PROMPT LITERACY AT WORK

A Practical Framework for Better Human–AI Collaboration

31 MARCH 2023 • ENTERPRISE AI SERIES



Paul Forrest

PAPER TWO · ENTERPRISE AI, 2023 TO 2024

About this paper

This is the second paper in a planned series of nine on the enterprise adoption of generative artificial intelligence.

The first paper argued that early advantage comes from organisational learning rather than early deployment, and it left a question open. If the useful unit of work is a bounded experiment run by named people, what do those people actually need to be able to do? This paper answers that.

Something like six parts in ten of this paper is a reading of published work. The remainder is my own, and I have tagged it so that nobody mistakes an instrument I designed for a finding somebody measured.

HOW TO CITE

> Forrest, P. (2023). *Prompt Literacy at Work: A Practical Framework for Better Human and Machine Collaboration.* Enterprise AI, 2023 to 2024, Paper Two. Version 1.0, 31 March 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.

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PREPRINT Research posted as a preprint or working paper, not peer-reviewed at this date.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 29 March 2023, and everything in it was current at that date. The literature on prompting is moving quickly enough that a reader six months from now should expect some of it to have been overtaken.

Contents

FRONT MATTER

Abstract and summary of the argument 4

THE PAPER

1 Why this became a workplace question in March 5

2 The anatomy of a prompt that can be checked 6

3 Iteration is the work, not a sign of failure 8

4 The verification ladder 9

5 The same skill, five different error costs 11

6 A rubric that assesses work rather than vocabulary 13

7 A four-week training sprint 15

END MATTER

Method, evidence and limits 16

Sources considered and set aside 17

About the author 18

References 19

Further sources consulted 20

ABSTRACT AND SUMMARY OF THE ARGUMENT

Half the skill is writing the prompt, and the other half is refusing to trust the answer

A substantially more capable model was released 17 days ago, and the practical question in most organisations changed shape.[1] VENDOR Until March the question was whether these systems were good enough to be worth learning. Few organisations are still asking it. What replaces it is harder, because it concerns people.

This paper defines prompt literacy as five things done in sequence. State a goal, supply the context the system cannot infer, constrain the task, specify the form of the output, and verify the result against something other than how the result reads. The first four are teachable in an afternoon. The fifth is where professional responsibility sits, and it is the part that no amount of prompting technique addresses.

The framing matters because the market has fixed on the phrase prompt engineering, which implies that the skill is technical and belongs to specialists. The failure I observe in these organisations is well-constructed prompts producing confident output that nobody checked, which is a question about accountability dressed up as a question about technique. AUTHOR

WHAT THIS PAPER ARGUES

Prompt literacy is a verification discipline with a drafting technique attached. Teaching the technique without the discipline produces faster production of work nobody has checked.

- The capability frontier moved on 14 March and the limitations did not. The vendor's own release states that the new model is still not fully reliable, that it hallucinates facts and makes reasoning errors, and it names social biases and adversarial prompts among known limitations.[1] VENDOR
- Availability is narrower than the coverage suggests. On release the model reached paying subscribers under a usage cap and a waitlisted developer interface, and image input was demonstrated with one partner and was not available to users.[1] VENDOR
- A prompt has seven parts, and most workplace prompts supply two. Goal and task are usually present. Context, source material, constraints, output form and the success test are usually absent, and their absence is what produces the output people then complain about. AUTHOR
- Verification has five rungs and the cost rises steeply. Surface review is nearly free and nearly worthless, while source confirmation, calculation, expert review and refusal each cost more, and the rung a task needs is set by the consequence of being wrong. AUTHOR
- Assessment should look at observed work rather than vocabulary. A four-level rubric that tests whether someone can produce a verified output under time pressure beats any written test of prompting terminology.

SECTION 1

Why this became a workplace question in March

A more capable model arrived on 14 March. What it changed was the range of tasks worth attempting, and what it left alone was every reason for checking the result.

OpenAI released GPT-4 on 14 March 2023, describing it as a large multimodal model accepting image and text inputs and emitting text outputs.[1] VENDOR The company reported that on internal adversarial factuality evaluations the model scored 40 per cent higher than its predecessor, that it was 82 per cent less likely to respond to requests for disallowed content, and that it responded to sensitive requests in accordance with company policy 29 per cent more often.[1] VENDOR

Those are the vendor's own internal evaluations and this paper treats them as such. The 40 per cent figure in particular deserves care, because it is a relative improvement on an adversarially constructed internal benchmark. It is not an accuracy rate, and an organisation that reads it as one will conclude that four in ten previous errors have gone away. Nobody has made that claim.

What was actually available on the day

Coverage of the release has been looser than the release itself. Text access on 14 March went to paying subscribers of the consumer product, under a usage cap, and to developers through a waitlist. The free tier of the consumer product continued to run the earlier model. Image input was demonstrated with a single accessibility partner and was not available to any user.[1] VENDOR An enterprise planning a March pilot on image understanding is planning around a capability it cannot obtain.

The limitations survived the upgrade intact. The release states that the model is still not fully reliable, that it hallucinates facts and makes reasoning errors, and it names social biases, hallucinations and adversarial prompts among its known limitations.[1] VENDOR The failure mode has become rarer and has not changed character, which is the worst combination for a human reviewer, because rare errors in fluent text are harder to catch than frequent ones.

One further piece of evidence landed three days later and bears on who this concerns. A study from OpenAI, OpenResearch and the University of Pennsylvania estimated that around 80 per cent of the United States workforce could have at least 10 per cent of their work tasks affected by language models, with roughly 19 per cent potentially seeing at least half of their tasks affected.[2] PREPRINT It is an exposure estimate built on task-level ratings, and the authors are explicit that it is not an observed outcome. Read as a forecast it is contestable. Read as an argument that this is a general workplace literacy, it is hard to dismiss.

SECTION 2

The anatomy of a prompt that can be checked

A working prompt has seven parts, and the two that people usually supply are the two that matter least.

The prompt is an instruction to a statistical process, which means it works by constraining what output is probable. It cannot specify what output is required. That distinction explains most of what frustrates new users. A vague prompt does not produce an error message. It produces a plausible answer to a question the user did not ask.

THE SEVEN PARTS OF A WORKING PROMPT



FIGURE 1 Most workplace prompts supply the goal and the task and stop there. The five remaining parts are what convert a plausible answer into one that can be checked.

My framework. AUTHOR

1. The goal is what the output is for, stated in terms of what the reader will do with it. The difference between write a summary and write a summary a credit committee will use to decide whether to read the full paper is the whole of the useful instruction.
2. Context is what the system cannot infer, which means audience, purpose, prior decisions, house conventions, and anything that would be obvious to a colleague and is invisible here.

3. Source material is the text the work is to be performed on, supplied in the prompt. Where the material is not supplied, the system is drawing on training data of unknown provenance, and the output cannot be verified against anything.
4. Constraints cover length, register, what to exclude, what must be preserved verbatim, and what the output must not claim.
5. Examples, meaning one or two worked instances of the form wanted, are the cheapest single improvement available and the one least often used.
6. Output form covers structure, headings, field names, and whether the answer should mark its own uncertainty. A request to flag any statement the system is unsure of is imperfect and still better than nothing.
7. The success test is what the user will check before using the result, and it is written before the prompt is sent, because a test designed after reading a persuasive output is not a test.

The seventh part is the subject of this paper, and it has nothing to do with prompting. Writing down what would make the answer acceptable, before seeing an answer, is the discipline that keeps a fluent output from setting its own standard of adequacy.

> **ON THE PHRASE PROMPT ENGINEERING**
>
> The term has stuck. It carries an unhelpful implication, which is that the work is technical and therefore somebody else's. In every organisation I have watched attempt this, the limiting factor has been domain judgement, and domain judgement cannot be centralised into a prompt team.

SECTION 3

Iteration is the work, not a sign of failure

The first output is a draft of the question. Treating it as a draft of the answer is the most common error in the first month of use.

Users who report that these systems are unreliable are frequently describing a single exchange. They asked once, read the answer, found a fault, and concluded. Users who report that the systems are useful are almost always describing a sequence, and the sequence has a shape worth teaching.

Decomposition

A task that fails as one instruction often succeeds as four. Ask for an outline, correct the outline, ask for one section against the corrected outline, then ask for the next. The gain comes from the human staying in the loop at each junction, which is where the errors are still cheap to catch.

The research literature supports the mechanism from a different direction. Prompting a model to produce its intermediate reasoning steps improves performance on multi-step problems, a result published in January 2022.[3] PEER-REVIEWED Sampling several reasoning paths and taking the most consistent answer improves it further.[4] PREPRINT Both findings are about model behaviour. The workplace lesson sits beside them, and I would sooner say that plainly than borrow more authority than the studies offer. AUTHOR

Asking for uncertainty

Asking the system which parts of its answer it is least confident about produces useful output surprisingly often, and it should be understood for what it is. The response is generated by the same process that generated the answer, so it is a statement about the text and carries no calibrated probability. It is a cheap way to direct attention. It is no substitute for checking.

Keeping the trail

Where an output contributes to a decision that anybody may later question, the prompt and the verification should be recoverable. This sounds bureaucratic and takes about 20 seconds, and the organisations that skip it discover the cost at the first challenge, when nobody can reconstruct what was asked or what was checked.

SECTION 4

The verification ladder

Five rungs, rising sharply in cost. The rung is chosen by what happens if the output is wrong, and it is chosen before the output exists.

Verification is the half of prompt literacy that training programmes leave out, partly because it is duller and partly because it cannot be demonstrated in a keynote. It is also the half that determines whether the organisation is safer or more exposed at the end of the year.

THE VERIFICATION LADDER



FIGURE 2 Five rungs in ascending order of cost and confidence. The rung a task requires is determined by what happens if the output is wrong, and it is set before the work starts, while the judgement is still cold.

My framework. AUTHOR

1. Surface review means reading it. That catches tone, obvious contradiction and format errors, and it catches essentially no factual errors, because a fabricated fact reads exactly like a correct one. It is adequate for internal drafting where the drafter holds the facts.
2. Source confirmation checks each material claim against a document. It is slow, it is the only rung that addresses fabrication directly, and it is required before any external use.
3. Calculation or test means re-running the arithmetic, executing the code or reproducing the query. It is available only where the task has a mechanical check, and where it is available it is the best value on the ladder.

4. Expert review has a person with independent domain knowledge read for the errors the previous rungs cannot see. It is expensive, and the cost is the point, since it forces an explicit decision about which work deserves it.
5. Refusal means the output is not used. It is a rung rather than a failure, and a team that has never used it is approving rather than verifying.

Two design rules govern the ladder. The rung is set by the consequence of error rather than by the difficulty of the task, which means a short paragraph going to a regulator sits higher than a long internal analysis. And the rung is set before the work begins, because a persuasive draft lowers the verification standard of the person holding it. AUTHOR

> **THE FAILURE MODE THIS LADDER EXISTS TO PREVENT**
>
> A workflow in which rejecting the output means redoing the work by hand will produce agreement regardless of quality. The reviewer is responding rationally to a design in which the cost of saying no falls entirely on them.
>
> Where that condition holds, add time to the task. Training the reviewer will not help. No amount of instruction in critical appraisal will overcome an incentive structure that punishes appraisal.

A useful contemporary reference point for the mechanical rung comes from software. A controlled study published in February 2023 reported that developers using an AI pair programmer completed a specific task around 56 per cent faster.[5] PREPRINT / VENDOR The number travels widely and the study's own limits travel less. The study recruited 95 developers and 70 completed, the task was a single synthetic one, and the confidence interval around the headline runs from roughly 21 to 89 per cent. It is a real result about a narrow thing, and it is not a productivity figure for software engineering.

SECTION 5

The same skill, five different error costs

Prompt literacy is general. What is not general is what it costs to be wrong, and the training has to be built around the second thing.

A single enterprise training programme teaching everyone the same prompting technique will be efficient to deliver and will miss the point in most of the roles it reaches. The technique is nearly identical across roles. The verification rung, the escalation route and the definition of an unacceptable error are not.

WHERE THE ERROR COST FALLS, BY ROLE

ROLE	TYPICAL FIRST USE	WHAT THE ERROR ACTUALLY COSTS
Communications and marketing	Drafting, reframing, adapting an approved message for a channel.	Reputational and external. The error leaves the building, and correction is public. Verification is editorial and the reviewer must be senior enough to say no.
Analysis and research	Summarising supplied material, generating candidate explanations.	Compounding and invisible. A wrong premise propagates into a conclusion that looks sound, and the error surfaces at the decision rather than at the draft.
Software engineering	Generating candidate code, tests, and explanations of unfamiliar code.	Usually cheap, because the code either runs or does not. The dangerous case is code that runs and is subtly wrong, which is where tests earn their cost.
Customer support	Drafting responses, summarising case history, suggesting resolutions.	Immediate and contractual. An invented policy stated confidently to a customer becomes a commitment the organisation may have to honour.
Management and planning	Structuring documents, drafting papers, generating options.	Slow and structural. Fluent strategy documents that say nothing specific are already common, and a system that produces them faster makes the problem worse.
Legal, clinical, financial advice	Out of scope at this stage.	The consequence sits outside the organisation and the verification cost approaches the cost of doing the work unaided. No published evidence supports assisted practice here.

The row worth dwelling on is management. Fluent documents that commit to nothing were already a feature of corporate life before any of this, and a system that produces them in seconds removes the one natural brake, which was that writing something vacuous at length used to be tiring. I am not aware of any evidence on this, and I raise it as a practitioner's concern. AUTHOR

Where the evidence on assisted work currently sits

Two studies published in the first quarter give a partial picture, and both concern narrow tasks. The software study above is one. The other, circulated as a working paper in early March, tested mid-level professional writing tasks and reported faster completion and higher rated quality, with the largest gains among the initially weaker writers.[6] PREPRINT It is a working paper at the date of this publication and should be cited as one.

Neither study measured what happens when the assisted output is wrong and nobody notices. That is the question an enterprise needs answered. That absence is not a criticism of either piece of work. It is a

statement about what the field has measured so far, and it is the reason this paper spends more space on verification than on technique.

SECTION 6

A rubric that assesses work rather than vocabulary

The rubric has four levels, and each is defined by something a manager can watch somebody do.

Assessment in this area has drifted quickly towards quizzes about terminology, which are cheap to build and measure nothing anybody cares about. Somebody who can define few-shot prompting and cannot tell that an output is wrong is a liability, and the assessment should be capable of saying so.

FOUR LEVELS OF PROMPT LITERACY

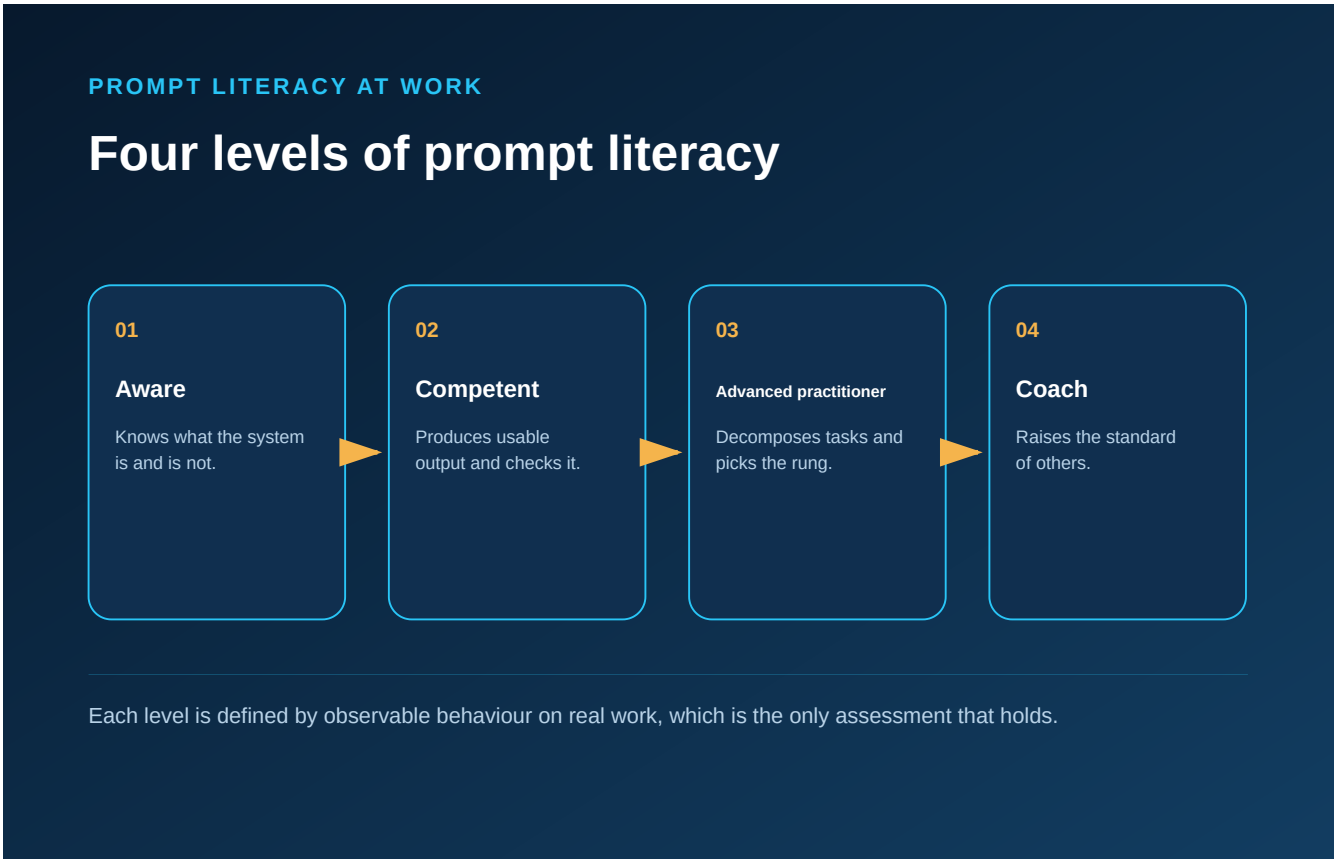


FIGURE 3 Progression from aware to coach. Each level is defined by observable behaviour on real work, because no other assessment method survives contact with an organisation.

My framework. AUTHOR

THE FOUR-LEVEL RUBRIC, ASSESSED ON OBSERVED WORK

LEVEL	WHAT THE PERSON CAN DO	HOW IT IS OBSERVED
Aware	Uses the system for low-consequence drafting. Knows what may not be entered into a prompt and knows that output can be confidently wrong.	Can state the data rule without looking it up, and can give one example of an output they rejected and why.
Competent	Supplies context, constraints and an output form. Verifies against a source before use. Escalates rather than guessing when out of depth.	Watched producing one real deliverable end to end, including the verification step, within the normal time allowed for the task.
Advanced practitioner	Decomposes a task, compares approaches, recognises where the system is unreliable in their own domain, and designs the check before the prompt.	Can predict in advance which parts of a given task the system will get wrong, and is right more often than not.
Coach	Teaches the above, reviews other people's work, maintains the domain examples, and is trusted to say that a use case should be refused.	Has taught at least one cohort and has stopped at least one proposed use, with a written reason that survived challenge.

The progression is deliberately slow at the top. Moving from competent to advanced practitioner requires enough exposure to a domain to predict where the system fails in that domain specifically, and that knowledge cannot be transferred by training. It accumulates by keeping records of what went wrong, so the log from the first paper matters more than it looked at the time.

What the coach level is for

Every organisation attempting this will designate champions, and most will define the role as advocacy. Defined that way it is worse than useless, because it puts the enthusiasts in charge of the appraisal. The coach level above is defined by the authority to refuse a use case and the obligation to write down why. It is a different job. AUTHOR

SECTION 7

A four-week training sprint

Baseline, guided practice, adversarial examples, assessment. The adversarial week is the week that gets cut, and it is the week that works.

What follows assumes a cohort of 12 to 20 from a single function, an internal facilitator, and real work. Mixed cohorts drawn from across the organisation are easier to fill and teach nothing transferable, because the error costs differ by function and the discussion never gets past technique.

Week one, baseline on real work

Each participant completes one representative task unassisted and timed, with the output kept. It takes an hour. That hour is the comparison for everything that follows, and teams that skip it spend the rest of the sprint arguing about whether anything improved.

Week two, guided practice against the seven parts

Work the same class of task with the full prompt anatomy and a nominated verification rung. The facilitator's job is to insist on the success test being written before the prompt is sent, which participants will resist because it feels like bureaucracy and is in fact the entire lesson.

Week three, adversarial examples

Give participants outputs that are wrong in ways characteristic of their domain, without telling them which ones. Include an output with a fabricated citation, one with a subtly wrong figure carried correctly through the arithmetic, and one that is accurate but has omitted the qualification that changes the conclusion. Ask them to mark which are safe to use.

Most participants catch the fabricated citation. Far fewer catch the omission, and the discussion that follows is the most valuable hour in the sprint, because it is where people discover that their verification habit was tuned to a failure mode the system does not have. AUTHOR

Week four, assessment and refresh

Assess against the rubric on observed work. Record who reached competent, who reached advanced, and who should not be using these systems on consequential material yet. That is a legitimate outcome and should be said out loud. Set a refresh date, because the models change and a rubric calibrated to March will describe a different system by the autumn.

> **WHAT THE SPRINT DOES NOT ACHIEVE**
>
> Four weeks produces competence at the second level of the rubric across a cohort, and it produces the beginnings of a domain-specific error catalogue. It does not produce advanced practitioners, because that level requires months of exposure to the same task class.
>
> It also does not produce a measured productivity gain, and a sprint sold to an executive sponsor on that basis will disappoint. The honest claim is that it produces people who can tell when the output is wrong, which is the precondition for any gain being real.

END MATTER

Method, evidence and limits

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

Sources

Sources are peer-reviewed studies, preprints labelled as preprints, and vendor announcements cited as vendor announcements. Where a figure originates in a vendor's internal evaluation, the text says so and does not convert it into an independent finding.

Dating

Every reference carries its original publication date. The literature on prompting and verification is growing week by week, and much of what a reader finds later will postdate this paper.

A note on the March studies

Several of the studies cited are preprints or working papers. They carry the preprint tag, which records their status in March and says nothing about their quality. A reader who wishes to discount them on that basis is entitled to, and the text does not lean on any one of them alone.

Gaps

No study measures the effect of verification practice on error rates in assisted professional work. No published rubric for assessing prompt literacy on observed work could be located. And no evidence exists either way on whether assisted drafting degrades the quality of thinking in document-heavy roles, which is the concern raised in section five and left there deliberately.

Where to be sceptical

The seven-part prompt anatomy, the verification ladder, the four-level rubric and the error-cost table are my own. They are tagged wherever they appear and none has been validated beyond the engagements it came from.

My interest in this

I deliver training of the kind described in section seven, which is a reason to read the recommendation sceptically. The paper argues for smaller programmes than the ones that are commercially easiest to sell, which is at least a point in its favour.

END MATTER

Sources considered and set aside

What I read and put down again, with the reason in each case.

SOURCE	PUBLISHED	WHY IT WAS SET ASIDE
Microsoft, Microsoft 365 Copilot announcement	16 March 2023	A demonstration with no availability date and no figures. There is nothing in it to cite yet beyond the fact of the announcement.
Google, Bard announcement	6 February 2023	Announced in February and opened to a waitlist on 21 March. No published evaluation and no enterprise terms at this date.
Bubeck and others, Sparks of Artificial General Intelligence	22 March 2023	An early evaluation of the new model by researchers at the company that funds its developer. It is qualitative, it is a preprint, and its title has done more work than its contents.
Goldman Sachs, The potentially large effects of artificial intelligence on economic growth	26 March 2023	A projection that generative tools could expose the equivalent of 300 million full-time jobs to automation. It is a model of exposure, circulated in the last week, and the press has cited it as though it were a finding.
Anthropic, Introducing Claude	14 March 2023	A vendor announcement of a competing system, with no figures a reader could check.
OpenAI, GPT-4 Technical Report	15 March 2023	Cited elsewhere as though it documented the model. It says in terms that it withholds architecture, training data and compute, so the evaluation figures in it cannot be reproduced by anyone outside the company.

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. The argument in this paper reflects both. Techniques for getting more out of a tool are the easy half of any adoption programme, and the half that determines the outcome is whether anybody can tell that the work got worse.

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

Six works cited in the text, each with its original date of publication. None is later than 29 March 2023.

1. **OpenAI.** *GPT-4.* Release announcement. Cited for the multimodal description, the internal evaluation figures of 40 per cent, 82 per cent and 29 per cent, the availability position on release, and the stated limitations.
   PUBLISHED 14 MARCH 2023

2. **Eloundou, T., Manning, S., Mishkin, P., and Rock, D.** *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.* arXiv:2303.10130. Task-level exposure estimates, not observed outcomes.
   PUBLISHED 17 MARCH 2023

3. **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

4. **Wang, X., Wei, J., Schuurmans, D., and others.** *Self-Consistency Improves Chain of Thought Reasoning in Language Models.* arXiv:2203.11171. Accepted for the 2023 learning representations conference, whose proceedings appear after this date.
   PUBLISHED 21 MARCH 2022

5. **Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M.** *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.* arXiv:2302.06590. A study of 95 recruited developers, 70 completed, one synthetic task, confidence interval from roughly 21 to 89 per cent.
   PUBLISHED 13 FEBRUARY 2023

6. **Noy, S., and Zhang, W.** *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.* Working paper. Cited in its working paper form, which is its status at the date of this publication.
   PUBLISHED 2 MARCH 2023

END MATTER

Further sources consulted

Nine 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 29 March 2023.

- **Kojima, T., Gu, S. S., Reid, M., and others.** *Large Language Models are Zero-Shot Reasoners.* arXiv:2205.11916.
  PUBLISHED 24 MAY 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
- **Ouyang, L., Wu, J., Jiang, X., and others.** *Training language models to follow instructions with human feedback.* arXiv:2203.02155.
  PUBLISHED 4 MARCH 2022
- **Liang, P., Bommasani, R., Lee, T., and others.** *Holistic Evaluation of Language Models.* arXiv:2211.09110.
  PUBLISHED 16 NOVEMBER 2022
- **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework 1.0.* NIST AI 100-1.
  PUBLISHED 26 JANUARY 2023
- **International Organization for Standardization and International Electrotechnical Commission.** *ISO/IEC 23894, Information technology, Artificial intelligence, Guidance on risk management.*
  PUBLISHED 6 FEBRUARY 2023
- **Bubeck, S., Chandrasekaran, V., Eldan, R., and others.** *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* arXiv:2303.12712. A preprint, not peer reviewed, and cited here only for its documentation of capability boundaries.
  PUBLISHED 22 MARCH 2023
- **Department for Science, Innovation and Technology.** *A pro-innovation approach to AI regulation.* Command paper.
  PUBLISHED 29 MARCH 2023
- **OpenAI.** *GPT-4 Technical Report.* arXiv:2303.08774. Published one day after the release announcement and cited separately from it.
  PUBLISHED 15 MARCH 2023

WHERE THIS GOES NEXT

The next paper moves from the individual to the workforce

This paper treated prompt literacy as something a person has. The third will treat capability as something an organisation designs, across the different populations inside it, and it is due at the end of May.

Employer survey evidence on these tools is thin at the moment, and I expect that to change during the year. Until it does, the honest position is the one taken here, which is that the workforce question is being answered from anecdote.

WHAT THE NEXT PAPER WILL ASSUME IS ALREADY IN PLACE

- One cohort taken through a four-week sprint, with baselines kept and the adversarial week actually run.
- A short domain error catalogue, recording the ways these systems fail on your work rather than in general.
- A named person with the authority to refuse a use case, and at least one instance of them having used it.

Paul Forrest · 31 March 2023 · Version 1.0 · Research closed 29 March 2023