FROM CURIOSITY TO CAPABILITY

The Early Enterprise Generative AI Adoption Playbook

31 JANUARY 2023 • ENTERPRISE AI SERIES



Paul Forrest

PAPER ONE · ENTERPRISE AI, 2023 TO 2024

About this paper

This is the first paper in a series on the enterprise adoption of generative artificial intelligence. Each paper is written to what was known when it was written, and each stands alone.

The series exists because the gap between what these systems appear to do and what an organisation can responsibly claim they do is wide, and it widens fastest in the weeks after a capable product reaches the public. This paper covers the first two months of that gap.

About seven parts in ten of what follows reports and analyses public evidence. The rest is my own framework, and it is tagged as mine wherever it appears.

HOW TO CITE

> Forrest, P. (2023). *From Curiosity to Capability: The Early Enterprise Generative AI Adoption Playbook.* Enterprise AI, 2023 to 2024, Paper One. Version 1.0, 31 January 2023.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence stands behind it. A claim resting on two kinds carries both.

PRIMARY An institutional or official document, cited from the issuing body.

PEER-REVIEWED A study in a peer-reviewed journal or conference proceedings.

PREPRINT Research posted as a preprint or working paper, not peer-reviewed at this date.

VENDOR A vendor announcement or a figure the vendor reports about its own product.

AUTHOR My own framework or judgement, proposed and not yet proven.

A NOTE ON DATES

Research for this paper closed on 27 January 2023 and everything in it was current at that date. The references carry the titles and addresses the sources had then.

Contents

FRONT MATTER

	Abstract and summary of the argument	4

THE PAPER

1	The interface breakthrough, and what it did not change	5
2	A vocabulary precise enough to argue with	7
3	The opportunity map as it stands in January	8
4	A filter that screens on consequence	10
5	Designing a pilot that can fail visibly	12
6	What leaders should measure, and what they should refuse to	14
7	A 30-day charter	16

END MATTER

	Method, evidence and limits	17
	Sources considered and set aside	18
	About the author	19
	References	20
	Further sources consulted	21

ABSTRACT AND SUMMARY OF THE ARGUMENT

Two months of public access, and what an enterprise can honestly conclude from it

A conversational system capable of producing fluent prose on demand became publicly available on 30 November 2022, and within weeks employees in most large organisations were using it without being asked to and without being told they could not.[1] VENDOR That is the situation this paper addresses. It is a question about organisational learning under uncertainty, and the technology is only the occasion for it.

The argument runs against the instinct of the moment. An organisation that moves quickly to deploy will find that it has deployed something it cannot evaluate, because the evaluation capability does not yet exist anywhere, including inside the companies building these systems. An organisation that prohibits use will find that use continues and that it has lost the ability to see it. The position worth occupying sits between those two, and it is a learning system.

What follows sets out what could responsibly be known on 31 January 2023, what could be attempted, and what should be refused. It proposes a use-case filter, a bounded-pilot design and a 30-day charter. None of these require new technology, a budget line or a reorganisation. They require somebody named to be accountable for a small number of experiments, and a written record of what happened, kept where the next person can find it.

WHAT THIS PAPER ARGUES

Early advantage comes from faster organisational learning. It does not come from earlier deployment, and the two are easy to confuse in the first quarter of a technology's public life.

- Conversational access changed who can reach a large language model, and nothing else. Accuracy, provenance and accountability are where they were in October.
- Fluency tells the reader nothing about quality. The system's own documentation says it writes plausible-sounding but incorrect answers, and that identical questions phrased differently produce different results.[1] VENDOR Any control that depends on a reader noticing an error by reading is already weak.
- Tasks should be screened by the consequence of error rather than by enthusiasm. The filter proposed here scores value, data sensitivity, verifiability, reversibility and consequence, and it rejects any task failing the last three however attractive the first two look. AUTHOR
- A pilot without a baseline, a named owner and a written stop condition is a demonstration, and demonstrations are the cheapest thing in this field.
- Task time, rework, error type and incident count can be observed in 30 days. Return on investment cannot, and any figure offered for it this quarter is a projection wearing the clothes of a result.

SECTION 1

The interface breakthrough, and what it did not change

A capability that had existed in research settings for several years became reachable by anyone who could type. That is a distribution event, and the distinction matters for what follows.

On 30 November 2022 OpenAI made a conversational system available as a free research preview. Its own announcement describes the interaction rather than the model, saying that the dialogue format makes it possible for the system to answer follow-up questions, admit its mistakes, challenge incorrect premises and reject inappropriate requests.[1] VENDOR Within a week its chief executive said publicly that the preview had passed a million users.[2] VENDOR

Read carefully, none of that is a claim about capability. It is a claim about access. Large language models trained on internet-scale text had been described in the research literature for several years, and the technique that made this particular system feel cooperative, reinforcement learning from human feedback, had been applied to instruction following and published in March 2022.[3] PEER-REVIEWED What changed in November was that the capability stopped requiring an API key, a budget line and somebody who could write the calling code.

The distribution event is the enterprise event. An organisation does not get to decide whether to begin evaluating this technology. Its employees began on the day the preview opened, using personal accounts, on work material, without a policy telling them what was permitted. The open question is what the organisation learns from that.

What did not change

Three properties of the model class sit exactly where they were in October. The system produces text by predicting what text should come next, which means fluency is generated directly and correctness is generated incidentally. It carries no record of where any particular statement came from, so a claim cannot be traced to a source. And it has no mechanism for knowing that it does not know, which is why its errors arrive in the same confident register as its correct answers.

OpenAI states the first of these plainly in the announcement, saying the system sometimes writes plausible-sounding but incorrect or nonsensical answers, and that it is sensitive to tweaks to the input phrasing, so that given one phrasing of a question the model can claim not to know the answer while a slight rephrasing produces a correct one.[1] VENDOR A vendor documenting that its own product is unreliable in this specific way is worth more than a great deal of independent commentary, and it has been available to read since the first day.

WHAT CHANGED, AND WHAT DID NOT



FIGURE 1 The left column is a distribution event and the right column is the reason it still needs governing. Nothing in the second column moved on 30 November, which is why a capability that arrived overnight produces a governance problem that did not.

Left column, my reading. Right column, the properties described in the research literature. AUTHOR / PREPRINT

The risk literature had reached the same place earlier and with more structure. A 2021 paper from DeepMind set out six areas of risk from language models, running from discrimination and information hazards to misinformation harms and malicious use.[4] PREPRINT A 2021 paper widely known by its short title raised the environmental and documentation costs of ever larger training corpora, and the difficulty of auditing what is inside them.[5] PEER-REVIEWED None of this is new information in January 2023. It is simply information that had no operational consequence while the capability sat behind an API.

A CAUTION ABOUT THE PHRASE ARTIFICIAL INTELLIGENCE

This paper uses the phrase where a source uses it and avoids it elsewhere, in favour of naming the specific thing under discussion. The reason is practical. An organisation that has an artificial intelligence strategy tends to produce a document. An organisation that has a position on what may be done with a text-generating system, by whom, on what material, tends to produce a decision.

SECTION 2

A vocabulary precise enough to argue with

Most of the confusion in the current discussion is lexical. Six terms used consistently remove a surprising amount of it.

Terms that are used loosely in public will be used loosely in a board paper, and a board paper that says the organisation is deploying artificial intelligence has told the board nothing it can govern. The following six are sufficient for the decisions in this paper, and they are the terms the underlying literature uses.

1. A large language model is a statistical model trained to predict the next unit of text across a large corpus. Its output is a probability distribution, and what a user sees is one sample from it.
2. Generative means producing new output, as distinct from selecting it from a fixed set. The word describes the mechanism and carries no claim about originality or quality.
3. A prompt is the text supplied to the model, including instruction, context and any source material. It is an input to a statistical process, and the process treats it as evidence about what should come next.
4. An inference is a single run of the model over a prompt. Two identical prompts may yield different outputs, which is a property of sampling.
5. Training is the process that produced the model, and it was completed before the user arrived. A conversation does not train the model, though what a user types may be retained for other purposes depending on the terms of service.
6. Fine-tuning is further training of an existing model on additional data. It is a distinct activity with distinct costs, and it is frequently confused with prompting.

Terms to avoid for now

Two families of language are best left alone at this stage. The first attributes mental states, so that a model knows, believes, understands or decides. Each of those words imports an assumption about reliability that the system does not support, and they tend to survive into the risk register where they do real damage. The second is the vocabulary of imminent transformation, which is abundant this month and will date badly. A paper written in January 2023 about what an organisation should do in February should still be legible in two years, and the way to achieve that is to describe mechanisms, which age slowly.

> **ON THE WORD HALLUCINATION**
>
> It has become the standard term for a fabricated output and this paper uses it, because a term already in use beats an invented one. The metaphor is poor, implying a malfunction in an otherwise truthful process. The process has no notion of truth at any point, and the fabricated output is produced by exactly the mechanism that produces the correct one.

SECTION 3

The opportunity map as it stands in January

Six task classes can be attempted now, each under a stated condition. Two are named here specifically so that they can be refused.

The useful question is which tasks have a shape that makes an unreliable assistant safe to use, since what the technology can do remains unsettled and will stay that way for some time. That shape has three features. The person using the output already holds the knowledge needed to judge it, the cost of a bad output falls inside the organisation, and the work can be discarded without consequence.

THE JANUARY 2023 OPPORTUNITY MAP, WITH ITS CONDITIONS

TASK CLASS	WHAT THE SYSTEM IS DOING	THE CONDITION THAT MAKES IT DEFENSIBLE
Drafting from a brief	Producing a first version of text whose facts the writer already holds.	The writer supplies the facts and reads every line. The system is drafting, and it is not researching.
Summarising supplied text	Compressing a document the user has provided in the prompt.	The source sits in front of the reviewer, who can check the summary against it in the same minute.
Reframing and tone	Rewriting an approved message for a different audience.	The substance is already signed off, so the change under review is register rather than content.
Brainstorming and option generation	Producing a wide spread of candidate ideas for a person to select from.	Nothing is adopted without a human judgement. Volume is the output, and correctness is not claimed.
Classification against a rubric	Sorting supplied items into categories the user defines.	The rubric is written down first and a sample is checked by hand against it.
Coding assistance	Producing candidate code, tests or explanations of existing code.	Output runs against tests in an isolated environment before any human reads it as correct.
Anything about a named person	Generating or inferring statements concerning an identifiable individual.	Out of scope for this stage. There is no evaluation method available that would make it defensible.
Anything a customer sees unreviewed	Producing text that reaches an external party without a person in between.	Out of scope. The reversibility test fails, and the consequence of error sits outside the organisation.

The pattern across the first six rows is that the system is doing something with material the user has supplied, for a reader who can check it. The pattern across the last two is that neither condition holds. An organisation which cannot tell the difference between those two patterns will place its first pilot in the second group, because the second group is where the visible saving looks largest.

The case that looks safest and is not

Summarisation deserves a paragraph of its own, because it is the task most often proposed first and its failure mode is the quietest. A summary of a document the reader has not read cannot be checked by reading the summary, and the reader will not notice the absence of a point that was never mentioned. Errors of omission produce no artefact. Where a summary is used to decide whether to read the source, the control has been removed at precisely the moment it was needed, and this holds however good the model is, because it is a property of the workflow. AUTHOR

SPECIFICATION

Conditions under which summarisation is defensible in an early pilot

- The source document is supplied in the prompt, so the system is not retrieving anything and the scope of the material is known.
- The reviewer has the source open and is checking the summary against it.
- The output is used to locate what matters in the document, the reading still happens, and the pilot records whether it did.
- Omission is in the error taxonomy from the first day, because a count of factual errors will otherwise report a clean result on a summary that left out the only paragraph that mattered.

SECTION 4

A filter that screens on consequence

Five tests, applied in order, with the last three carrying a veto. The order is the part that does the work.

Most screening approaches in circulation this month score value and sensitivity, which produces a ranked list dominated by high-value tasks on non-sensitive data. That list is wrong in a specific and predictable way, because it omits the three tests that determine whether anyone will be able to tell that the pilot failed.

THE FIVE-PART USE-CASE FILTER



FIGURE 2 Value and data sensitivity are the tests organisations apply by instinct. Verifiability, reversibility and consequence of error are the tests that decide whether a task belongs in an early pilot at all.

My framework, read against the trustworthiness characteristics in the NIST framework. AUTHOR / PRIMARY

1. Value asks what the task is costing today in time, rework and delay. An estimate is acceptable at this stage provided it is written down before the pilot, because a remembered baseline flatters the tool by roughly the margin the pilot is trying to measure.
2. Data sensitivity asks what would enter the prompt. Treat anything a supplier's terms permit them to retain as disclosed, and read the terms themselves.
3. Verifiability asks whether the person receiving the output can tell whether it is right, using knowledge they already hold, in the time the task allows. A task failing this test cannot be piloted safely at any scale.

4. Reversibility asks what it costs to undo the output if it is wrong and has been acted upon. Reversible errors are a learning opportunity, and irreversible ones are an incident.
5. Consequence of error asks who bears the cost, and whether it falls outside the organisation. Where it does, the task waits.

Tests three, four and five each carry a veto, and that is the whole of the design. A task scoring highly on value and sensitivity but failing verifiability is exactly the task an enthusiastic team will propose, and exactly the task that will produce a result nobody can interpret. AUTHOR

> **WHERE THIS SITS AGAINST PUBLISHED FRAMEWORKS**
>
> The filter is my own and is offered as a working instrument. It is consistent with the framework the United States standards institute published on 26 January 2023, which organises risk work around four functions and describes trustworthy systems as valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.[6] PRIMARY That framework is voluntary, rights-preserving, non-sector-specific and use-case agnostic by its own description, which makes it a reference point. It supplies no procedure. The filter above is one attempt at the procedure.

One structural point about that framework is worth carrying, because it is commonly misread. Of its four functions, three describe activities performed on a system and the fourth describes the conditions under which those activities happen at all. Govern is cross-cutting rather than sequential, and an organisation that treats it as the first of four steps will complete it, move on, and discover later that nobody owns the result.[6] PRIMARY

SECTION 5

Designing a pilot that can fail visibly

Everything below is written down before access is granted, and in my experience the stop condition is what gets left out.

A pilot is an instrument for producing evidence. Most of what is currently described as a pilot is an instrument for producing enthusiasm, and the two are distinguishable in advance by whether anyone has written down what would count as a failure.

THE BOUNDED EXPERIMENT LOOP

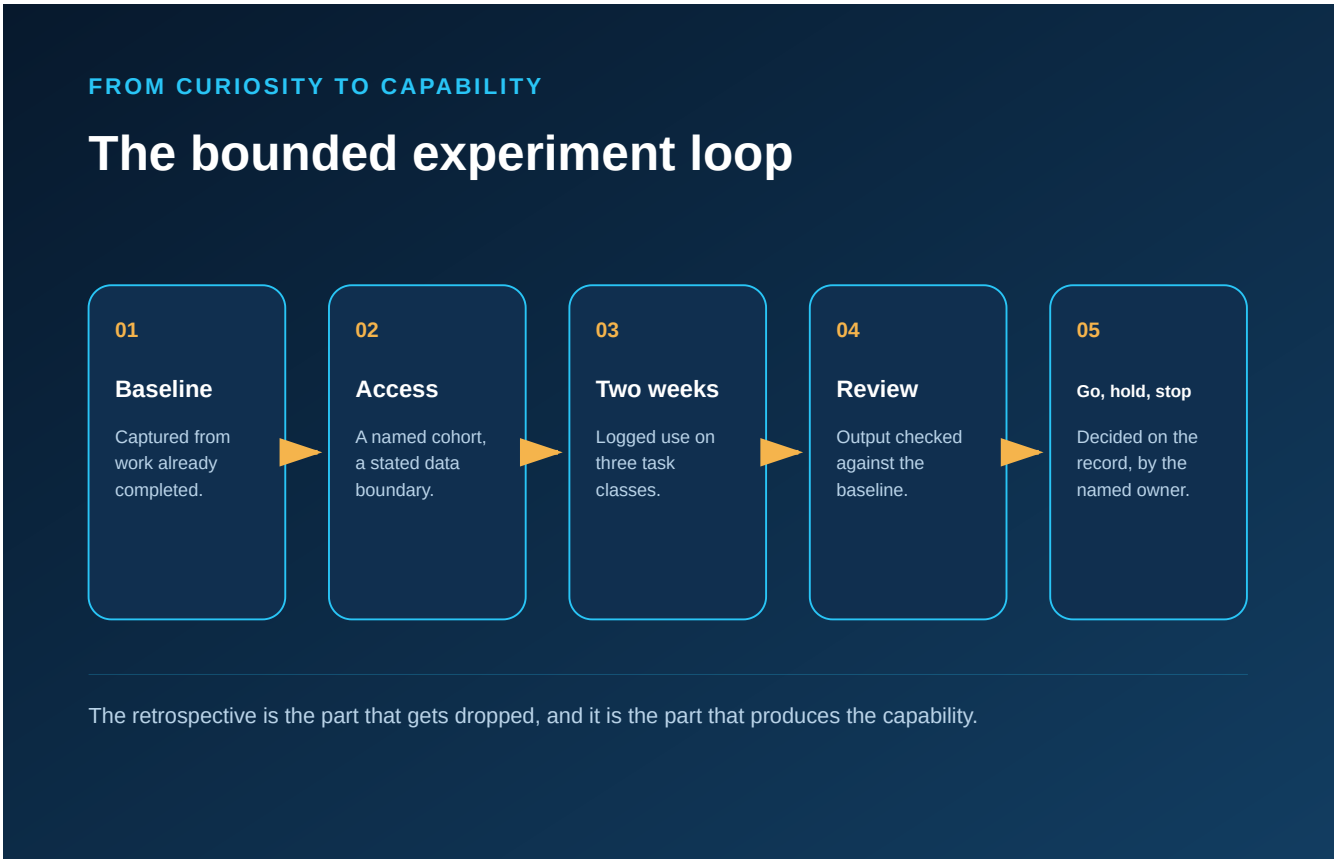


FIGURE 3 A two-week cycle with a baseline at the front and a go, hold or stop decision at the back. The retrospective is the part organisations drop, and it is the only part that produces transferable knowledge.

My framework. AUTHOR

SPECIFICATION

What to fix before the first person is given access

- A named owner, meaning one person rather than a committee or a function, who answers for the record at the end.
- An approved cohort of named people, small enough that the owner can speak to each of them in a week.
- A short list of what may go into a prompt, which beats a long list of what may not, because the long list will always be incomplete and the reader will infer permission from the absence of a prohibition.
- A baseline captured from work already completed, before access is granted. This is the element most often skipped and the one without which nothing else can be interpreted.
- A review method, stating who checks the output, against what, and how a disagreement is resolved.
- Stop conditions written before the pilot begins. Confidential material entering a prompt is the obvious one, and a rework rate above the baseline is the one people forget.
- A retrospective date two weeks out, in the calendar, with the owner present.

Why the stop condition matters more than it appears to

A pilot with no stop condition cannot produce a negative result, because there is no state of the world in which anyone declares it over. What happens instead is that it continues, quietly loses its baseline, acquires additional users who were never in the cohort, and eventually becomes a thing the organisation is doing. The test ended without anybody deciding it should. At that point the decision to deploy has been taken, by nobody in particular, on no evidence. I have watched this happen with three previous technology waves and the mechanism is identical each time. AUTHOR

A pilot that cannot end is not a pilot. It is an unannounced deployment with a flattering name.

THE AUTHOR

SECTION 6

What leaders should measure, and what they should refuse to

Five measures are available inside 30 days. Return on investment is not one of them, and saying so early is cheaper than retracting a number later.

The pressure to produce a financial figure will arrive before the evidence that would support one, and it will arrive from people who are entitled to ask. The answer worth giving is that the first month produces observations about work, the quarter after it produces a defensible estimate of cost, and anyone offering a return figure this month is extrapolating from a demonstration.

WHAT TO MEASURE IN THE FIRST 30 DAYS, AND WHAT TO REFUSE TO MEASURE

MEASURE	HOW IT IS CAPTURED	WHY IT SURVIVES SCRUTINY
Task time against baseline	The same task class timed before the pilot and during it, by the same people.	Observable inside two weeks and comparable across sites.
Rework rate	Proportion of outputs returned for a second pass before the work is accepted.	Catches the case where speed at the first draft is paid for later.
Error type, not error count	A short taxonomy filled in by the reviewer when something is wrong.	Counts tell you how often. Types tell you what to change.
Confidence against verified quality	User rates confidence before the reviewer checks the output.	The gap between the two is the number that predicts trouble at scale.
Incidents and near misses	Any confidential material entered, any output used unverified, any complaint.	Near misses arrive earlier than incidents and cost nothing to record.
Return on investment	Refused at this stage.	No baseline of sufficient length exists, and a 30-day figure would be an extrapolation.
Headcount equivalence	Refused at this stage.	It converts an unmeasured time saving into a financial claim, and it poisons the pilot politically.

The measure most likely to be informative

Of the five, the gap between user confidence and verified quality is the one I would watch hardest, and it is the one nobody collects. Ask the user how confident they are in the output before the reviewer sees it, then record what the reviewer found. Where confidence tracks quality, the workflow is self-correcting and can bear more volume. Where confidence runs ahead of quality, every additional user makes the problem larger, and the review step that was supposed to catch it will be the first thing dropped when the work gets busy. AUTHOR

The evaluation literature offers a useful discipline here, even though its subject is the model and ours is the workflow. A November 2022 project from Stanford evaluated 30 language models across 42 scenarios and seven metric categories, and its central methodological argument is that a single headline score conceals more than it reveals, because performance varies by scenario in ways that matter.[7] PREPRINT The same

holds inside an organisation. One number describing how well the pilot went will be true on average and useless everywhere.

A measurement problem nobody has solved yet

There is a difficulty underneath all five measures and this paper does not solve it. Quality in professional work is largely tacit, held by experienced people who can tell that a draft is wrong without being able to say which sentence is at fault, and a pilot that reduces quality to a rubric will measure the part of quality the rubric happens to capture. Every organisation I have watched attempt this has reached the same wall at roughly the same point, usually around week three, when the reviewers start disagreeing with each other and the disagreement turns out to be more interesting than the scores. AUTHOR

The practical response is to record the disagreements. Averaging them away destroys the finding. Where two experienced reviewers differ about whether an output was acceptable, that is a finding about the task, and it will still be a finding after the model has been replaced. Whether anybody has built a defensible quality measure for this kind of work by the end of 2023 is an open question, and I would not want to predict the answer from where the field sits this month.

SECTION 7

A 30-day charter

The charter runs for four weeks over three use cases with one cohort and ends in a decision. It fits on a page, and that is deliberate.

What follows assumes no budget, no procurement and no new supplier. It assumes a named owner with enough authority to stop the thing, and a policy notice that can be issued in a week. Where an organisation cannot manage those two, the problem is not technological and this paper will not help with it.

Week one, close the gap that is already open

Issue a short notice to all staff stating what may and may not be entered into a public system, in plain terms and without a compliance register. The notice exists because employees are already using these tools on work material and have been told nothing, which means the organisation currently has the risk without the visibility. Name the owner in the notice and give people a route to say what they have been doing, without consequence for having done it. That last clause is what determines whether the notice produces information or silence.

Week two, select three and baseline them

Run the five-part filter over the proposals and keep three task classes. Three is the right number because one produces an anecdote, two produce a comparison nobody trusts, and five exceed what a single owner can watch properly. Capture the baseline from work already completed, and resist the version of this step where somebody estimates the baseline from memory in a meeting.

Weeks three and four, run and log

Grant access to the named cohort only. Keep a daily log, which should take a user under a minute and should record the task, the time, whether the output was used, and what was wrong with it if anything. Hold a review at the end of each week with the owner present. The log is the deliverable of the whole exercise, and it will be worth more in six months than any conclusion drawn from it now.

> **THE DECISION AT DAY 30**
>
> Three outcomes are available and all three are successes. **Go** means one task class showed a measurable improvement with no unresolved incidents, and it earns a larger cohort and a longer baseline. It does not earn an enterprise licence. **Hold** means the evidence is ambiguous, which is the most likely result and the least embarrassing, and it earns another 30 days with a tightened method. **Stop** means a stop condition was met or the task class failed verifiability in practice, and it earns a written note of what was learned so that the next team does not repeat it.
>
> What is not available at day 30 is a decision to deploy across the organisation, because nothing that could be observed in 30 days would support one.

The organisations that will be in the strongest position by the middle of this year are not the ones that deployed earliest. They are the ones that can say, with a record to back it, which tasks in their own operation got better, by how much, and where the attempt failed. Few can say that today about any technology. That is why the capability is worth building on something small first.

END MATTER

Method, evidence and limits

How the sources in this paper were chosen and dated, and what the reader should distrust.

Where the sources come from

Sources were drawn from peer-reviewed literature, primary institutional documents and vendor announcements cited as vendor announcements. Vendor figures about vendor products are labelled as such. No consultancy estimate of market size or economic impact appears here, because none I could find rests on observed enterprise deployment.

Dating

Every reference carries its original publication date, and the title and address the source carried when I cited it. Pages on this subject are being revised week by week, and a reference that follows the revisions stops being a citation of anything in particular.

What I looked for and did not find

Three things were looked for and not found. No independent measurement of enterprise productivity effects from conversational language models, because the technology has been publicly available for nine weeks and no such study could exist yet. No published enterprise policy for the use of these systems on confidential material that I could find. And no evaluation method for output quality in professional work that an organisation could adopt without designing it, which is the gap this series will keep returning to.

The parts that are mine

The use-case filter, the bounded-pilot design and the confidence-against-quality measure are my own and are tagged wherever they appear. They are offered as instruments. None has been tested across enough organisations to be called validated. The reading of the January standards framework as cross-cutting is the issuer's own, paraphrased here from its text. The claim that organisations commonly misread it as the first of four steps is my own observation and rests on nothing beyond my own engagements.

A declaration of interest

I advise organisations on technology adoption, so several of the things this paper recommends are things I could be paid to help with. Against that, the paper argues against the enterprise-wide deployment that would be the easier thing to sell this quarter. Weigh it accordingly.

END MATTER

Sources considered and set aside

Six things I read while writing this and decided not to lean on, with the reason for each.

SOURCE	PUBLISHED	WHY IT WAS SET ASIDE
McKinsey, The state of AI in 2022	6 December 2022	A survey of adoption whose fieldwork closed months before the November release. It measures the previous generation of tools and says nothing about this one.
PwC, Sizing the prize	June 2017	A projection of $15.7 trillion of economic value by 2030 that is still quoted as though it were a measurement. It is a model, and the model predates every system discussed here.
GitHub, Research: quantifying GitHub Copilot's impact on developer productivity and happiness	7 September 2022	A vendor surveying users of its own product. Useful as a signal of satisfaction and useless as a measurement of productivity.
Brown and others, Language Models are Few-Shot Learners	28 May 2020	The paper describing the base model family. It is cited constantly in place of anything about the deployed system, which it does not describe.
Sequoia Capital, Generative AI: A Creative New World	19 September 2022	An investor's essay about a market, widely circulated. It says nothing about what an organisation should do on Monday, and it was partly written by the technology it describes, which its authors say and most people who quote it do not.
Stack Overflow, 2022 Developer Survey	June 2022	Fieldwork in May 2022. It records attitudes to tools that predate the one this paper is about.

END MATTER

About the author

Paul Forrest advises boards and executive teams on the adoption of new technology in operating businesses, with a practice grounded in delivery rather than in advisory work alone.

His background is in business process reengineering, which he has practised since the 1990s, and in applied natural language processing, which he has worked on since 2014. That combination informs the argument in this paper. The recurring failure in technology adoption is rarely the technology, and it is almost always the absence of a method for telling whether the change worked.

He writes on the condition that a claim carries its evidence, and this series is an attempt to hold that line while a subject moves quickly.

END MATTER

References

Seven works cited in the text, each with its original date of publication. None is later than 27 January 2023.

1. **OpenAI.** *ChatGPT: Optimizing Language Models for Dialogue.* Research preview announcement, published at openai.com/blog/chatgpt. Cited for the description of the dialogue format and for the stated limitations.
   PUBLISHED 30 NOVEMBER 2022

2. **Altman, S.** Public statement reporting that the research preview had passed one million users in its first week.
   PUBLISHED 5 DECEMBER 2022

3. **Ouyang, L., Wu, J., Jiang, X., and others.** *Training language models to follow instructions with human feedback.* arXiv:2203.02155, published in Advances in Neural Information Processing Systems 35, 2022. The method underlying instruction-following behaviour in the deployed system. Reinforcement learning from human feedback itself was published earlier, by Christiano and others in 2017.
   PUBLISHED 4 MARCH 2022

4. **Weidinger, L., Mellor, J., Rauh, M., and others.** *Ethical and social risks of harm from Language Models.* arXiv:2112.04359. A preprint at this date. Six risk areas and 21 specific risks.
   PUBLISHED 8 DECEMBER 2021

5. **Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S.** *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of FAccT 2021.
   PUBLISHED MARCH 2021

6. **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework 1.0.* NIST AI 100-1. Voluntary, rights-preserving, non-sector-specific and use-case agnostic, organised around Govern, Map, Measure and Manage.
   PUBLISHED 26 JANUARY 2023

7. **Liang, P., Bommasani, R., Lee, T., and others.** *Holistic Evaluation of Language Models.* arXiv:2211.09110. A preprint at this date, covering 30 models across 42 scenarios and seven metric categories.
   PUBLISHED 16 NOVEMBER 2022

END MATTER

Further sources consulted

Seven sources I read for this paper and did not cite. They are listed so that the reading behind it can be seen in full, and none is later than 27 January 2023.

- **Wei, J., Wang, X., Schuurmans, D., and others.** *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* arXiv:2201.11903, published in Advances in Neural Information Processing Systems 35, 2022.
  PUBLISHED 28 JANUARY 2022
- **Liu, P., Yuan, W., Fu, J., and others.** *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.* arXiv:2107.13586.
  PUBLISHED 28 JULY 2021
- **Bommasani, R., Hudson, D. A., Adeli, E., and others.** *On the Opportunities and Risks of Foundation Models.* arXiv:2108.07258.
  PUBLISHED 16 AUGUST 2021
- **Bai, Y., Kadavath, S., Kundu, S., and others.** *Constitutional AI: Harmlessness from AI Feedback.* arXiv:2212.08073.
  PUBLISHED 15 DECEMBER 2022
- **White House Office of Science and Technology Policy.** *Blueprint for an AI Bill of Rights.* Five principles for the design and deployment of automated systems.
  PUBLISHED OCTOBER 2022
- **European Commission.** Proposal for a Regulation laying down harmonised rules on artificial intelligence, COM(2021) 206 final. Cited as a legislative proposal, which is its status at the date of this paper.
  PUBLISHED 21 APRIL 2021
- **Council of the European Union.** General approach on the proposed artificial intelligence regulation.
  PUBLISHED 6 DECEMBER 2022

WHERE THIS GOES NEXT

The next paper takes up prompting as a workplace skill

This paper argued that the first useful capability is a learning system rather than a deployment. The second will take the next question in order, which is what the people running those experiments need to be able to do, and it is due at the end of March.

The policy for the series is that each paper is dated to the evidence available when it was written. If a later paper contradicts this one, this one will stay as published, because a record of what it was reasonable to believe at the time is worth more than a document quietly corrected.

THREE THINGS TO HAVE RUNNING BEFORE THE NEXT PAPER IS USEFUL

- A policy notice issued, naming an owner and a route for employees to report what they have already been doing.
- Three task classes selected through a written filter, with a baseline captured from completed work.
- A daily log running, however rough, because the log is what makes the next 30 days interpretable.

Paul Forrest · 31 January 2023 · Version 1.0 · Research closed 27 January 2023