INTELLIGENCE · ARTIFICIAL AND COLLECTIVE INTELLIGENCE

# Large Language Models as Research and Intelligence Assistants

Large language models can summarise documents, draft reports and answer questions in seconds. This article explains how they work, what the research shows about their effect on knowledge work, where they fail, and how intelligence and research teams can use them safely.

ModQu research article · 2,195 words · 22 references (APA 7th edition)

Within a few years, **large language models** (LLMs) have moved from research laboratories into everyday office software. They can summarise a long report, translate a foreign news article, draft a briefing note, suggest survey questions or answer a question in conversational language. For research and intelligence teams, whose work consists largely of reading, organising and writing, the potential is obvious. So are the risks. A model that writes fluently can be wrong in ways that are hard to detect, and an analyst who relies on it without checking can pass errors straight into a decision.

This article explains what these models are, what the evidence says about their effect on professional work, where they fail, and how to build them into a research process without lowering its standards.

## How language models work

A language model is a system trained to predict the next word, or more precisely the next fragment of text, given the text that comes before it. Modern models are built on the **transformer** architecture introduced by Vaswani et al. (2017), which lets the model weigh the relevance of every earlier word when predicting the next one. Devlin et al. (2019) showed that a transformer pre-trained on large amounts of text could be adapted to many language tasks, from classifying sentences to answering questions, with little additional training.

The step to today's general-purpose assistants came with scale. Brown et al. (2020) showed that a very large model, trained on hundreds of billions of words, could perform new tasks when given only a few examples in its prompt, without any further training. Ouyang et al. (2022) then described how models can be fine-tuned with human feedback to follow instructions and give answers people judge helpful. This combination of scale and instruction-tuning produced the chat assistants now widely used.

Two features of this design matter for anyone using a model in research. First, the model has no direct access to facts. It has learned statistical patterns in text, and it produces text that is plausible given those patterns. Most of the time, plausible text is also accurate, because accurate statements are common in the training data. But plausibility and truth are not the same thing. Second, the model's knowledge stops at the date its training data were collected, unless it is connected to other sources.

Bender et al. (2021) warned early that models which produce fluent language without grounding in meaning invite people to attribute understanding and authority to them that they do not have. That warning is the starting point for using them well.

## What the evidence shows about productivity

Several careful studies have measured the effect of language models on real professional tasks.

Noy and Zhang (2023) ran an experiment with college-educated professionals writing realistic documents such as press releases and short reports. Those given access to a chat assistant finished faster and produced work that independent evaluators rated higher. The largest gains went to the weaker writers, narrowing the gap between workers.

Brynjolfsson, Li and Raymond (2023) studied customer support agents at a large software company as an AI assistant was introduced. Productivity, measured by issues resolved per hour, rose on average by about 14%, with the biggest gains for new and less skilled agents. The assistant appeared to spread the practices of the best agents to the rest.

The most relevant study for analysts is by Dell'Acqua et al. (2023), who worked with management consultants at a large firm. For tasks within the model's capabilities, consultants using it completed more work, faster and at higher quality. But for a task deliberately designed to fall just outside those capabilities, a business problem requiring careful reading of interview notes and spreadsheet data, consultants using the model were markedly *less* likely to reach the correct answer than those working without it. The authors describe a "jagged technological frontier": the boundary between what models do well and badly is irregular and hard to see, and people tend to trust the model on both sides of it.

For intelligence and research teams, the lesson is that gains are real but uneven, and the costs fall precisely on the hardest and most important analytical tasks unless people know where the frontier lies.

## Where models fail

### Hallucination

The best-known failure is **hallucination**: the production of statements that are fluent and confident but false or unsupported. Ji et al. (2023) review the research and distinguish two kinds. *Intrinsic* hallucinations contradict the source the model was given, for example a summary that reverses a figure in the report. *Extrinsic* hallucinations add information that cannot be verified from the source at all. Maynez et al. (2020) found that a large share of machine-written summaries in a standard benchmark contained material not supported by the original article.

Invented references are a well-documented case. Models asked for sources often produce citations with plausible authors, titles and journals that do not exist. In a widely reported American case, lawyers submitted a court filing that cited several non-existent judicial decisions generated by a chat assistant, and the court sanctioned them (Mata v. Avianca, Inc., 2023). The same failure in an intelligence report would be harder to catch, because readers rarely check the sources behind each claim.

### Overconfidence

Models do not reliably signal when they are unsure. Guo et al. (2017) showed that modern neural networks tend to be poorly **calibrated**: the confidence they express does not match how often they are right. A model writes a wrong answer in the same assured tone as a right one. Analysts trained to express uncertainty carefully need to supply that discipline themselves.

### Bias and simulated opinion

Because models learn from large collections of text, they reproduce the views, gaps and biases of that text. Argyle et al. (2023) showed that, when prompted with demographic descriptions, a language model could produce survey-like answers that matched some patterns in real American opinion data. This has prompted interest in using models to "simulate" respondents. It is a useful tool for drafting and testing questions, but it is not a substitute for asking real people: the model reflects what has been written, which differs from what a defined population currently thinks, and its errors are hard to detect without the real data to compare against.

### Reasoning errors

Wei et al. (2022) showed that prompting a model to set out its reasoning step by step improves its performance on multi-step problems. But a written chain of reasoning can still contain errors, and a well-structured explanation can make a wrong conclusion more persuasive. Analysts should read the steps, not only the conclusion.

## Human factors

Many of the risks come from how people respond to the tool. Parasuraman and Riley (1997) described **misuse** of automation: relying on it in conditions where it is unreliable because it is usually right. Logg, Minson and Moore (2019) found that people often give more weight to advice they believe comes from an algorithm than to the same advice from a person. Combined with fluent, confident text, these tendencies make uncritical acceptance likely.

Heuer (1999) showed that analysts tend to search for and accept information that fits their existing view. A model that responds to the framing of a question can amplify this. Ask it why a competitor's launch will fail and it will produce reasons; ask why it will succeed and it will produce those too. The model reflects the question back. Analysts who use it to test their views should ask it for the strongest case *against* their current judgement.

## Grounding the model in sources

The most effective technical remedy for factual errors is to give the model the right documents and require it to work from them. Lewis et al. (2020) introduced **retrieval-augmented generation**, in which a system first retrieves relevant passages from a defined collection and then generates an answer based on them. Shuster et al. (2021) found that retrieval substantially reduced hallucination in conversational systems. In practice, this means an analyst supplies the reports, transcripts or survey results that matter, asks the model to answer only from them, and requires it to cite the passage behind each statement. Every citation can then be checked.

Grounding does not eliminate errors. Models can still misread or misquote a source. But it changes the analyst's task from checking claims against the whole world to checking them against specific documents, which is far more practical.

## Choosing suitable tasks

Two questions help decide whether a task suits a language model: how costly would an error be, and how easily can the output be checked? Figure 1 places common research tasks against these two dimensions.

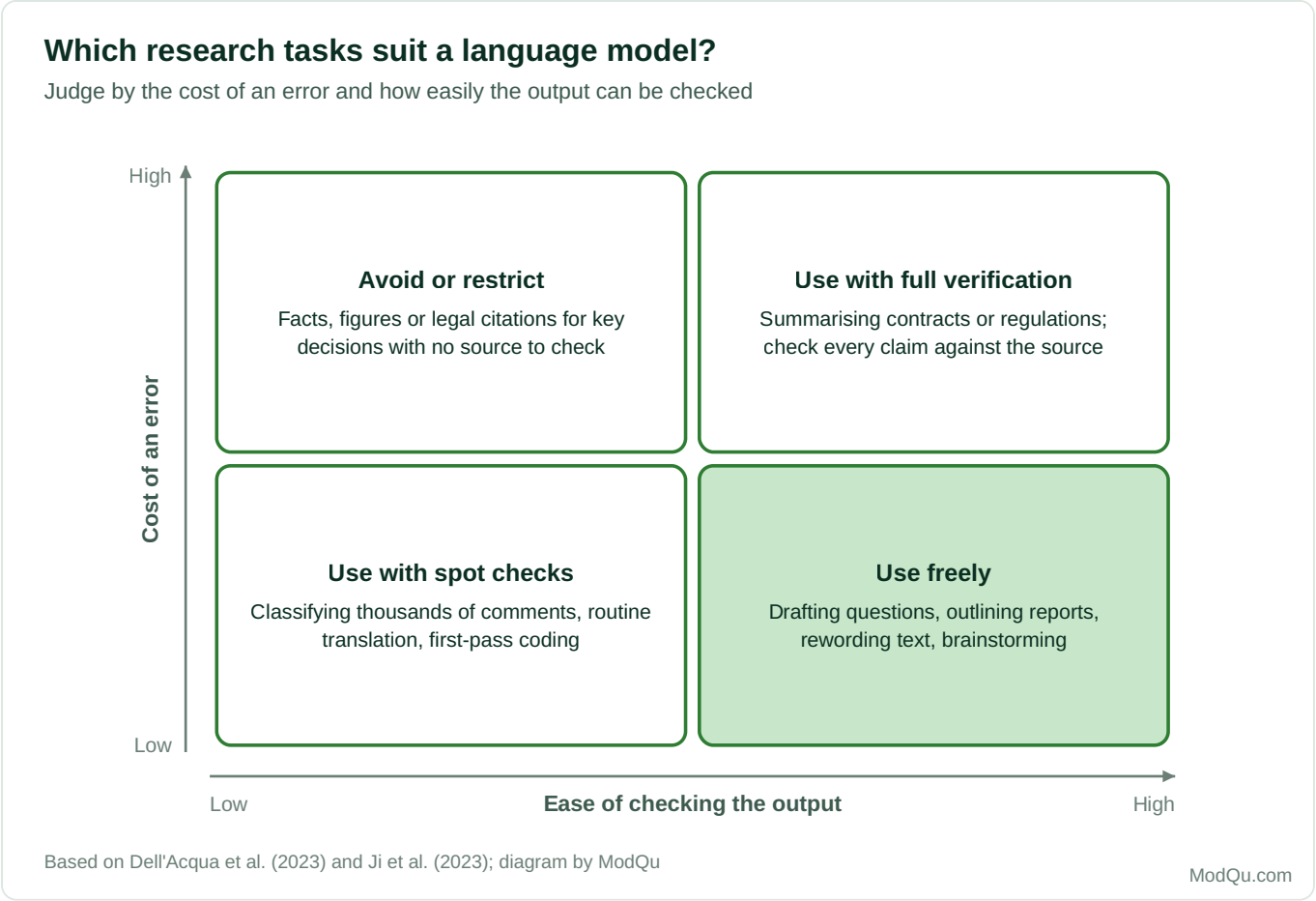



Figure 1. Which research tasks suit a language model? Tasks where errors are cheap and easy to see can be delegated freely; tasks where errors are costly and hard to detect should not rely on a model. Based on Dell'Acqua et al. (2023) and Ji et al. (2023); diagram by ModQu.

Drafting and rewording are low-risk: the analyst reads the output and errors are visible. Summarising a long contract is useful but must be checked line by line against the original, because an omitted clause is invisible in the summary. Classifying thousands of open-ended survey comments is efficient, and a sample can be checked by hand to estimate the error rate. Asking a model for facts or figures with no source to check them against is the riskiest use and should be avoided for anything that informs a decision.

## A safe workflow

Figure 2 sets out a workflow that captures most of the benefits while controlling the main risks.

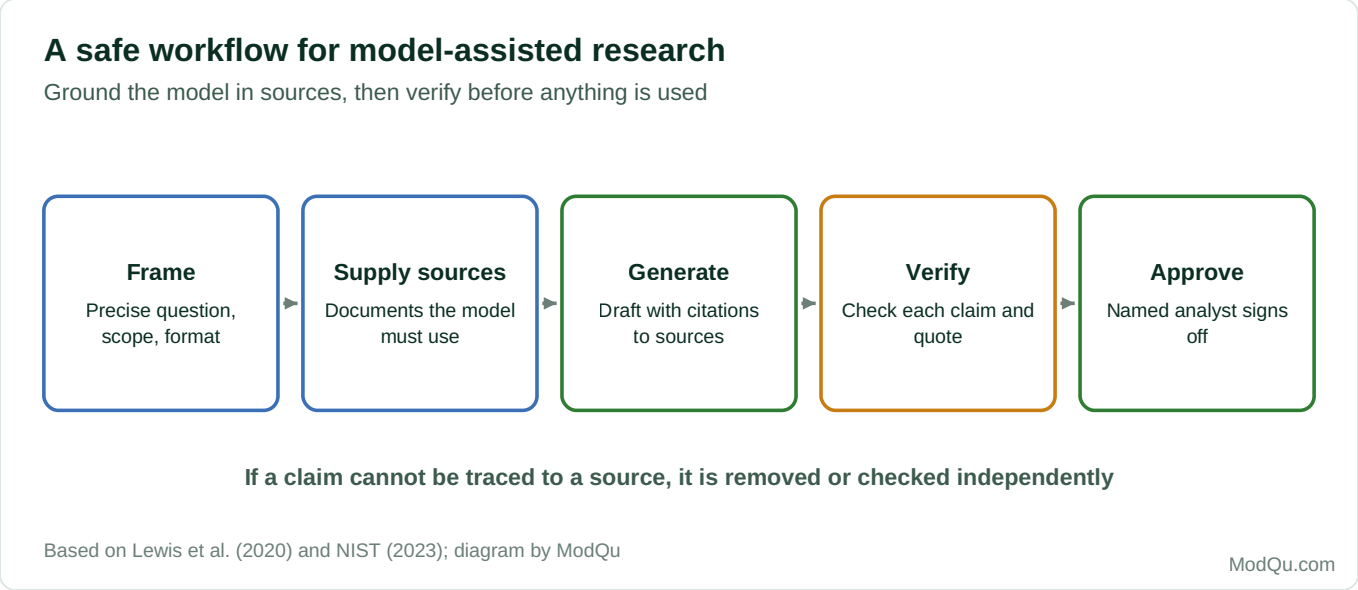



Figure 2. A safe workflow for model-assisted research. Grounding the model in specified sources makes verification practical, and a named analyst remains accountable. Based on Lewis et al. (2020) and NIST (2023); diagram by ModQu.

1. **Frame the question precisely.** State the scope, the time period, the audience and the format. Vague prompts produce vague answers.
2. **Supply the sources.** Provide the documents the answer should be based on and instruct the model to use only those.
3. **Generate with citations.** Ask for each claim to be linked to the passage that supports it.
4. **Verify.** Check every factual claim, figure and quotation against the source. Remove anything that cannot be traced.
5. **Approve.** A named analyst reviews the final product and takes responsibility for it.

This mirrors the guidance of the American National Institute of Standards and Technology, whose AI risk management framework emphasises mapping where a system is used, measuring its performance, and keeping clear human accountability (National Institute of Standards and Technology [NIST], 2023). The Office of the Director of National Intelligence (2020) similarly requires that people remain responsible for intelligence judgements made with AI support.

## Security and confidentiality

Intelligence and research work often involves confidential material: client data, unpublished survey results, commercial plans. Material entered into an external service may be stored or processed outside the organisation's control. Babuta, Oswald and Janjeva (2020), writing about national security uses, stress that data handling, security and legal compliance

must be settled before AI tools are adopted, not afterwards. Businesses should apply the same discipline: know where data go, use services with appropriate contractual protections, and keep sensitive material out of tools that are not approved for it.

## Common mistakes

Teams new to language models tend to make the same errors:

- **Treating the model as a search engine.** It generates text; it does not look facts up unless connected to a retrieval system.
- **Accepting references without checking them.** Every citation must be confirmed to exist and to say what is claimed.
- **Asking leading questions.** The model will usually agree with the premise. Ask neutral questions or ask for both sides.
- **Skipping the hard tasks' checks because the easy ones were right.** Accuracy on routine tasks says little about accuracy on difficult ones.
- **Losing the audit trail.** Record which tool was used, with what inputs, for which part of a product.

## An illustration

A market research team must brief a client on consumer attitudes to plant-based foods across three countries. An analyst gathers the team's recent survey reports, two industry studies and a set of open-ended responses. Using a language model connected only to those documents, the analyst asks for a summary of the main attitudes by country, with each point linked to its source. The draft arrives in minutes. Checking it takes an hour: two figures had been transposed between countries, and one claim about price sensitivity came from a study the team had decided was unreliable. The analyst corrects the draft, adds the team's own judgement about which differences are meaningful, and signs it off. The model saved most of a day; the checking prevented two errors from reaching the client.

## Conclusion

Large language models are powerful assistants for reading, summarising and drafting, and the evidence shows real productivity gains. Their weaknesses, invented facts, misplaced confidence and a tendency to reflect the question back, fall hardest on the complex judgements that matter most in intelligence and research. Used within a disciplined process, grounded in known sources, verified line by line and owned by a named analyst, they let teams cover more ground without lowering their standards. Used as oracles, they put those standards at risk.

## References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 324–335. https://doi.org/10.1017/pan.2023.2

Babuta, A., Oswald, M., & Janjeva, A. (2020). *Artificial intelligence and UK national security: Policy considerations* (Occasional Paper). Royal United Services Institute.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33* (pp. 1877–1901).

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work* (NBER Working Paper No. 31161). National Bureau of Economic Research.

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Working Paper No. 24-013). Harvard Business School.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330).

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33* (pp. 9459–9474).

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). https://doi.org/10.18653/v1/2020.acl-main.173

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Office of the Director of National Intelligence. (2020). *Principles of artificial intelligence ethics for the intelligence community*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (pp. 27730–27744).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35* (pp. 24824–24837).