript (from ~2 hours to ~1 hour). Anecdotally, GPT-4 is considerably more useful, but I haven't seen any systematic study of it.
    - I'm especially interested in studying capabilities in the domain of programming and ML research, because I am interested in understanding whether we should expect to see a feedback loop in which progress accelerates further because LLMs are used to more quickly and easily deploy and develop LLMs.
    - I would like to see comparisons between language models of different sizes, as well as between language models and other productivity tools.
- **Data collection and analysis.** This is work that collects and analyzes existing data relevant to the question of what systems built out of language models can and can't do in the real world. For example:
    - **Polling people** to ask them whether they use language model products, how much, for what tasks, how useful they are, etc. I've seen informal surveys from e.g. Business.com and FishBowl, but so far I haven't seen anything from a reputable survey firm using best practices. I'd be especially interested in user surveys that do more of a deep dive into the types of tasks they are helpful and unhelpful for than these informal surveys provide.
    - **Collecting case studies** of "in the wild" use of language models, for example by scraping Reddit (e.g. r/chatGPT), or by asking people to submit case studies to a dedicated database, or even partnering with a company to systematically collect examples from consenting customers. While there are a lot of individual case studies on the internet, I'm not aware of existing work that collects and analyzes them. Even though they are not going to be a representative sample, I think seeing thousands of examples of attempts to use language models by real people in the course of real jobs could be helpful for understanding qualitative patterns of language model strengths and weaknesses.
    - **Gathering key numbers** into one convenient place to support analysis. For example, HELM evaluates a wide variety of language models on a wide variety of

[6] For example, in Kinniment et al 2023, an agent built out of GPT-4 was able to draft a phishing email that could plausibly get a phishing target to click a link — but it was not able to generate a website realistic enough that it would plausibly fool a target into entering their login credentials.

existing benchmarks, and Papers with Code also provides a similar handy reference. Epoch similarly provides a handy reference for numbers related to AI *inputs* (such as hardware price performance and spending on large training runs). I'd be interested in similar data estimation and collection efforts for key economic indicators, such as revenues of LLM products, valuations of LLM-exposed companies, number of users of LLM products, etc. As more real-world evaluations of language agents and human assistance RCTs are conducted, I would also like to see data collection on those results.

- **Synthesizing and summarizing** the various lines of evidence that are already out there about what language model systems can and can't do (including benchmark evaluations, market analysis, qualitative information, etc) and arriving at a qualitative overview of "the state of language model systems." There are existing overviews of the AI field, such as the AI 100 report or market reports like this from McKinsey, as well as occasional news articles like this recent one from TIME. I would be most excited about a systematic, frequently-updated qualitative overview which is narrowly focused on the capabilities of systems built out of language models.

I'm very excited about **kicking off a field trying to understand the full range of real-world capabilities of agents and other systems built out of language models** (and how quickly they are improving). In addition to research projects like the ones listed above, I'm interested in supporting:

- Workshops, conferences, and other collaborations on this topic.
- Projects aimed at communicating language model capabilities, such as Nicholas Carlini's game which I referenced above (I'm especially interested in communicating potentially dangerous capabilities to policymakers).
- Efforts to forecast indicators of real-world capabilities (for example Jacob Steinhardt's 2021 forecasting contest, some questions in the Existential Risk Persuasion Tournament, and some questions in AI Impacts' expert survey).