https://doi.org/10.1038/s41551-024-01228-0

The advent of human-assisted peer review by AI

The Internet didn't disrupt academic publishing. Audiovisual generative AI might do.

Picture this: you receive an invitation to assess a new manuscript in your areas of expertise. The manuscript has already been peer reviewed by expert AI agents, and your task is to review the peer-review outcomes. The manuscript-review interface is multimodal: you can interact with a chatbot editor in whichever ways suit you — via videoconference, voice only, through text prompts, or by writing or drawing on paper or on the screen of your computing device. You don't need to check for clarity of language, for reporting accuracy and thoroughness or for the reliability of code; any such shortcomings would have been ironed out during earlier interactions between the authors and the AI agents.

Instead, you focus on higher-level matters. You can ask the chatbot for a summary of the work, and then prompt it about the main findings. The chatbot will also provide you with a summary of the shortcomings that would need to be addressed as already agreed by the authors and the AI agents, and of any disagreements among them. By marking up plots and diagrams that the chatbot shows you (and sometimes generates specifically for you), you can query unexplained relationships in the data and the necessity of a specific step in the diagram of a workflow. You can ask whether the AI agents have assessed or queried the authors about control groups that you deem necessary, and you can detail mechanistic pathways, characterization methods and validation efforts that should be considered. You can also ask for a list of the caveats and limitations that the authors included in the manuscript, and provide feedback on these.

Although you have quickly inspected the manuscript, you haven't really read it nor checked every figure, diagram and table. Yet, you don't fret about it; you are satisfied with the peer-review process and your contribution to it, and trust that the chatbot will faithfully and clearly relay all your feedback to the authors or the human editor. You have reasons to believe that the expert AI agents are helping the authors to improve methodological thoroughness, language clarity and presentation of the data and visuals. You also know that the expert AI agents will consistently find the sort of inadequacies in the work that most scientists working in the same research area would identify. You have assisted to improve the work both conceptually and at the deeper technical levels, and the mental effort has also provided you with new ideas and has shaped your views of a scientific specialty to which you are willing to contribute more.

This imagined scenario is idealistic, yet not far-fetched. Generative AI models with real-time multimodal capabilities are now available for wider use, as OpenAI showed on 13 May with their unveiling of GPT-4o (GPT stands for generative pre-trained transformer, '4' for version four and 'o' for omni). GPT-4o was trained end-to-end with text, images, videos, audio and speech, and it can process and generate content through all these modalities. According to demo videos released by OpenAI, discussions with the model are fluid (latencies are low and the model allows for interruptions, as in human-to-human conversations), voice tone and emotion can be used to shape the interactions (and make them more pleasant and creative), the model can 'understand' and explain content in videos and images, including handwritten text, drawings and plots (Fig. 1), and instances of the model can interact between them (that is, chatbot–chatbot conversations are also fluid).

That futuristic scenario is, therefore, not difficult to envisage. Large language models are increasingly being broadly used in scientific writing[1] and can provide constructive feedback on manuscripts that complements that of human reviewers[4]. Multimodal AI chatbots will soon help authors to craft better research outputs[1] (drafts and revisions, with review in real time, and by suggesting fitting journals) and will aid reviewers in assessing the outputs more productively and with a positive experience. Most of the feedback given during traditional peer review may soon be provided by foundation models that are fine-tuned to process information from specific research fields or according to particular needs in technical expertise. Chatbot instances of the models may interact among themselves and with authors before human experts and editors are brought into the peer-review process.

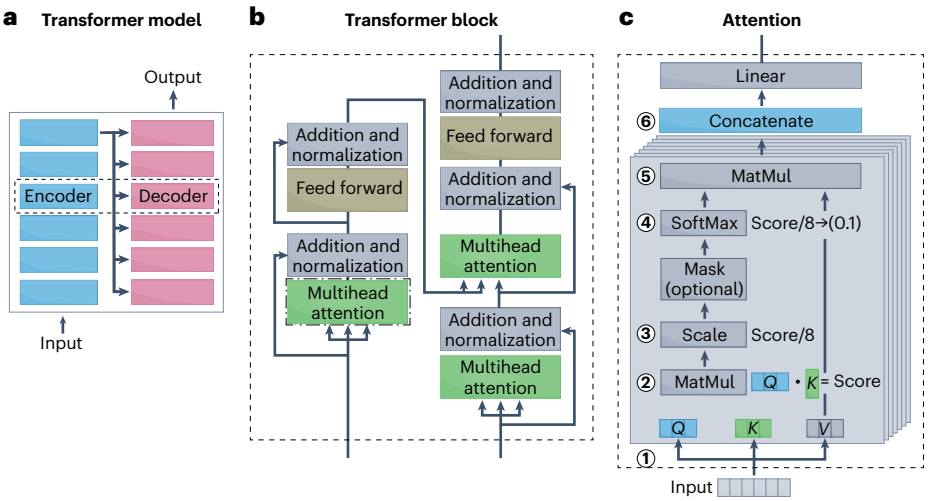
How peer review is carried out may be disrupted by AI. Today's peer-review processes are asynchronous and often involve too many steps, they can be unnecessarily slow because of the substantial effort required and the need to get specialist expertise (peer reviewers are not always available), and they require the poring over a seemingly ever-increasing amount of text, figures, data and, increasingly, code. In the future, AI agents could be developed to carry out most of the publishers' checks, help human editors assess the scientific context, perform peer review before human peer review, and relieve human experts from the drudgery of figuring out where the authors have placed a piece of data or what they meant to convey with a confusing schematic.

Moreover, emotion-aware avatar multimodal AI chatbots could make peer review synchronous when needed (by enabling authors, reviewers and editors to interact directly to, for instance, seek agreement) while preserving anonymity if desired. The AI model may also implement agreed textual and visual changes in real time.

AI agents might be adopted as the go-to interface for reading papers, with the version of record becoming generally less useful for human consumption. Will we prefer to each have our own multimodal AI chatbot that creates a version of each paper according to our preferences and needs (including needs for visual accessibility and preferences arising from neurodivergence), and to which we can query and ask to perform actions, from simple text annotation to the generation of alternative visualizations of the data? Will each paper have its own persistent 'author chatbot' ready to answer any queries? Will experts have their own AI avatar, open for consultations?

This may all seem promising because of the seemingly unbound opportunities of pre-trained large multimodal AI models[2]. But current problems — such as paper mills, falsification of data, plagiarism of ideas, the distortion of truth and the amplification of harm[3], and accessibility of state-of-the-art technology — may be exacerbated in the short term. At present, most publishers, including Springer Nature, ask peer reviewers not to upload manuscripts into generative AI tools. Also, currently unknowable challenges, which are most likely to involve humans intentionally

Editorial



Please briefly explain this schematic (in less than 200 words) so that I can understand the concept and interpret the meaning of the symbols and the interrelationships between them.

This schematic illustrates the architecture of a Transformer model, a powerful type of neural network used primarily for natural language processing tasks.

Panel (a) shows the overall Transformer model, consisting of an encoder and a decoder. The encoder processes the input, while the decoder generates the output based on the encoder's representations.

Panel (b) details a single Transformer block, which is repeated multiple times in both the encoder and the decoder. Each block includes:

1. **Multihead attention**: Computes attention scores across multiple heads, allowing the model to focus on different parts of the input sequence simultaneously.
2. **Addition and normalization**: Adds the input to the attention output (residual connection) and normalizes the result.
3. **Feed-forward network**: Applies a fully connected neural network to the normalized output.
4. **Addition and normalization**: Adds the feed-forward output to its input and normalizes again.

Panel (c) breaks down the attention mechanism:

1. Input sequences are transformed into query (Q), key (K), and value (V) vectors.
2. MatMul: Calculates the dot product between Q and K.
3. Scale: Divides the scores by the square root of the dimension of K to stabilize gradients.
4. (Optional) Mask: Masks out future positions in the decoder to prevent information leakage.
5. SoftMax: Normalizes the scores to a probability distribution.
6. MatMul: Multiplies the scores with V to get the weighted sum of values.
7. Concatenate: Combines the outputs of multiple attention heads.
8. Linear: Applies a linear transformation to the concatenated output.

This sequence allows the model to learn complex relationships and dependencies in the input data.

Fig. 1 | GPT-4o, by OpenAI, can answer queries about scientific figures and diagrams, as exemplified by a schematic of the transformer architecture that underpins all GPT models. Schematic reproduced with permission from ref. 5, Springer Nature Ltd. The explanatory text was generated by GPT-4o via the ChatGPT interface, for illustrative purposes only. The accuracy of the text has been verified by the journal's Chief Editor, Pep Pàmies.

seeking an unfair advantage, will arise. History, however, has shown that we will find ways to deal with these problems through the implementation and tweaking of safeguards within a regulatory environment that shouldn't stifle innovation. As for the models, the hope is that they can learn to refrain from hallucinating also when interacting among themselves.

The potential upsides may be truly worthy. Will human-assisted peer review by AI become substantially more efficient and of higher quality? Will peer review become more inclusive, by improving the skills involved in constructive judgement and critical thinking, particularly for younger researchers? And will the incumbents in scientific publishing (including Springer Nature) – for which the editorial and peer-review management of manuscripts are a substantial investment, and which embraced the Internet two-to-three decades ago and are now developing AI tools – be disrupted by entrants quickly adapting to take advantage of the eyes, ears and mouths of AI?

Published online: 12 June 2024

References

1. Liang, W. et al. Preprint at https://arxiv.org/abs/2404.01268 (2024).
2. *Nat. Biomed. Eng.* **7**, 85–86 (2023).
3. *Nat. Biomed. Eng.* **7**, 705–706 (2023).
4. Liang, W. et al. Preprint at https://arxiv.org/abs/2310.01783 (2023).
5. Zhang, A., Xing, L., Zou, J. & Wu, J. C. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).