
Will AI Be Built Around Language?

Odysseus Kaloidas
UC Berkeley

Abstract

Could language become the common means of organizing AI, even when much of its computation is not linguistic? A scientific question can require literature, numerical models, and experiments. Language may connect these activities by expressing goals, describing tools, and relating their results. A system organized this way could draw on a language model’s learned knowledge while leaving specialized calculations to other methods. Existing systems show that such combinations are possible; they do not establish a universal foundation for AI. Will language remain useful for combining capabilities as those capabilities change, or is its prominence mainly a consequence of today’s models, data, and users?

1 One Question, Several Kinds of Work

A scientist asks why two groups of cells respond differently to the same treatment. Answering may require earlier studies, measurements, a numerical model, and an experiment that separates rival explanations. The question is expressed in language. Much of the work is not. Could language nevertheless provide the common means of organizing it?

In this hypothetical investigation, language connects capabilities that remain specialized internally. Its role includes specifying questions, relating results to prior knowledge, and constructing procedures. That is more substantial than placing a conversational front end on an otherwise unchanged tool. An agent, a system that chooses and carries out steps toward a goal, must still make those connections reliably.

If this coordinating role endures, AI could be built around language without doing everything in words. That is the possibility at issue here. It leaves open both how much of the reasoning will remain linguistic and how much of today’s dependence on language models is temporary.

2 What Does It Mean to Be Built on Language?

A system can receive instructions in language while performing its task with a specialized model. It can instead inherit the learned numerical parameters of an already trained language model and extend that model with new inputs and outputs. A stronger claim is that its intermediate reasoning must itself use natural-language statements. These arrangements can coexist, but none entails all the others.

Tokens are the units into which a model divides its input. A sequence of them need not constitute human language: they can represent image patches, molecular units, or action coordinates. Calling every machine representation a language would make the proposed foundation universal by definition and explain very little.

The distinction is already visible in prior work. Hao and colleagues explicitly propose language models as a general-purpose task layer connected to different encoders [3]. Socratic Models compose pretrained systems through language-based exchanges [7]. HuggingGPT uses a language model to plan tasks, select specialized models from their descriptions, and combine their results [6]. These

earlier systems already demonstrate forms of linguistic coordination. The question here is why it might remain important as the coordinated capabilities change.

3 Why a Common Medium Could Matter

Language supplies a reusable way to express a task without first choosing every computational step. A request can refer to a constraint, an exception, or a relationship between results from different tools. If a system can interpret those descriptions and connect them to executable operations, its components need not have been designed for precisely that combination.

In the cell investigation, one tool might report an average for a group while another predicts a response for an individual cell. Calling both quantities a treatment effect would not make them interchangeable. Tools can also differ in units and assumptions. A common language helps only if the system understands enough about the operations to connect their results correctly.

Pretrained models offer a second kind of continuity. PaLM-E incorporates numerical representations of visual and sensor observations into a language-model-based system and reports benefits from training jointly across these kinds of input [2]. Such results support reusing learned resources across tasks; they do not isolate natural language as the only useful ingredient. Language can matter through inherited knowledge even when information enters the model in another form.

These routes suggest a possible cumulative advantage. Existing descriptions of scientific concepts and tools make new capabilities easier to connect, while useful connections create further applications for the common medium. This is an explanation for a possible pattern, not evidence that every system will converge on one architecture.

4 Biology: Connecting Knowledge to New Evidence

C2S-Scale develops the Cell2Sentence approach by training on sequences of gene names derived from cell measurements, alongside biological text and information about the samples [5]. Its authors report a model-generated hypothesis followed by validation in human cell models. The work is a preprint, and this case does not establish general scientific autonomy. It nevertheless illustrates a constructive role for language-model resources in a discovery process.

The stages contribute different kinds of evidence. A gene-name sequence reorganizes a measurement; it does not add information to that measurement. Prior biological knowledge can help turn it into a hypothesis. The experiment that tests the hypothesis supplies evidence that was not available when it was proposed. Language-model resources can help connect these stages without replacing any of them.

Reliable execution remains part of that connection. A June 2026 first-party research account from Anthropic reports substantial improvements in biological agents' data retrieval after introducing a retrieval tool that executes specified data lookups through fixed software procedures [4]. The example supports a layered interpretation: an agent can formulate and coordinate work while specialized software handles exact operations. It does not show that language alone provides their reliability.

5 A Linguistic Foundation Can Have Nonlinguistic Machinery

Robotic control makes the distinction tangible. In the robot-control system π_0 , a model trained on images and language works with a specialized component that generates robot movements [1]. Language can help determine the task while the system produces movements through a representation suited to control. The system's connection to language does not require verbal descriptions of each movement.

Transfusion supplies a related example within one multimodal model. It learns to predict the next text token and to generate images by gradually removing noise, a process called diffusion [8]. A common model can therefore accommodate different kinds of data and learning objectives. Neither shared parameters nor a shared interface proves that every modality has been reduced to natural language.

There are reasons for such diversity. A short instruction can identify which object to move while leaving its precise geometry unspecified. Converting every relevant detail into prose could add cost

or lose distinctions the controller needs. These are arguments about the suitability of representations, not proof that linguistic reasoning cannot contribute to physical tasks.

6 A Lasting Foundation or a Historical Starting Point?

Language may also be prominent for reasons that will not last. Widely available text, reusable model weights, and demand for human-facing interaction make it an attractive starting point. Training jointly on several kinds of input from the outset, using specialized models, or communicating through learned numerical representations could eventually make that starting point less important.

A system’s origins may then tell us less about what it continues to need. A model initialized from language weights may acquire capabilities that no longer depend on explicit words. Equally, a system whose internal models exchange numerical representations may still rely on language to formulate unfamiliar goals and connect their results to scientific knowledge.

The enduring role, if there is one, may therefore concern how capabilities become usable together. AI could become more diverse in its internal machinery while retaining a common linguistic means of directing and combining it. That future would give language a foundational role without making it the substance of everything the system knows or does.

References

- [1] Kevin Black, Noah Brown, Danny Driess, et al. π_0 : A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025. URL <https://www.roboticsproceedings.org/rss21/p010.html>.
- [2] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 8469–8488. PMLR, 2023. URL <https://proceedings.mlr.press/v202/driess23a.html>.
- [3] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces, 2022. URL <https://arxiv.org/abs/2206.06336>. arXiv:2206.06336.
- [4] Laura Luebbert. Paving the way for agents in biology. Anthropic Science, 2026. URL <https://www.anthropic.com/research/agents-in-biology>. June 8. First-party research account.
- [5] Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, et al. Scaling large language models for next-generation single-cell analysis, 2026. URL <https://doi.org/10.1101/2025.04.14.648850>. bioRxiv preprint, January 1, version 4; first posted 2025.
- [6] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL <https://arxiv.org/abs/2303.17580>.
- [7] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=G2Q2Mh3avow>.
- [8] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025. URL <https://arxiv.org/abs/2408.11039>.