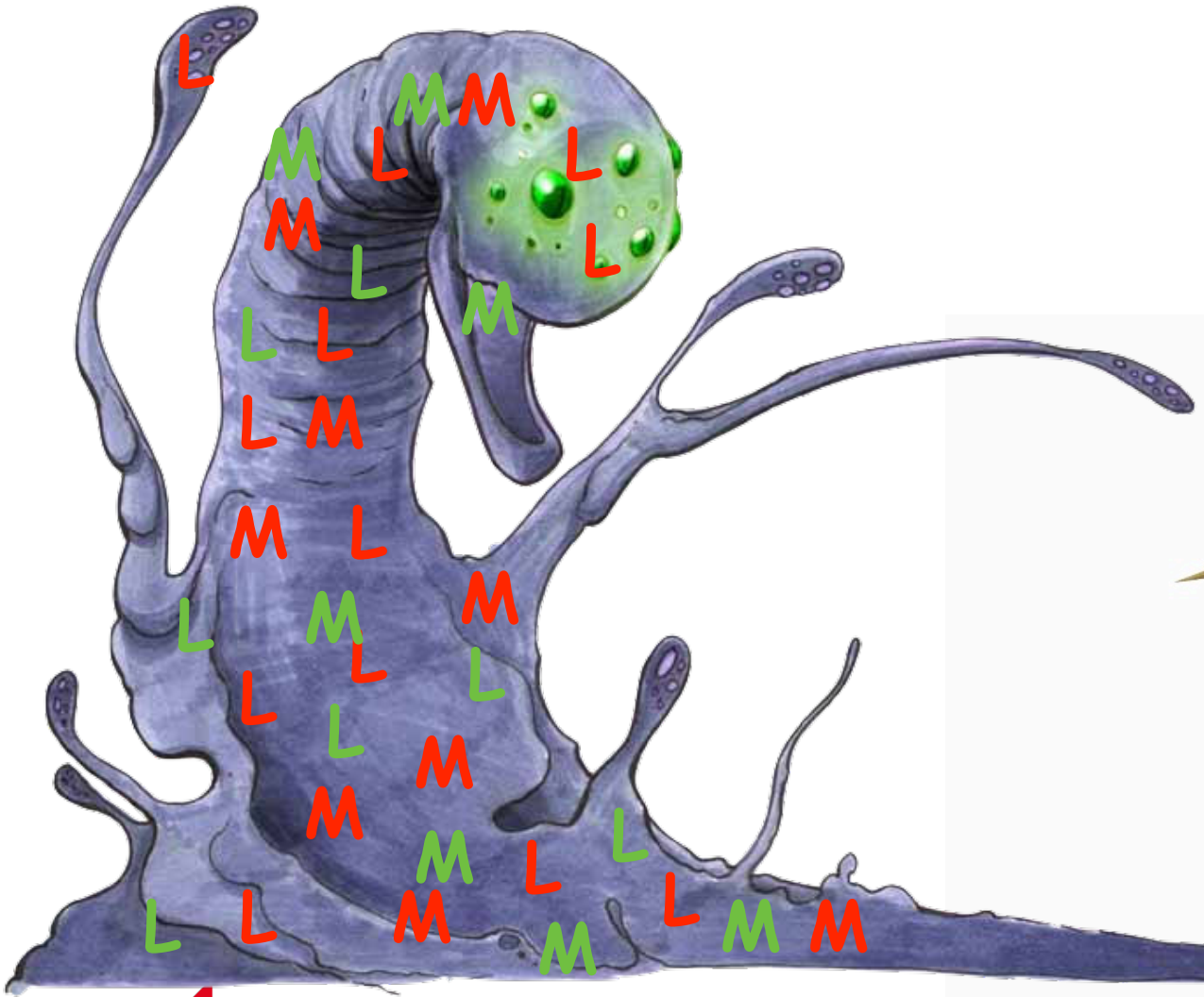
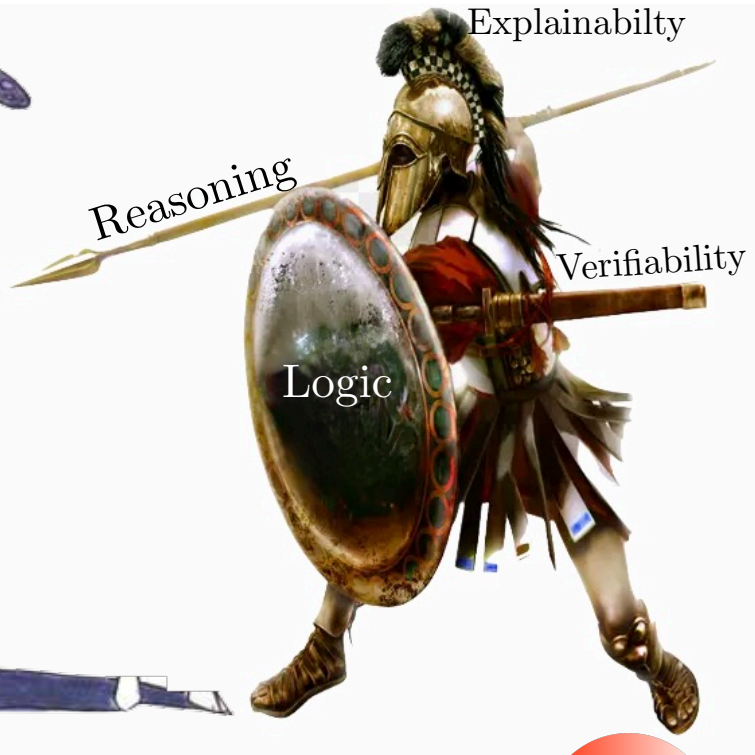
Theorem Provers and the Future AI Math Ecosystem

Stephan Schulz



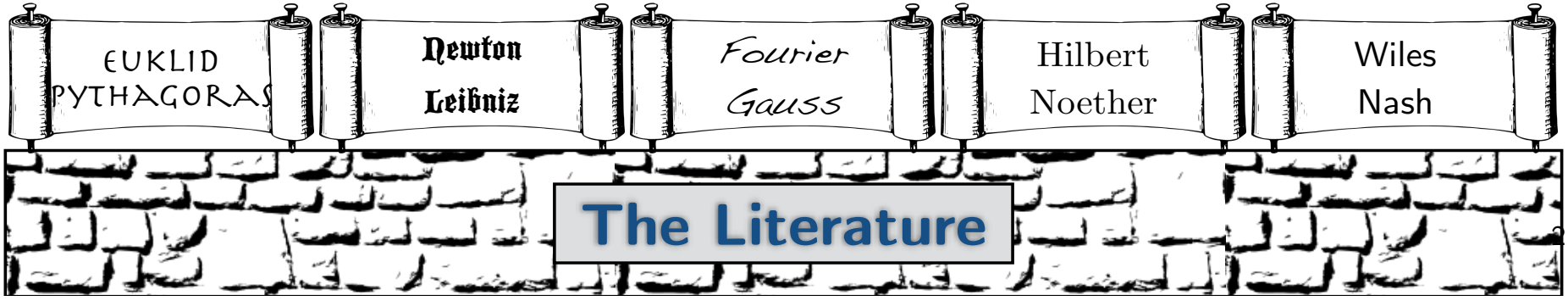


DHBW

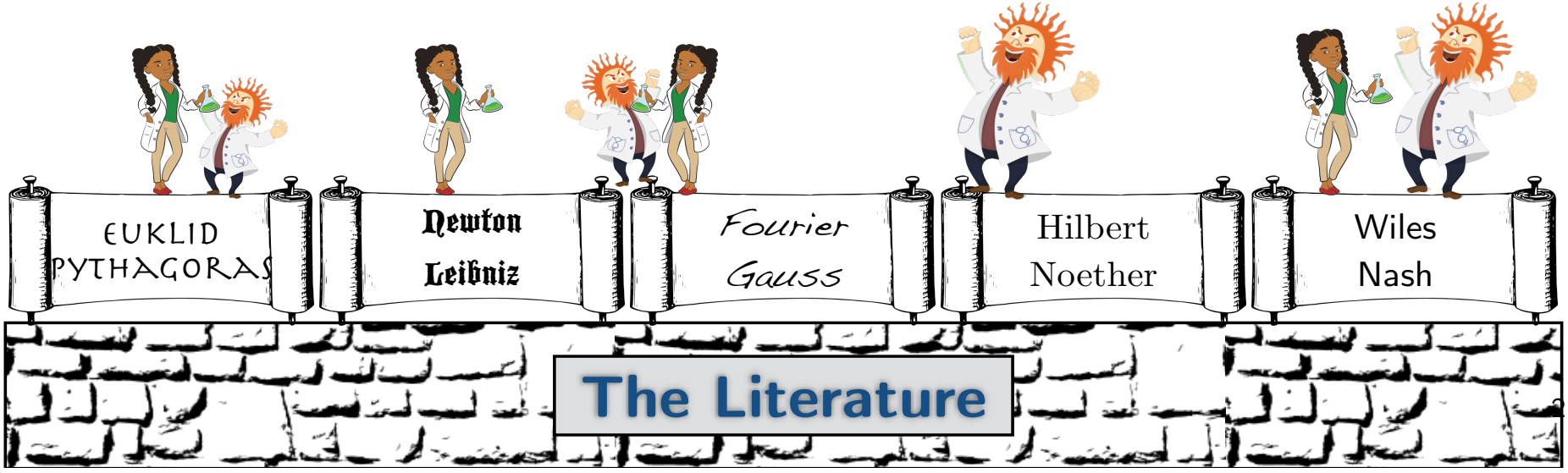
ZfKI

E

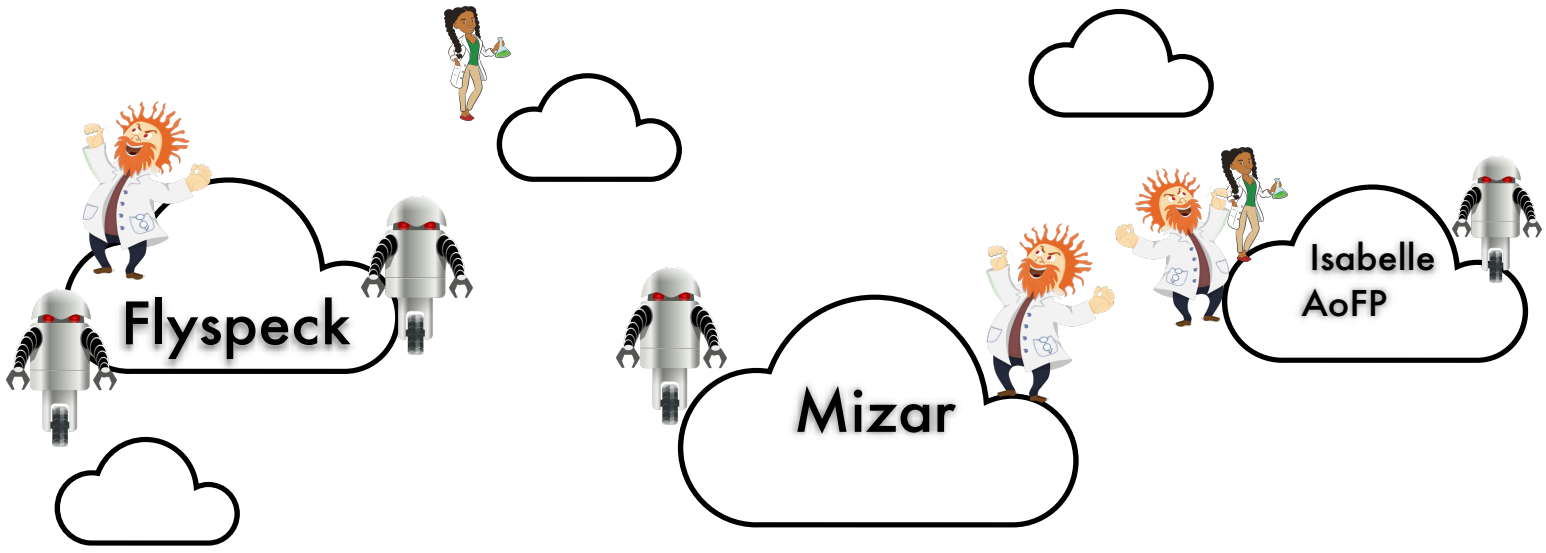
The math ecosystem (ca. now)

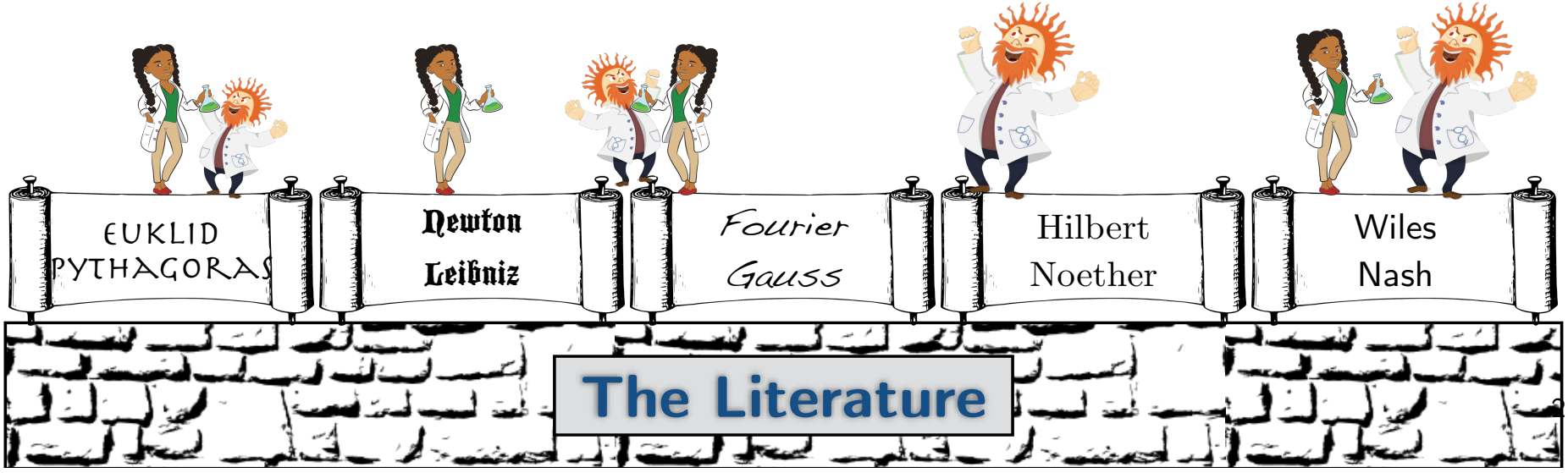


The math ecosystem (ca. now)



The math ecosystem (ca. now)

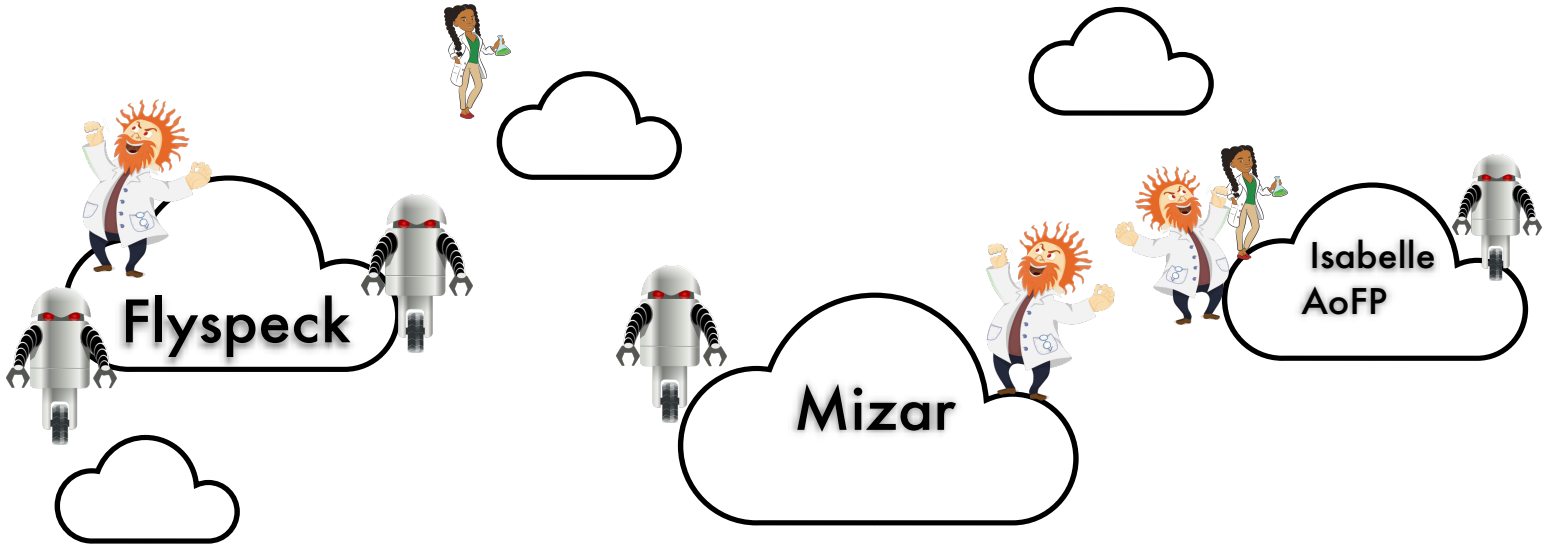


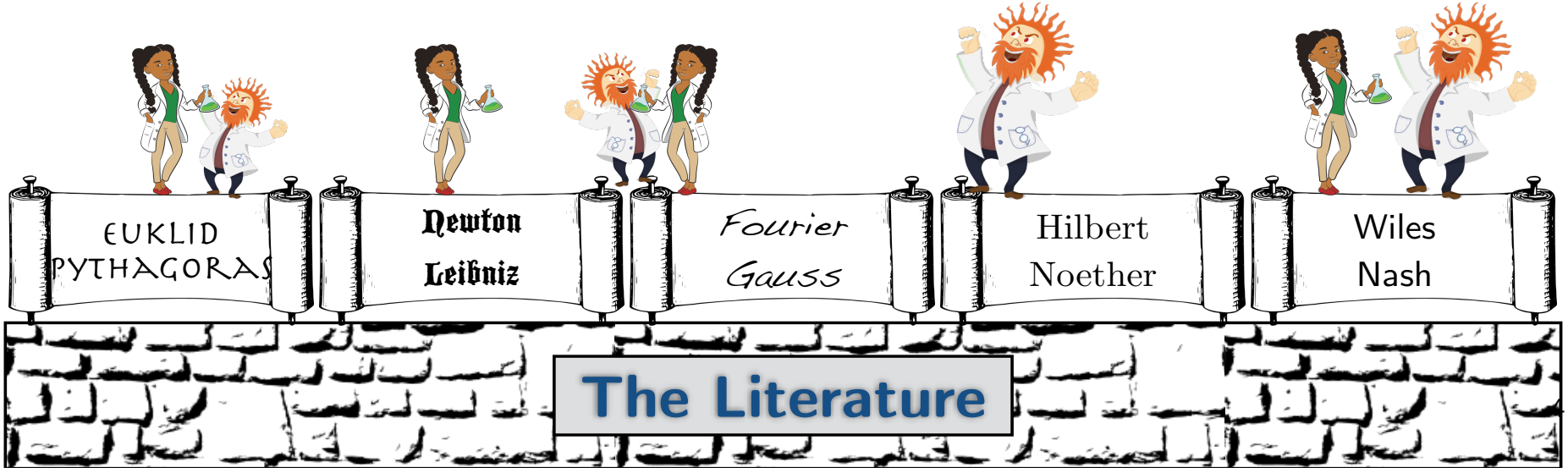


Enter

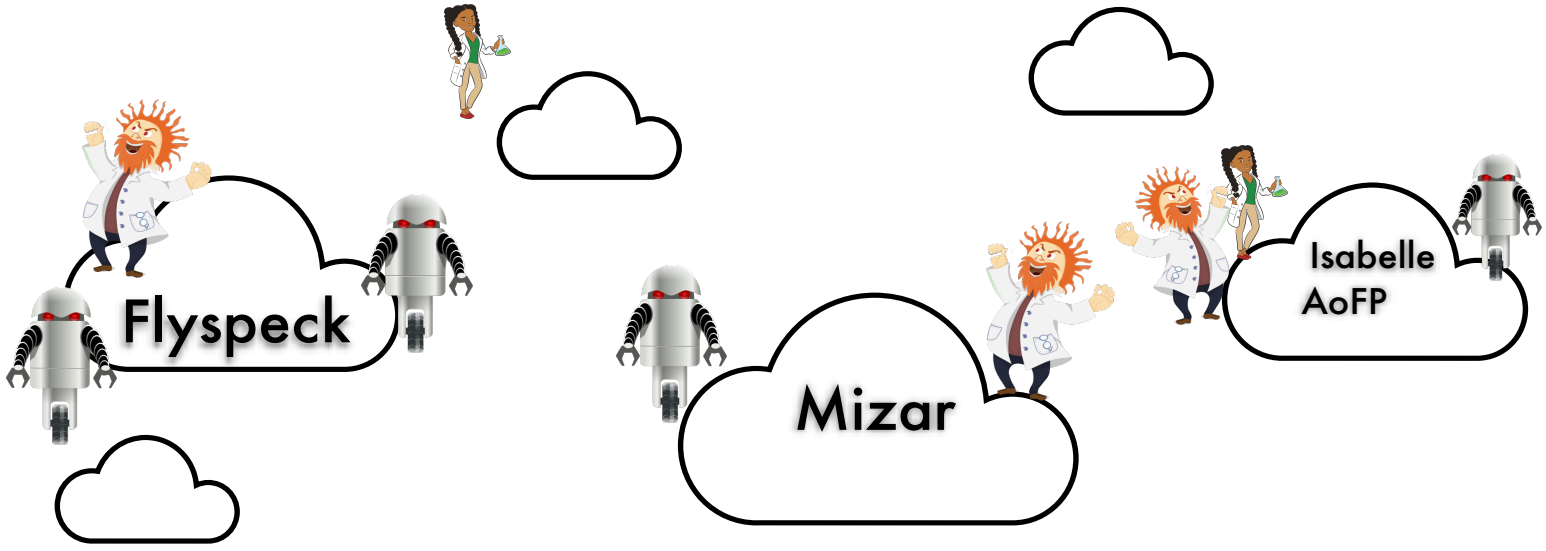
THE AGE OF ARTIFICIAL INTELLIGENCE

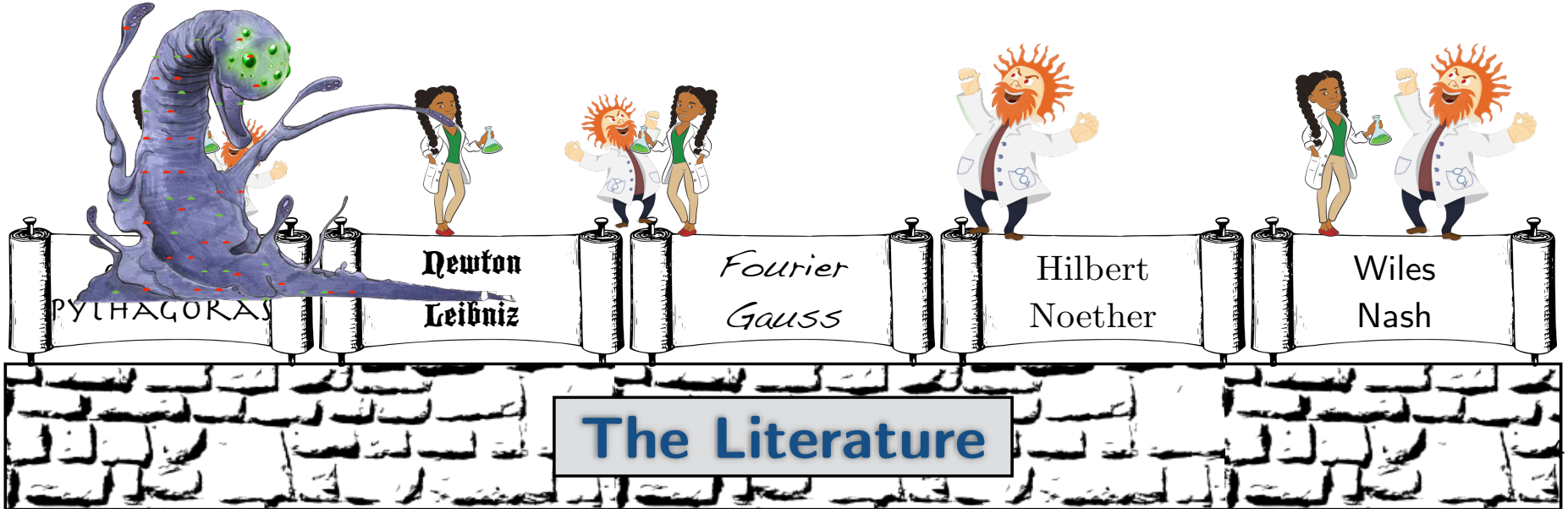
The future math ecosystem (OpenAI/xAI/...version)



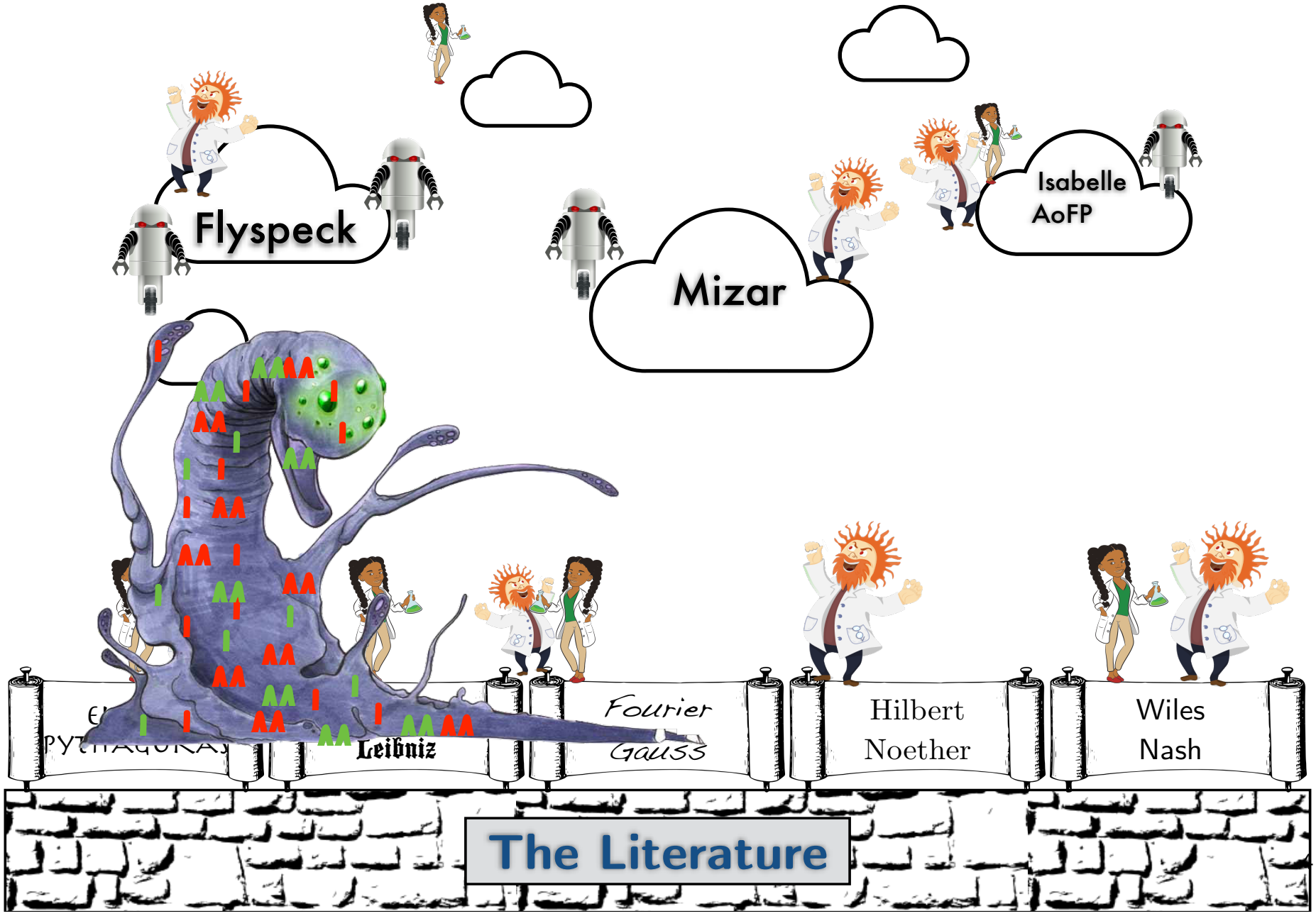


The future math ecosystem (OpenAI/xAI/...version)

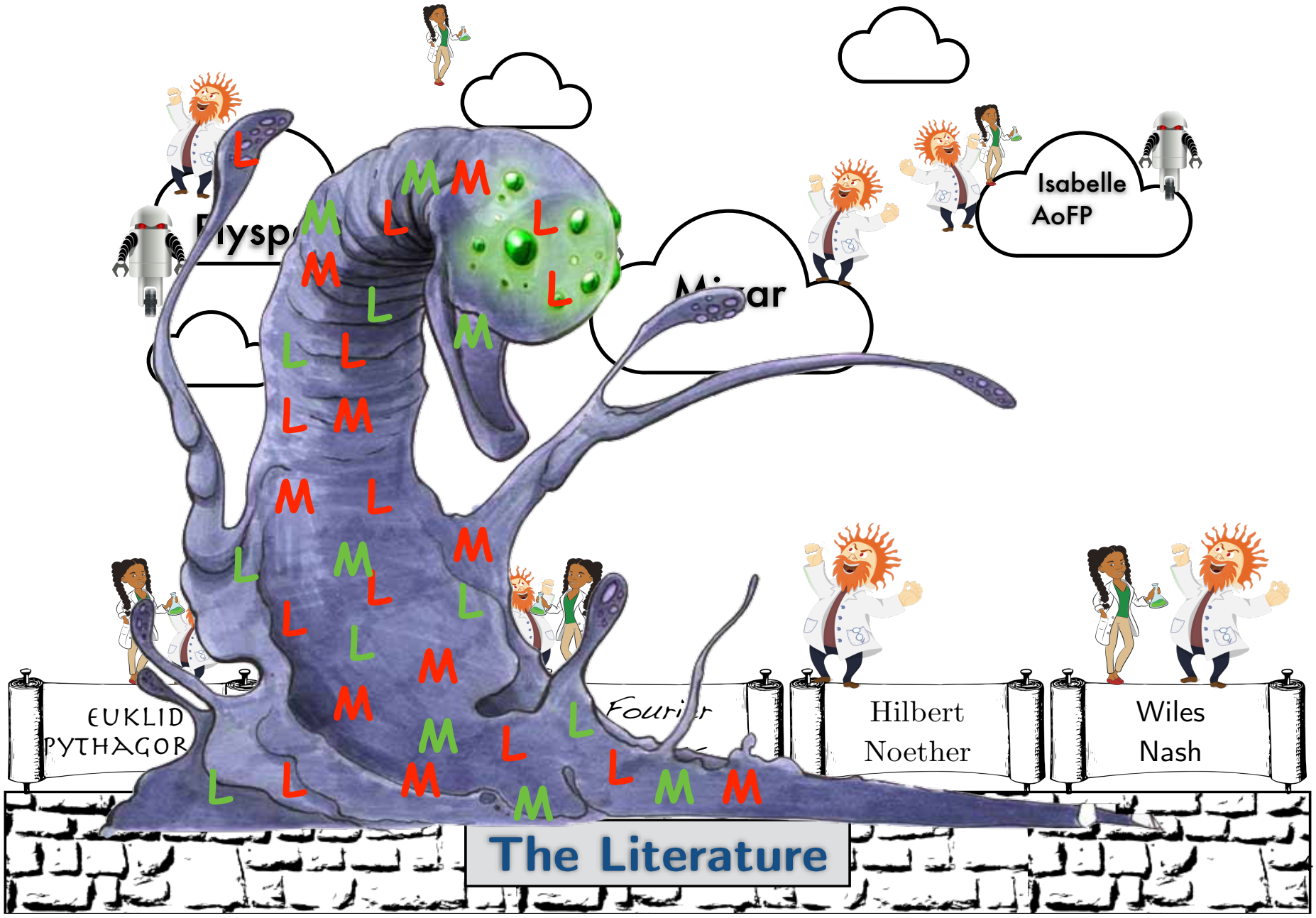




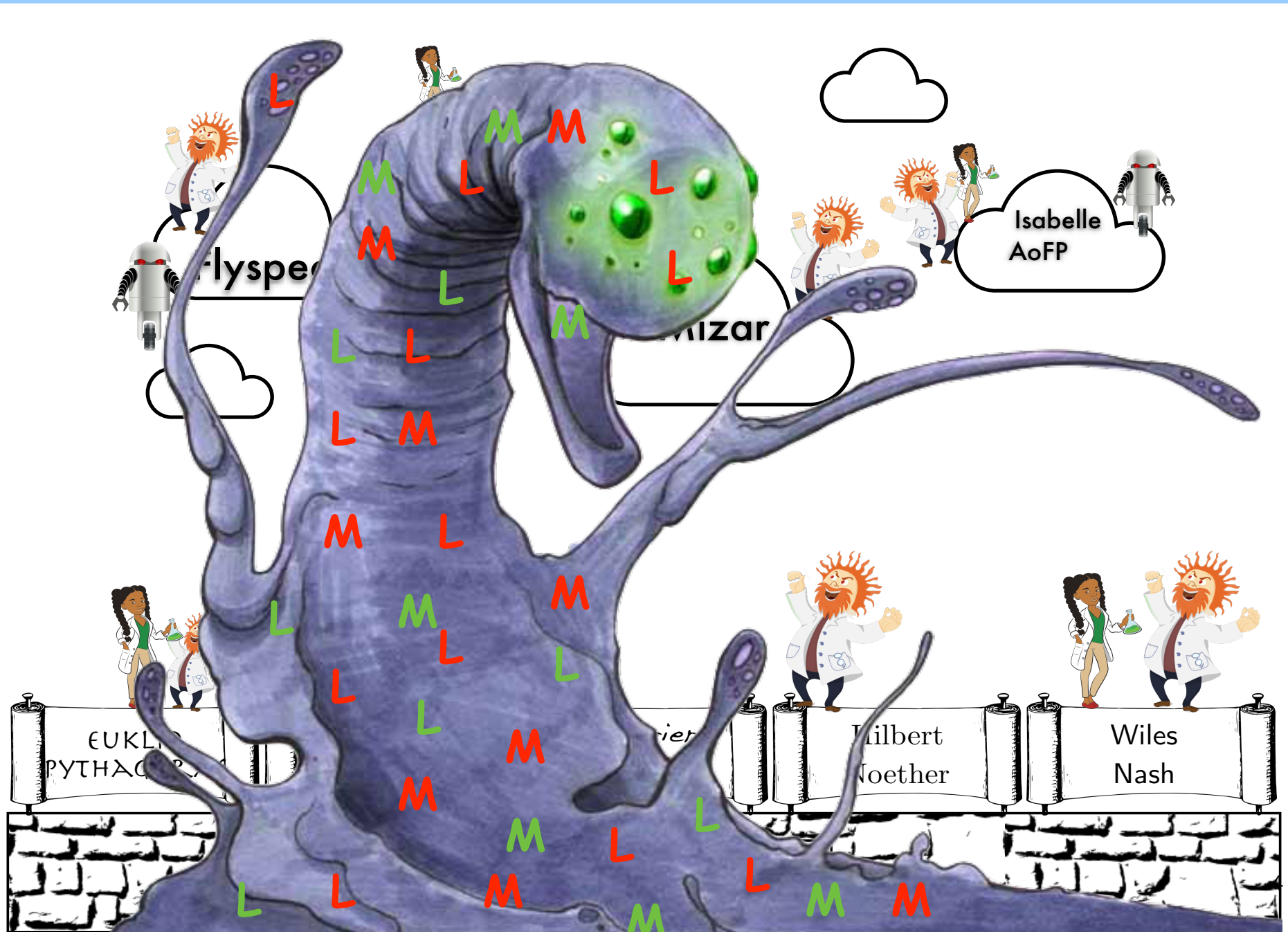
The future math ecosystem (OpenAI/xAI/...version)



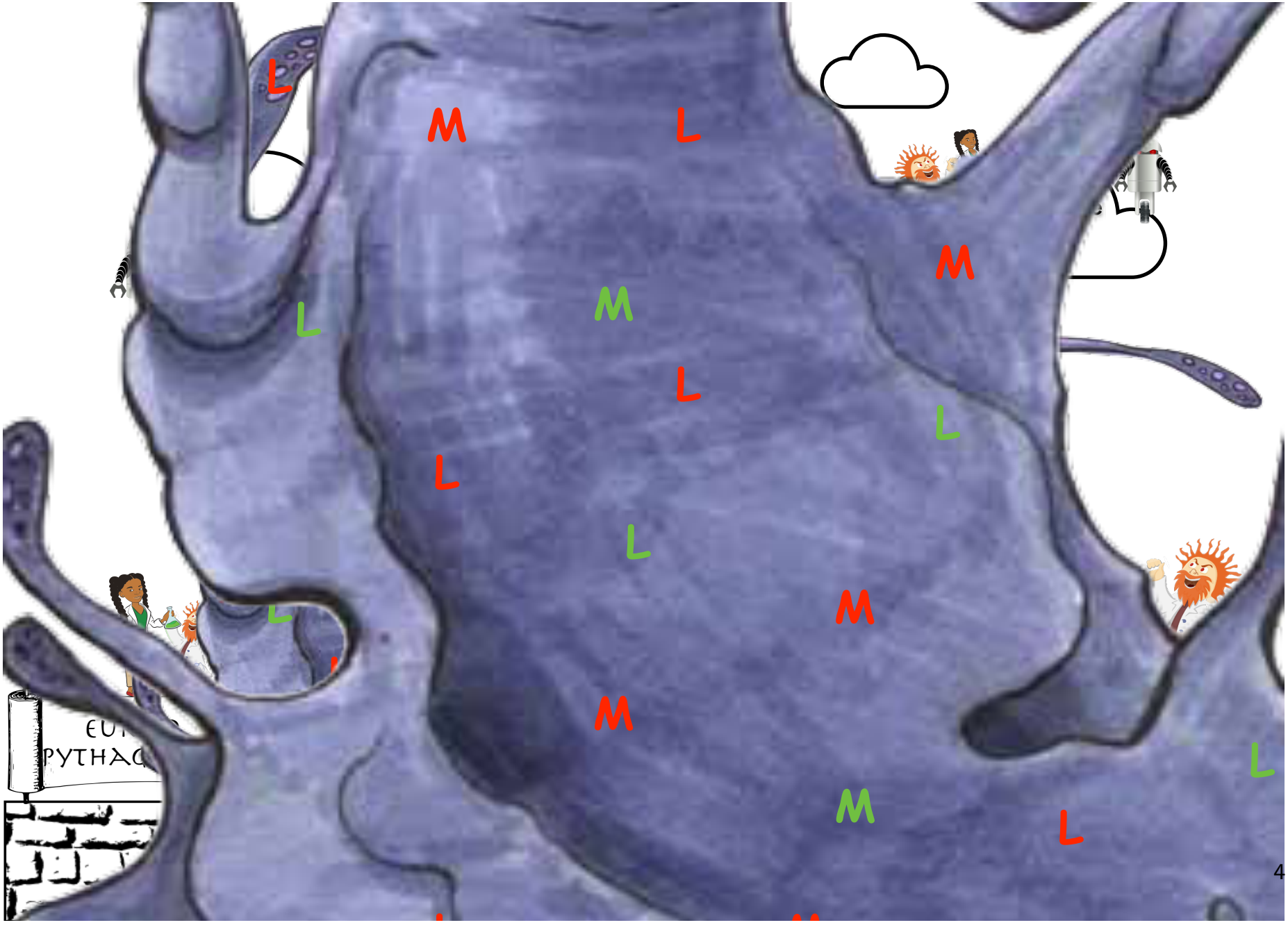
The future math ecosystem (OpenAI/xAI/...version)



The future math ecosystem (OpenAI/xAI/...version)



The future math ecosystem (OpenAI/xAI/...version)



This is not the future!

Large language models are amazing. . .

Large language models are amazing. . .

- . . . at processing *language*

Bold claims

Mathematics is *not* a language

Claiming otherwise is the same confusion as that between being able to program, and knowing a programming language (which gave us the horror that is COBOL). Or between knowing the rules of chess and being a competent chess player.

Bold claims

Mathematics is *not* a language

Claiming otherwise is the same confusion as that between being able to program, and knowing a programming language (which gave us the horror that is COBOL). Or between knowing the rules of chess and being a competent chess player.

- Of course there is *a* mathematical language
  - . . . with a lot of national dialects
  - . . . with a lot of field-specific dialects

Bold claims

Mathematics is *not* a language

Claiming otherwise is the same confusion as that between being able to program, and knowing a programming language (which gave us the horror that is COBOL). Or between knowing the rules of chess and being a competent chess player.

- Of course there is *a* mathematical language
  - . . . with a lot of national dialects
  - . . . with a lot of field-specific dialects

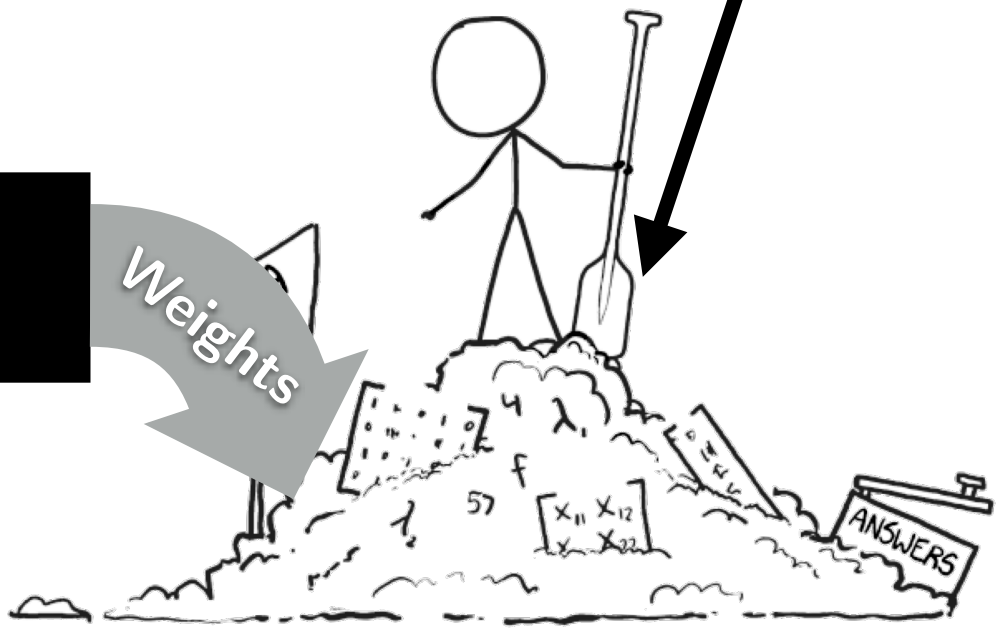
Mathematics is, at the core, the definition and exploration of new structures and their properties.

Genesis of an LLM

All the world's GPUs

Weights

Giant pile encoding the conditional probability of string continuations

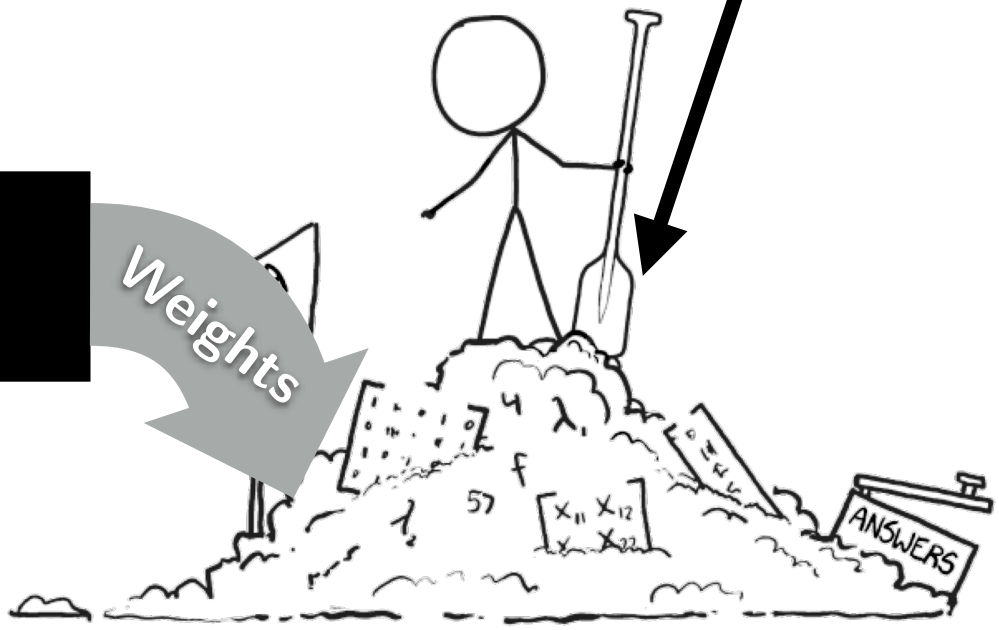


Genesis of an LLM

All the world's GPUs

Weights

Giant pile encoding the conditional probability of string continuations



On LLMs

- LLMs learn conditional probability distributions on language
  - They do not learn abstract theory
  - They do not learn abstract reasoning
  - They do not learn calculations (except by memorization)
- LLMs are essentially reproductive, not creative

On LLMs

- LLMs learn conditional probability distributions on language
  - They do not learn abstract theory
  - They do not learn abstract reasoning
  - They do not learn calculations (except by memorization)
- LLMs are essentially reproductive, not creative

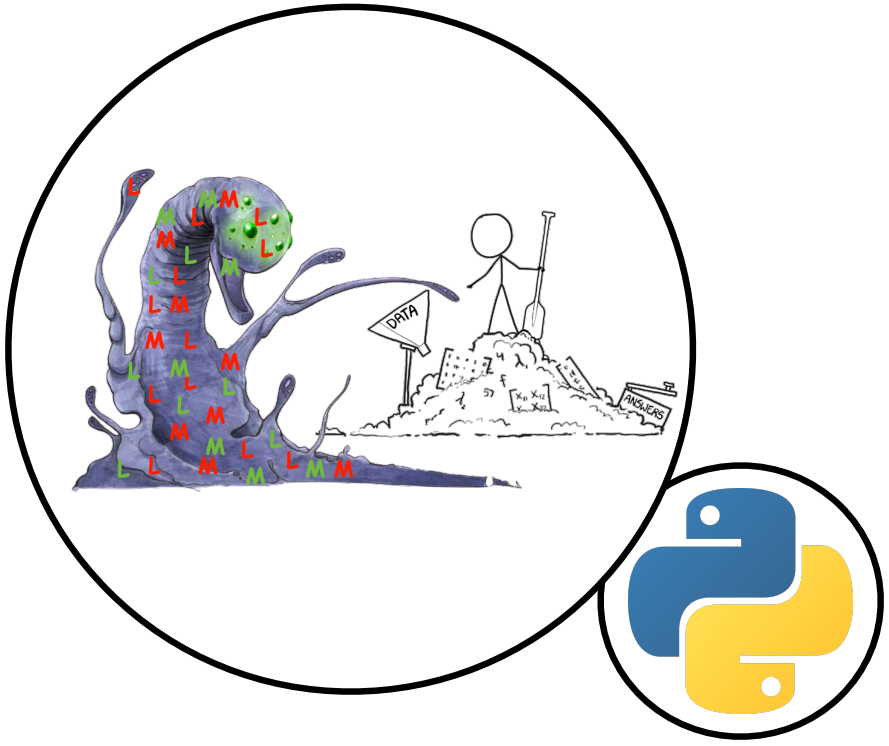
LLMs can't do (new) mathematics!

LLMs can't handle abstract reasoning

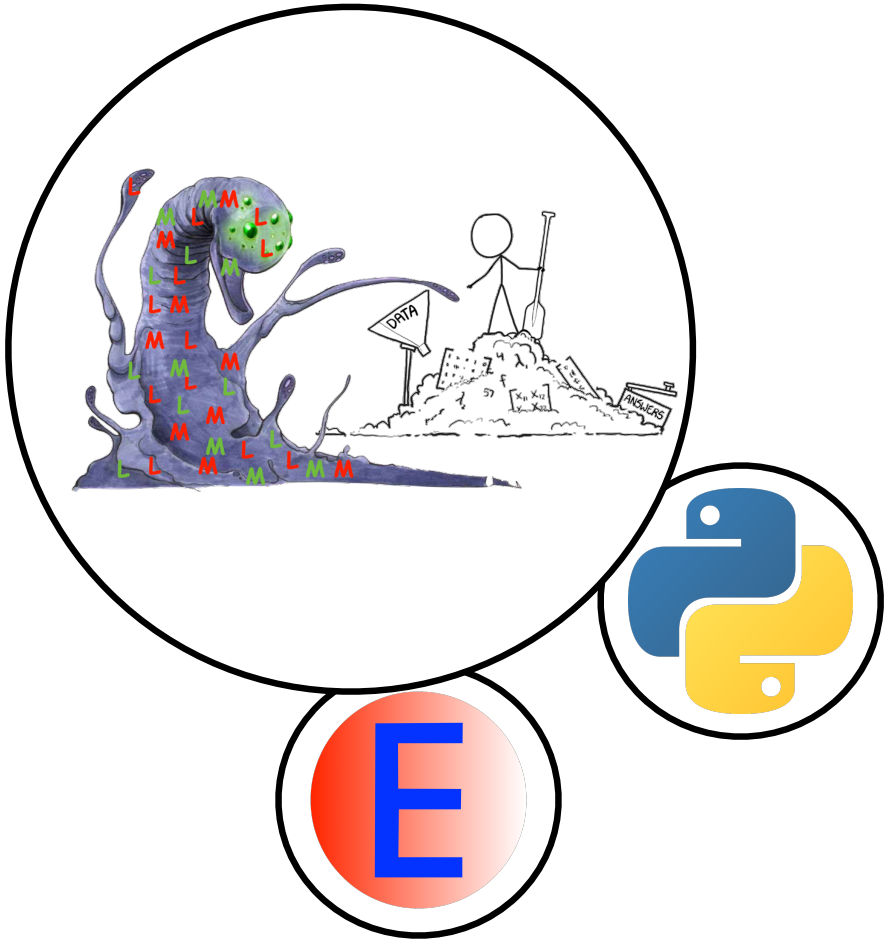
- Systematically renaming terms kills performance (*A chat with Bard*, [SMS25])
- Adding unrelated information leads to failure (*GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*, [MAS+24])
- Answers are not robust (*Fidelity of Medical Reasoning in Large Language Models*, [BJC+25])
- Hallucinations are likely unavoidable (*LLMs Will Always Hallucinate, and We Need to Live With This*, [BAS24])
- Models only "fake" understanding (*Potemkin Understanding in Large Language Models*, citeMWVM:Potemkin-2025)
- Even so-called *reasoning models* break gracelessly (*The Illusion of Thinking : Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, [SMA+25])

Fixing LLMs (?): The LLM+ Architecture

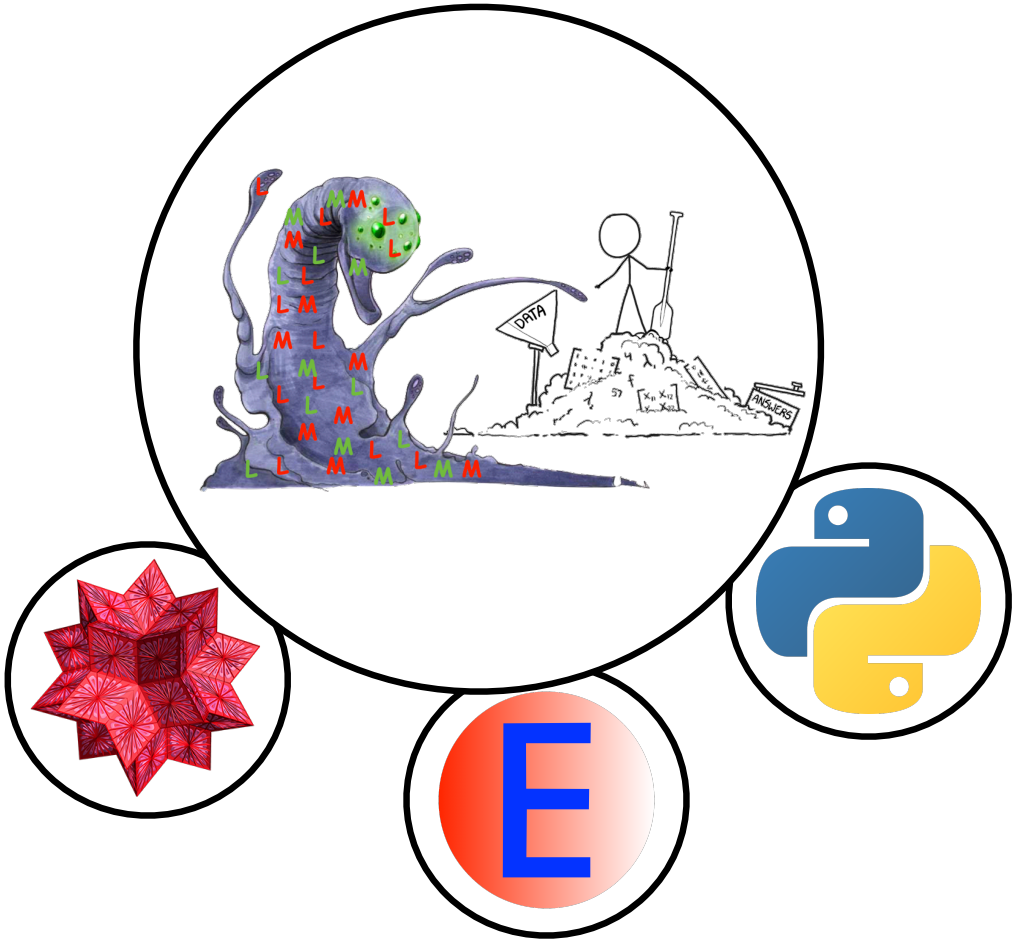
Fixing LLMs (?): The LLM+ Architecture



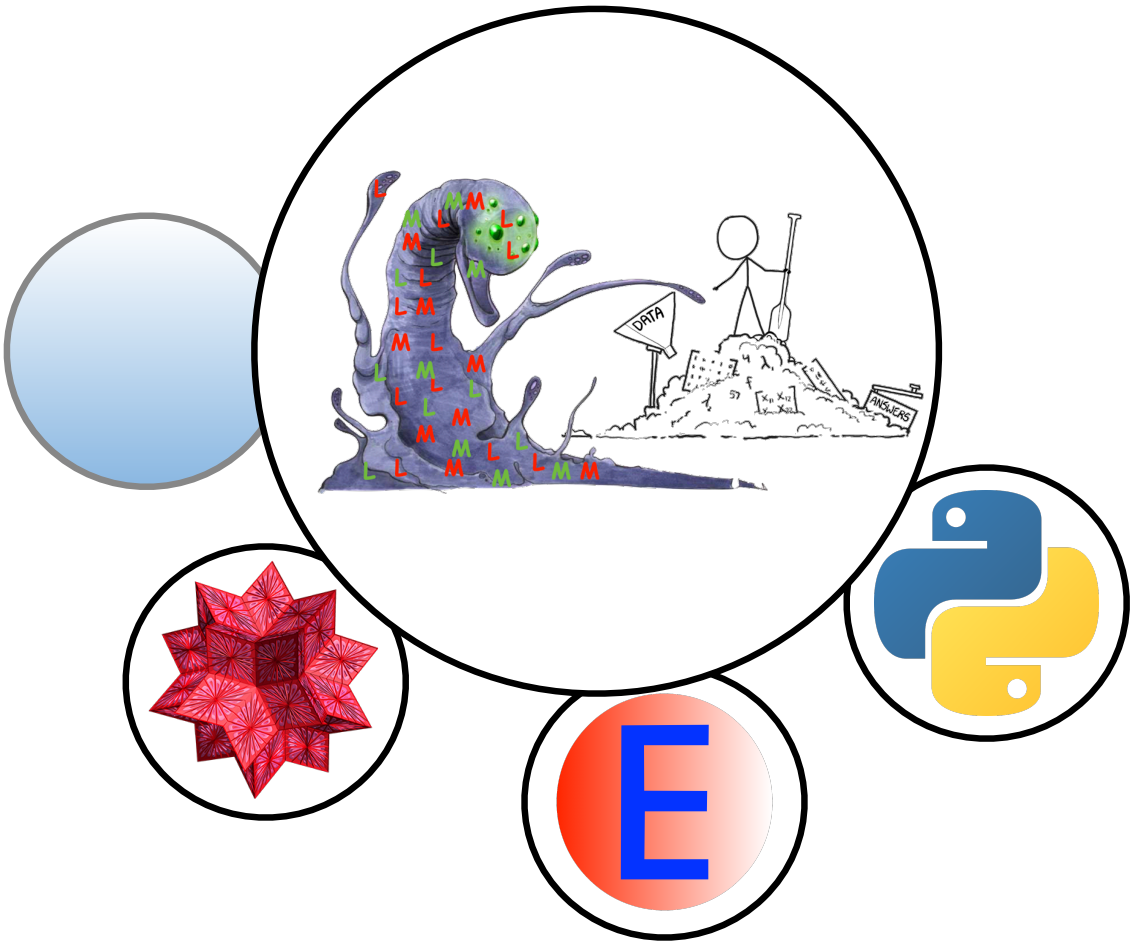
Fixing LLMs (?): The LLM+ Architecture



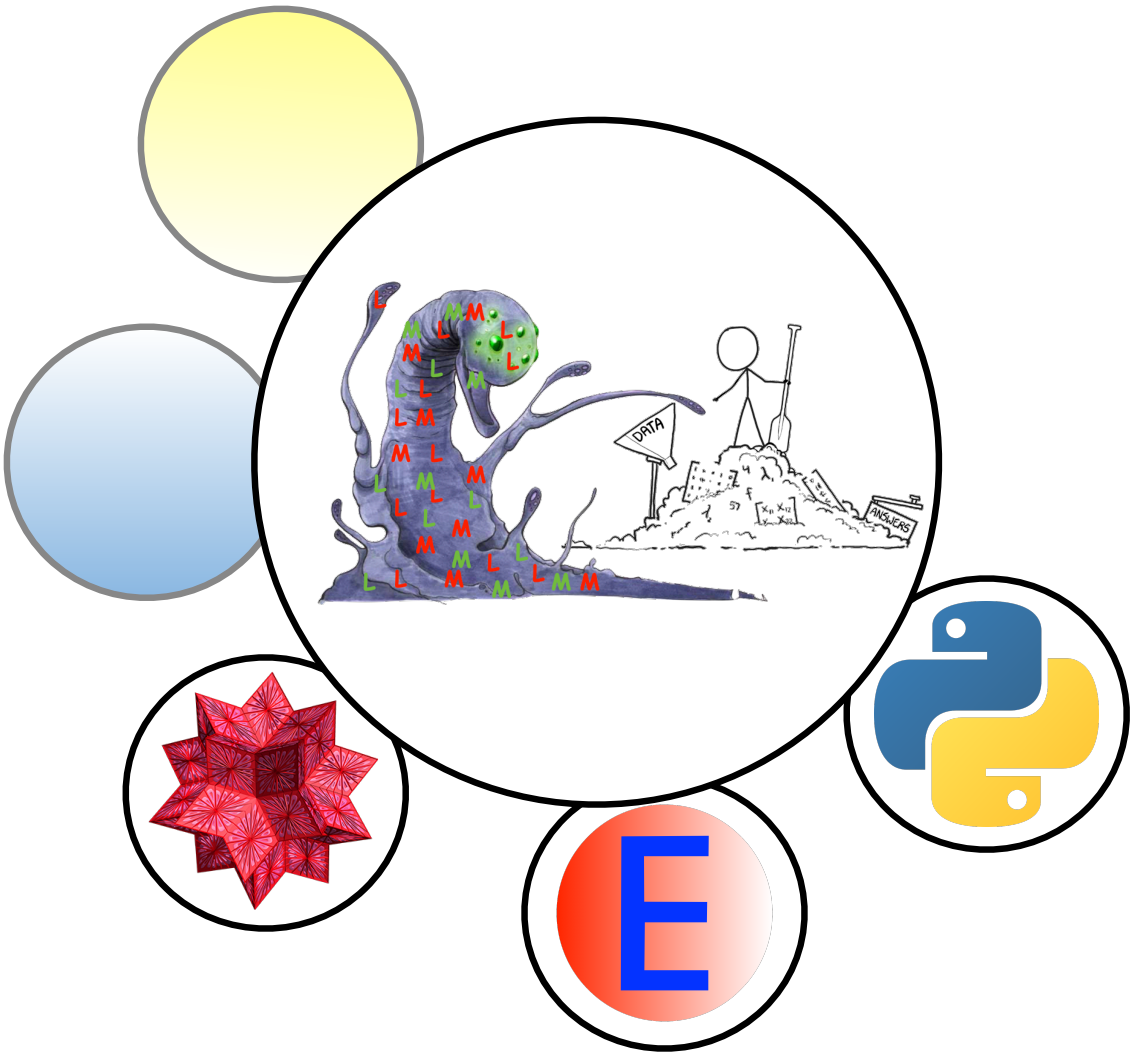
Fixing LLMs (?): The LLM+ Architecture



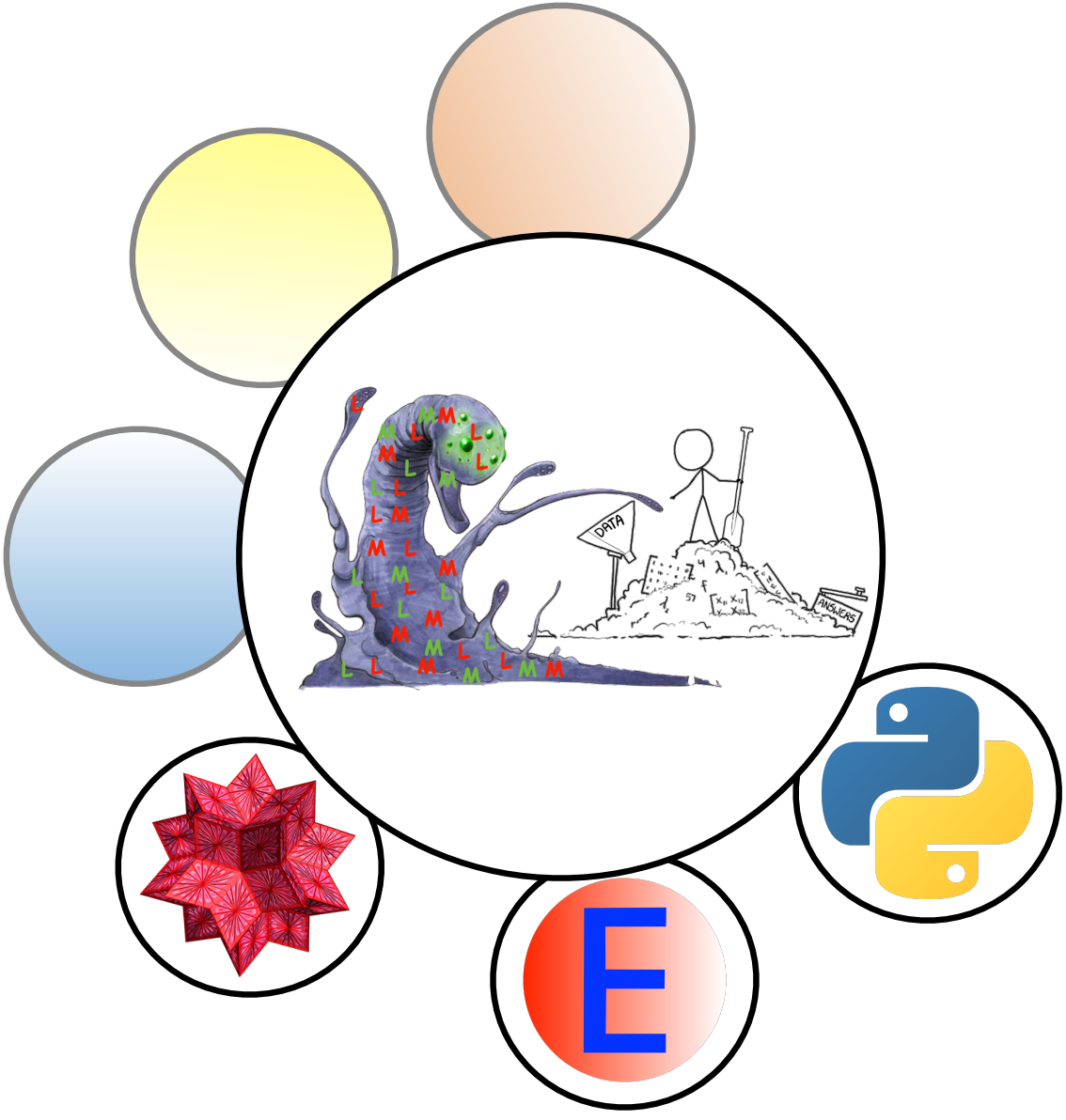
Fixing LLMs (?): The LLM+ Architecture



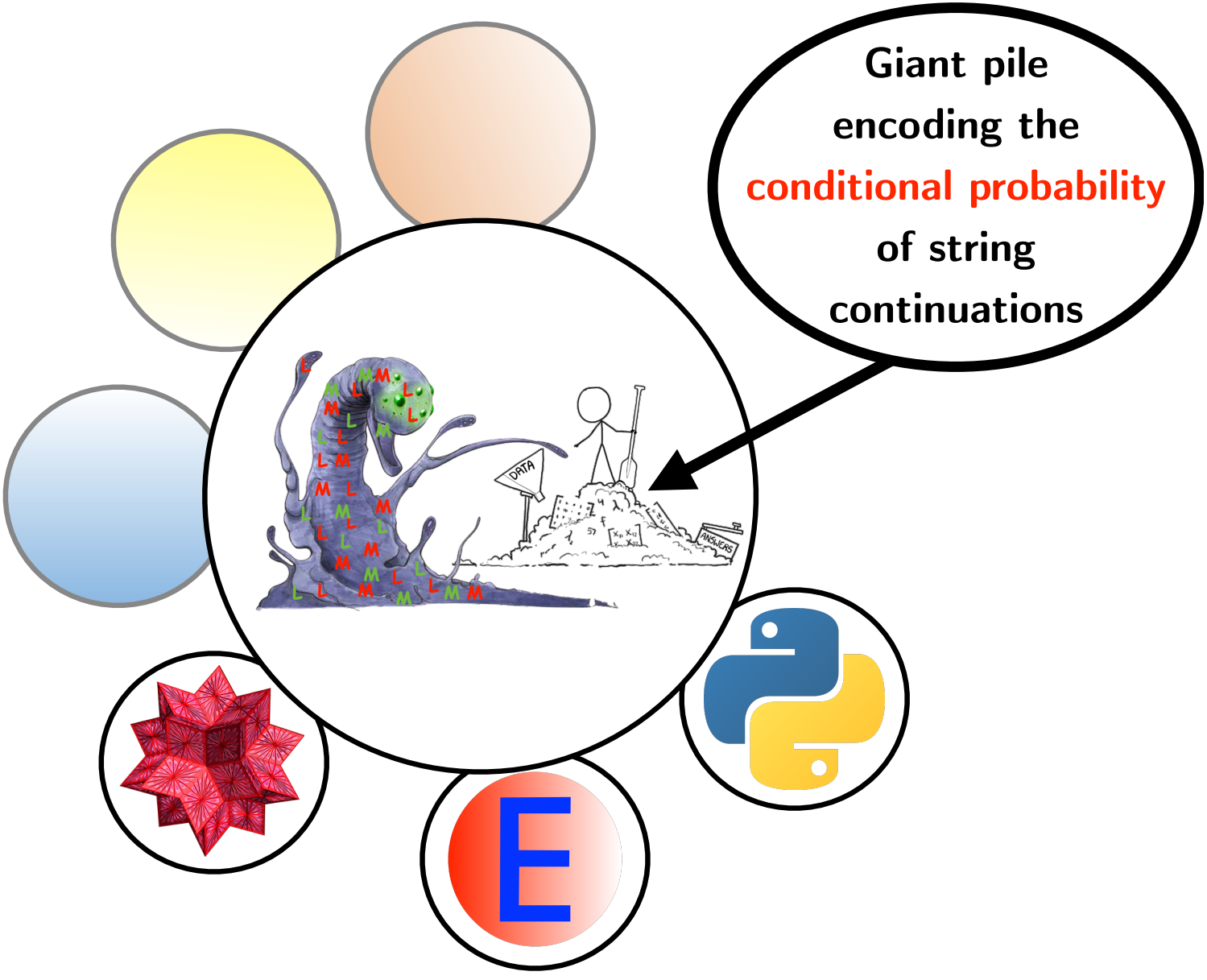
Fixing LLMs (?): The LLM+ Architecture



Fixing LLMs (?): The LLM+ Architecture

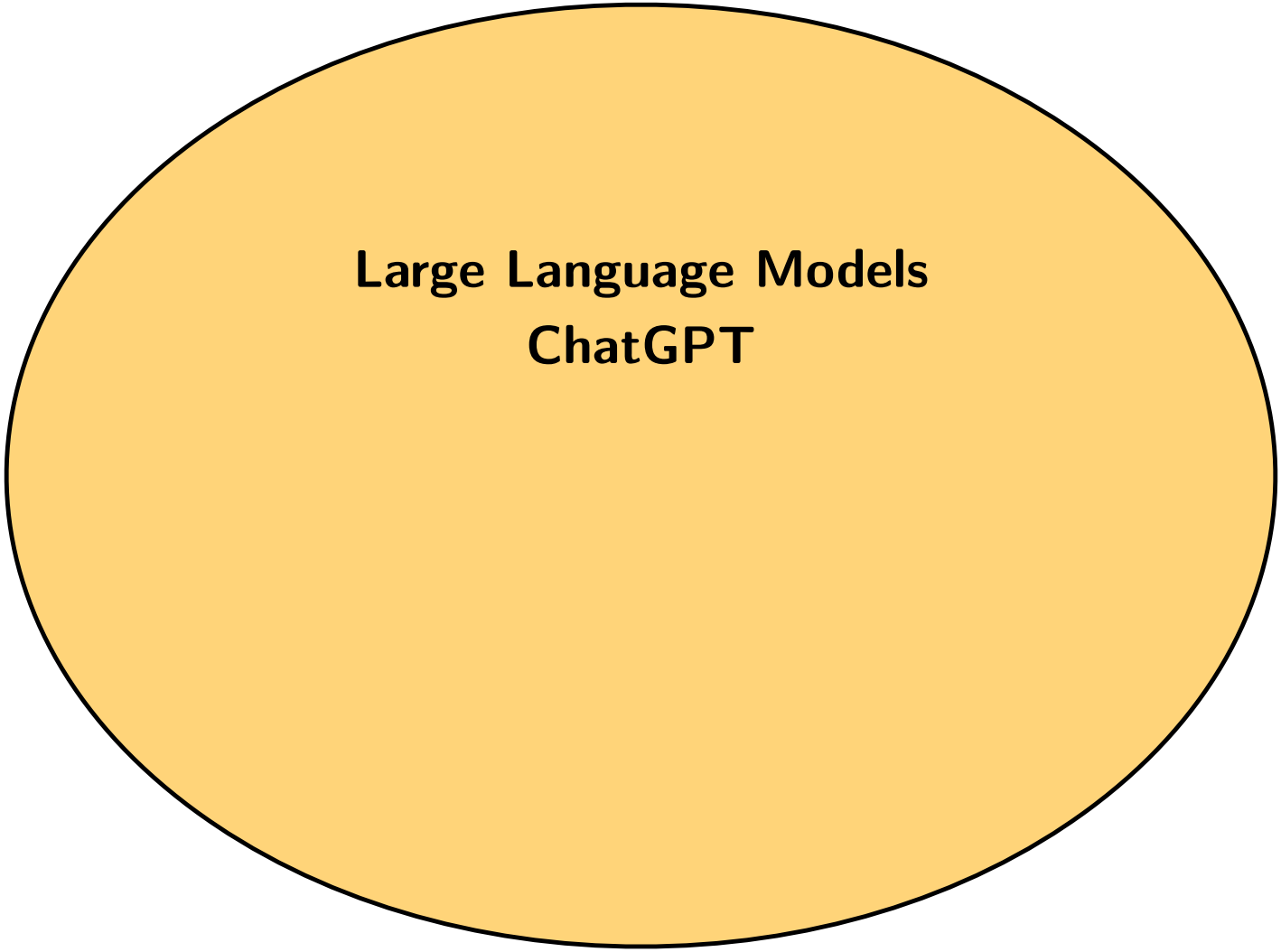


Fixing LLMs (?): The LLM+ Architecture

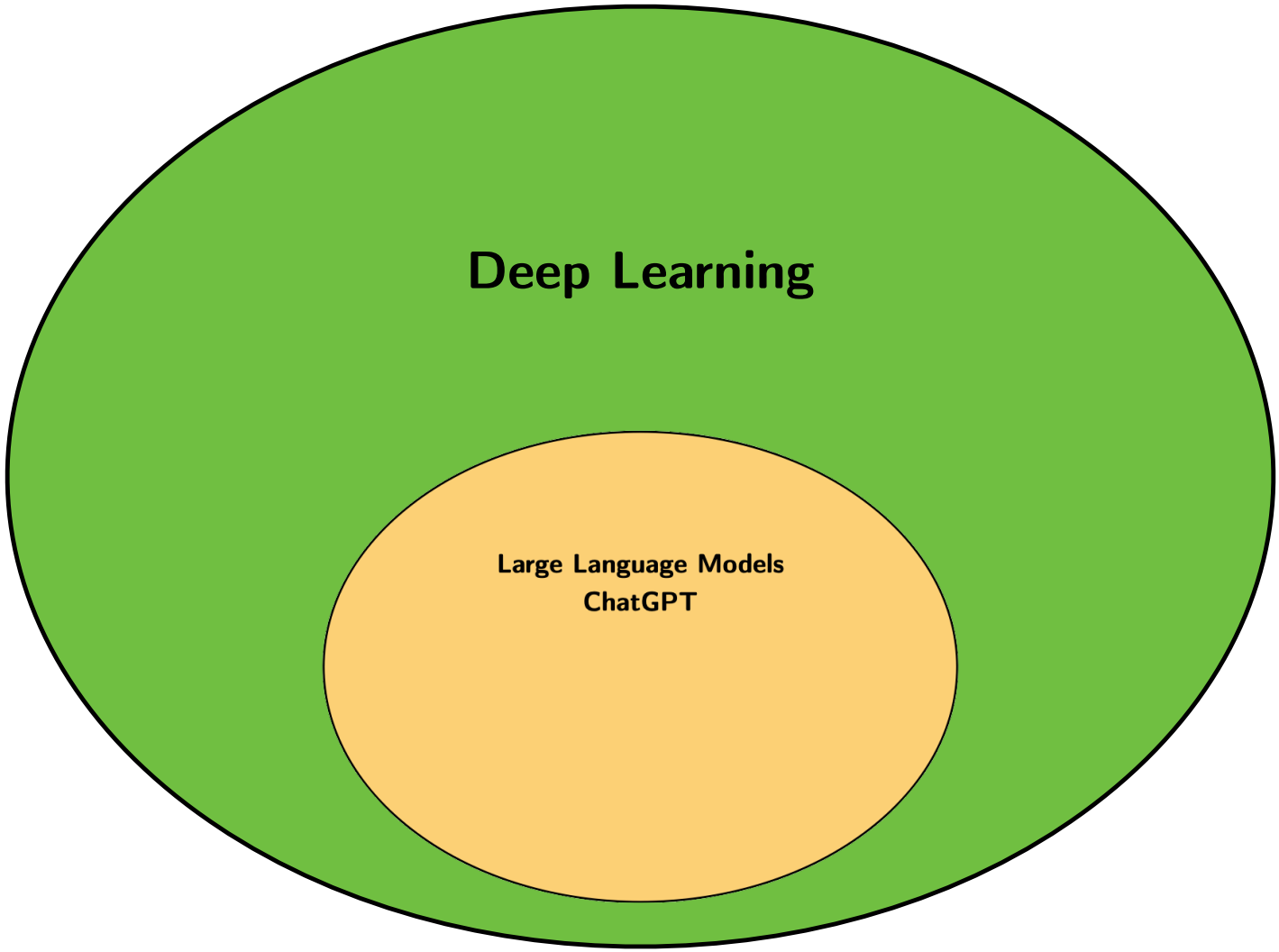


If LLM-based systems are not the future, what is?

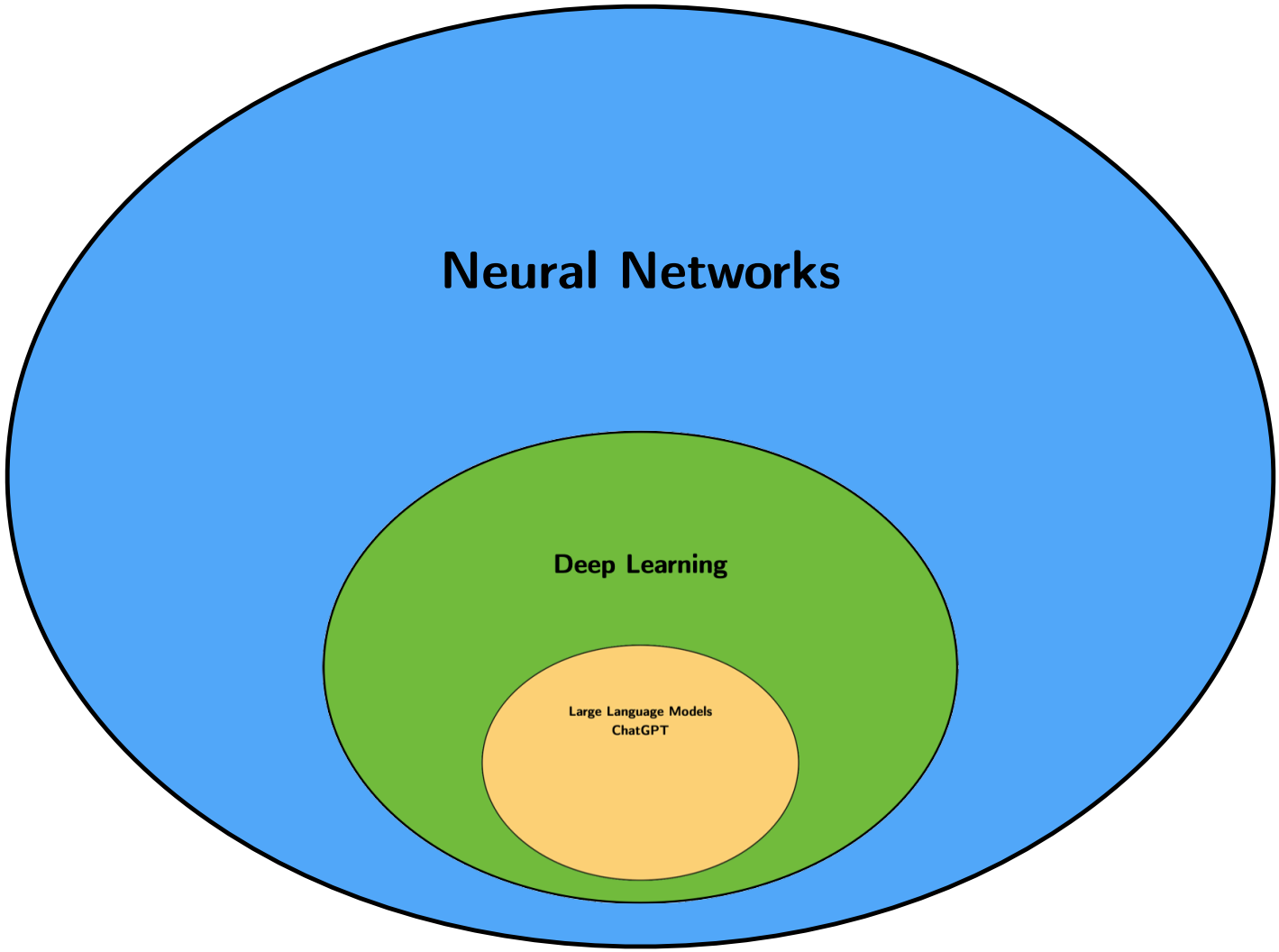
ChatGPT is not AI



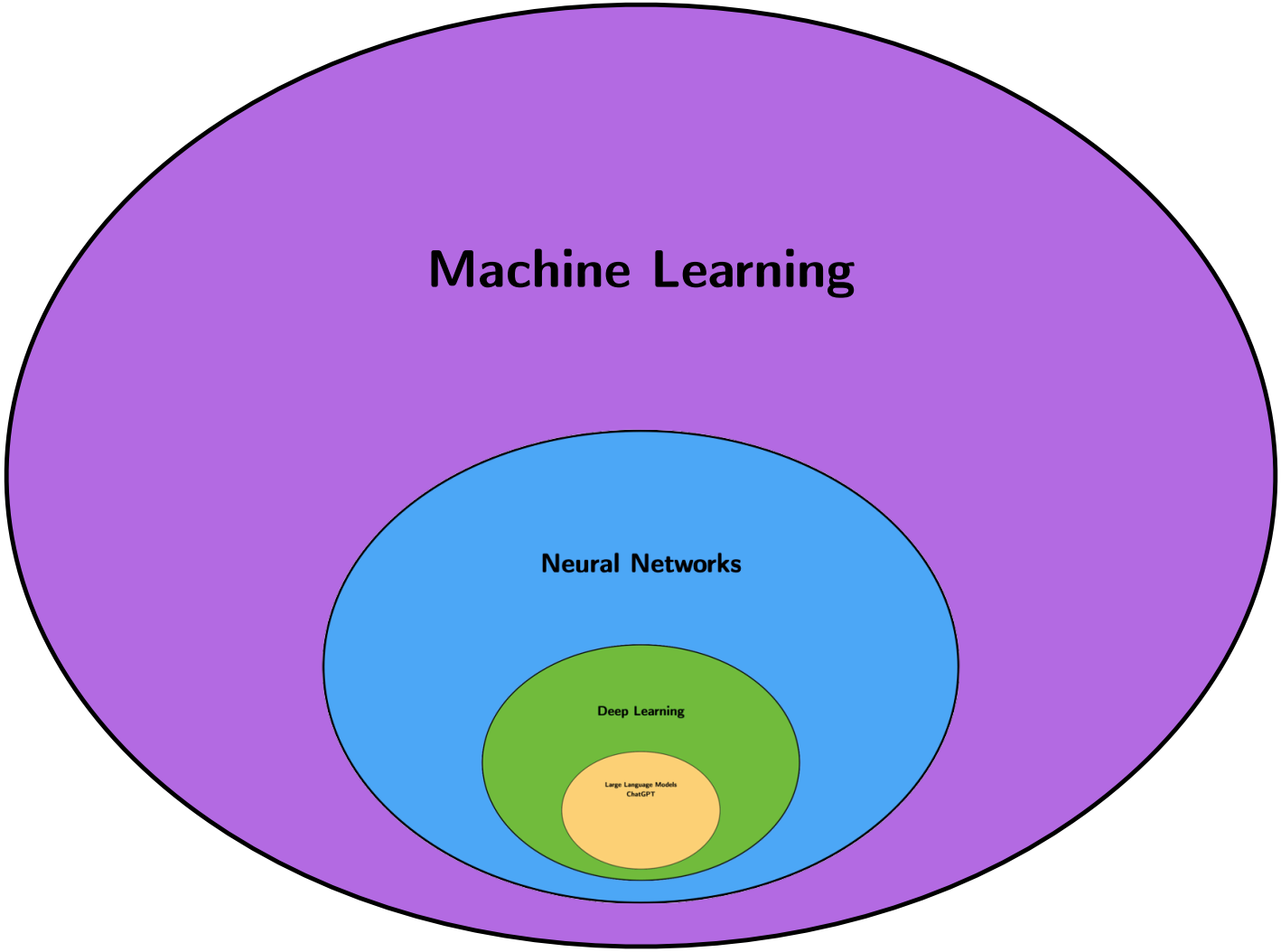
ChatGPT is not AI



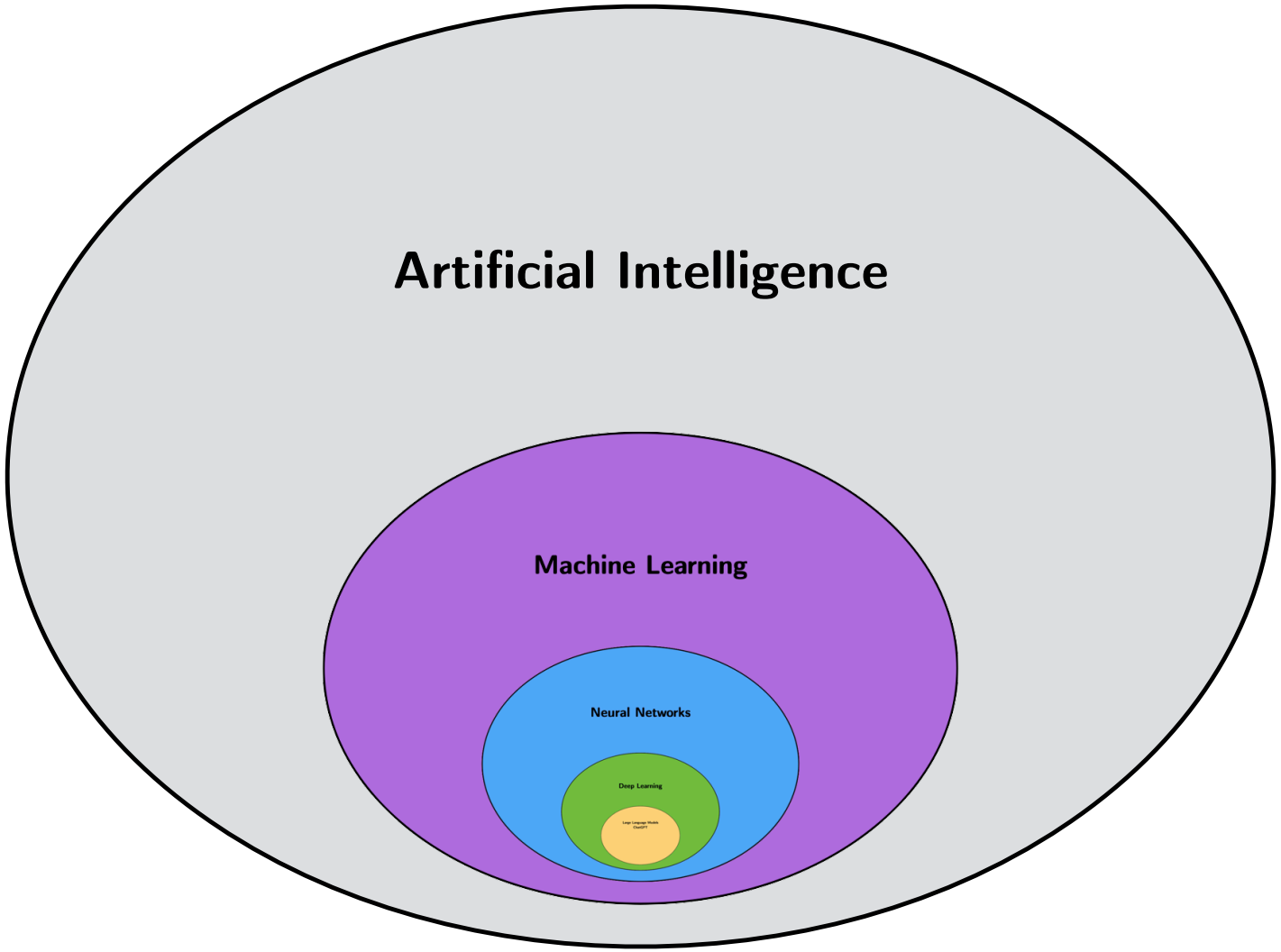
ChatGPT is not AI



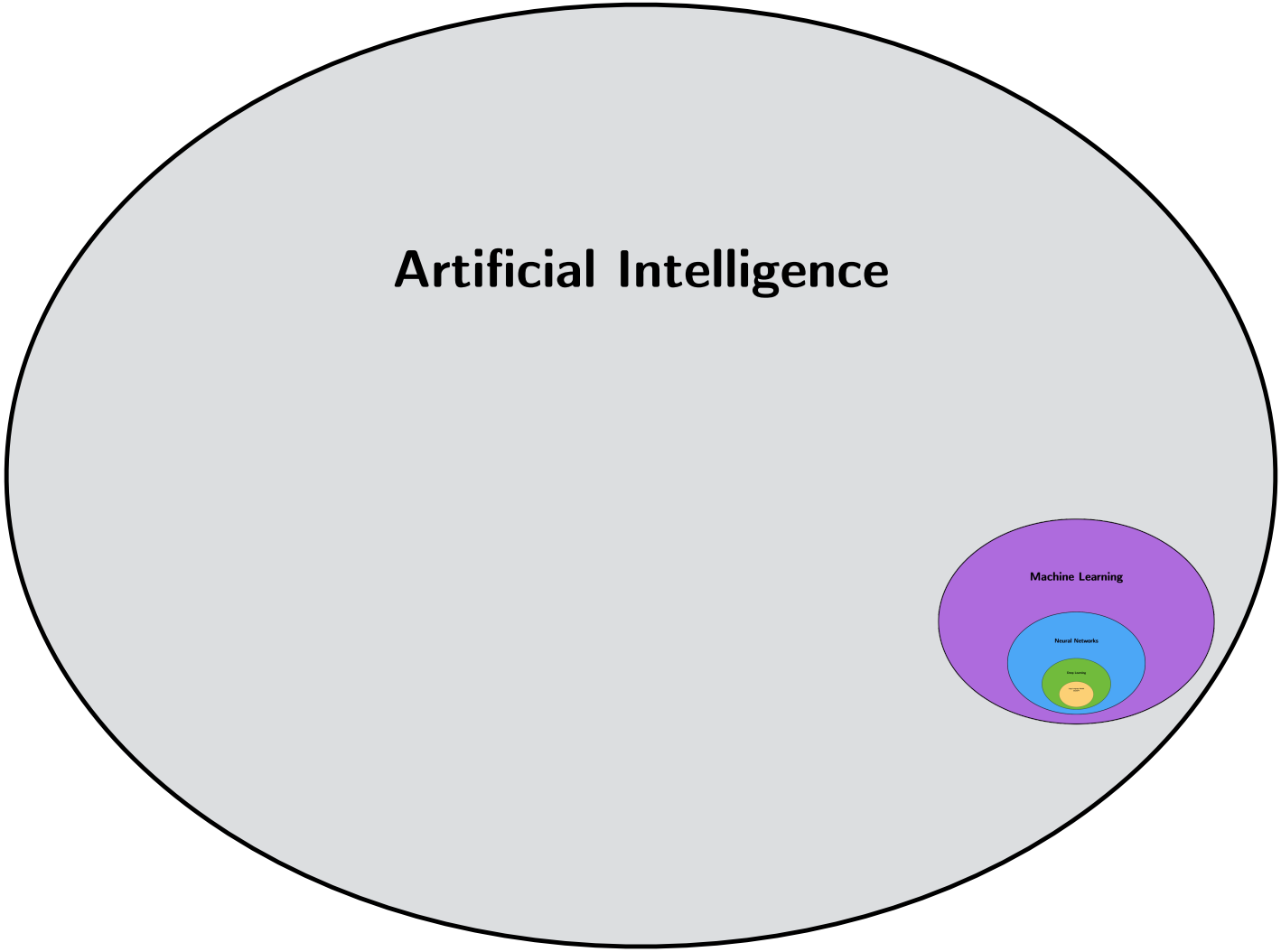
ChatGPT is not AI



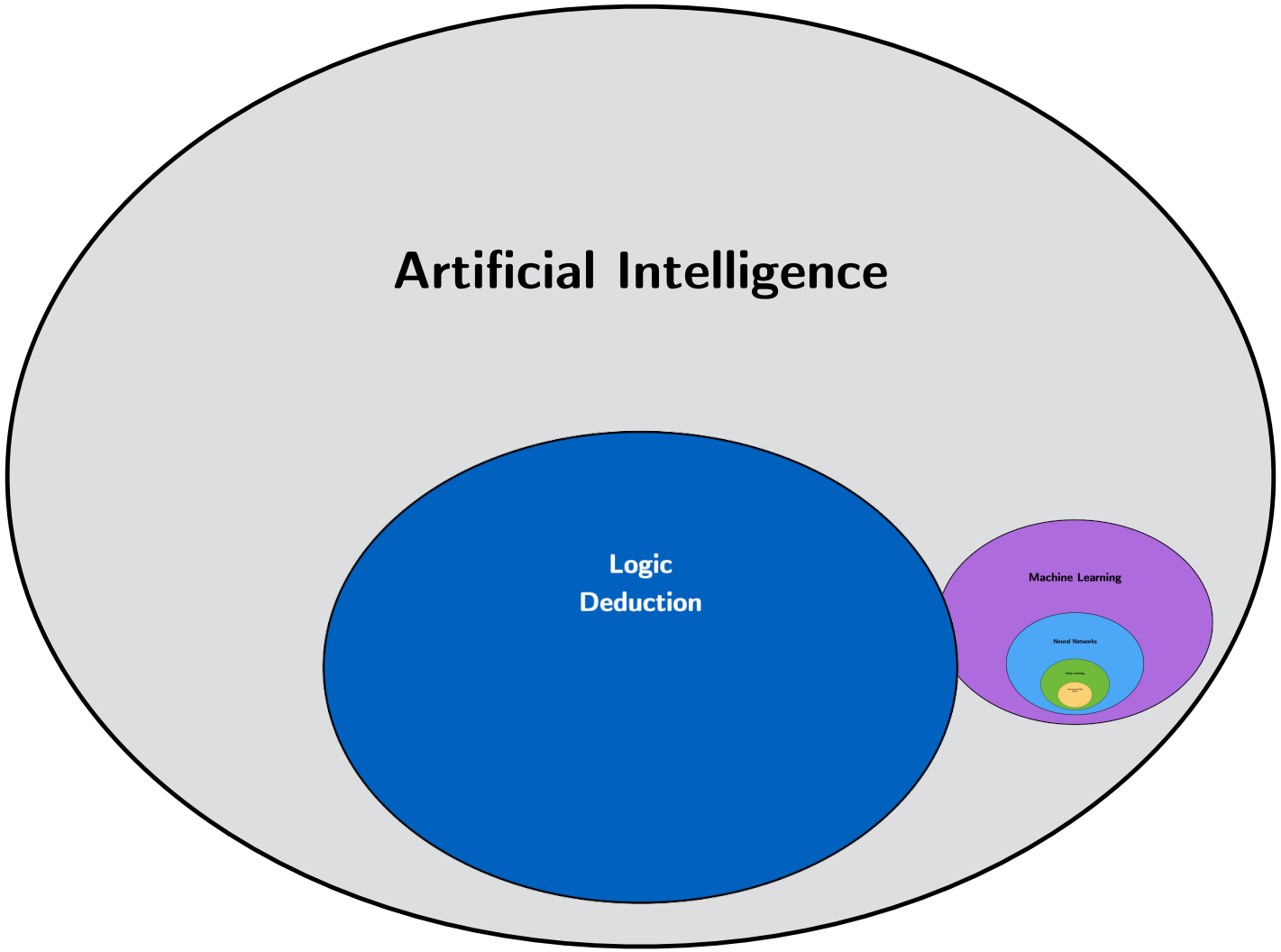
ChatGPT is not AI



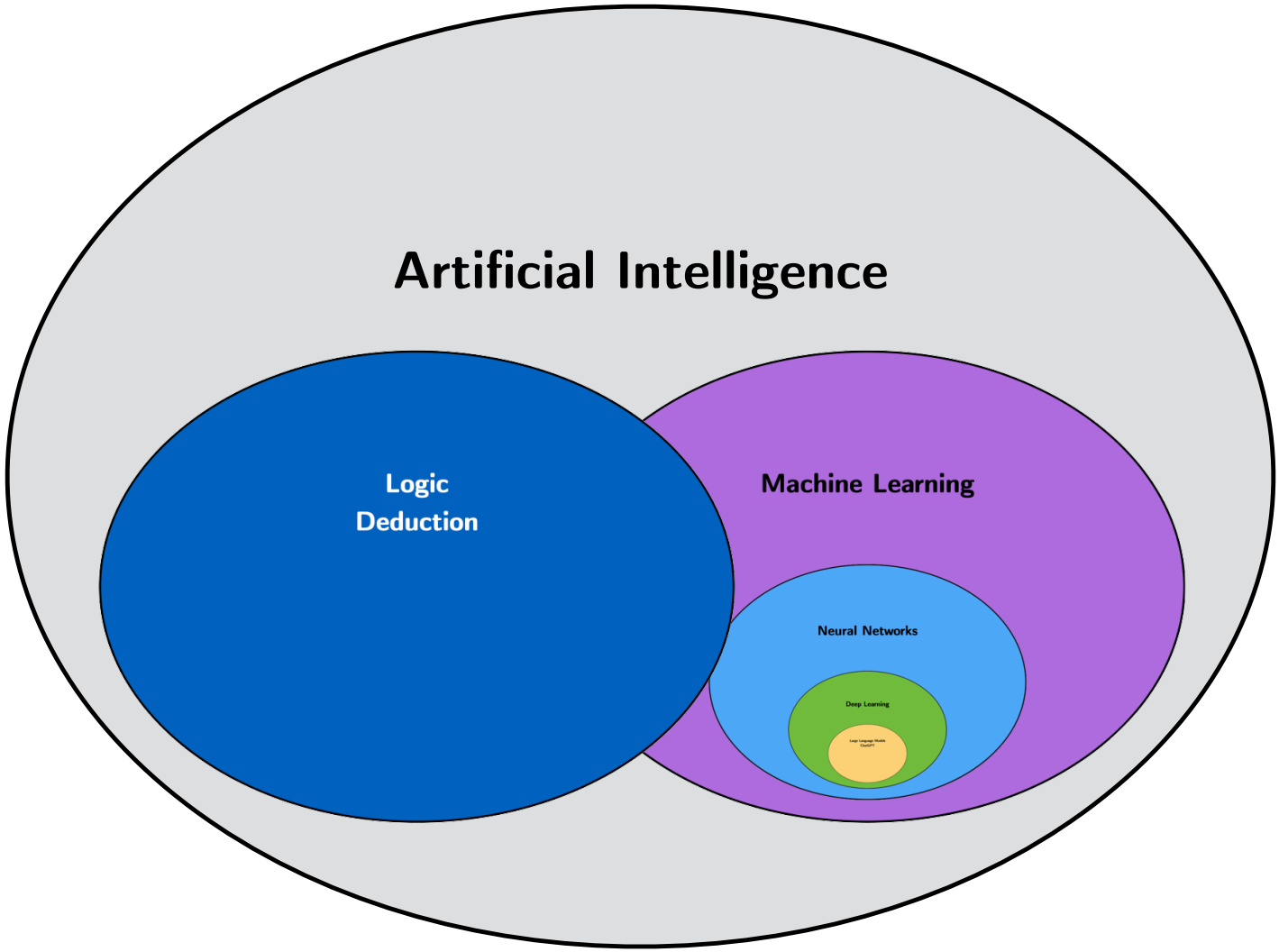
ChatGPT is not AI



ChatGPT is not AI



ChatGPT is not AI



2500 Years of History

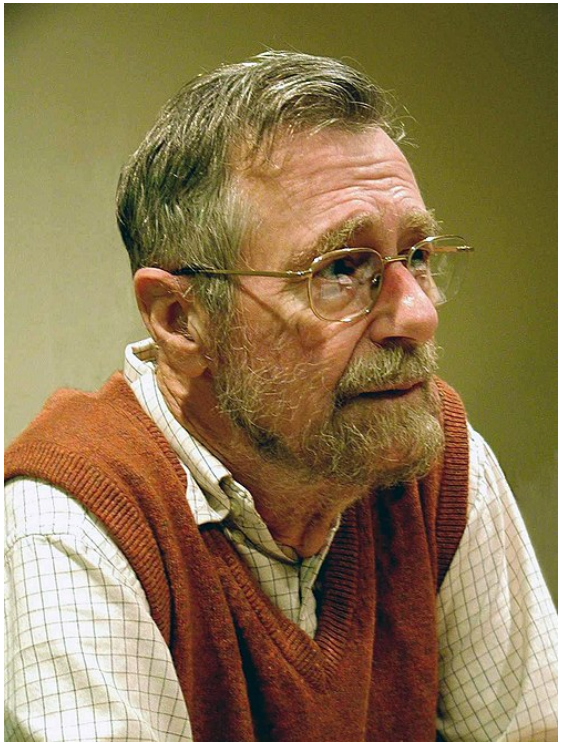
- LLMs are great at processing (natural) language
  - Natural language is amgiguous
  - Natural language is imprecise
- Mathematics has moved towards more formal language for its entire history
  - Euklid (formal definitions and arguments)
  - Al-Khwarizmi et al. (Algebraic formula language)
  - Newton and Leibniz (Formalisms for infinitesimals, *Calculemus*)
  - Bool, Frege, Church, . . . (Formal logics)
  - Russel and Whitehead (*Principia Mathematica*)
  - Mizar/Isabelle/HOL/Lean/FlySpeck . . .

2500 Years of History

- LLMs are great at processing (natural) language
  - Natural language is amgiguous
  - Natural language is imprecise
- Mathematics has moved towards more formal language for its entire history
  - Euklid (formal definitions and arguments)
  - Al-Khwarizmi et al. (Algebraic formula language)
  - Newton and Leibniz (Formalisms for infinitesimals, *Calculemus*)
  - Bool, Frege, Church, . . . (Formal logics)
  - Russel and Whitehead (*Principia Mathematica*)
  - Mizar/Isabelle/HOL/Lean/FlySpeck . . .

Future Mathematics will be based on formal logic!

Argument by Authority (and analogy)

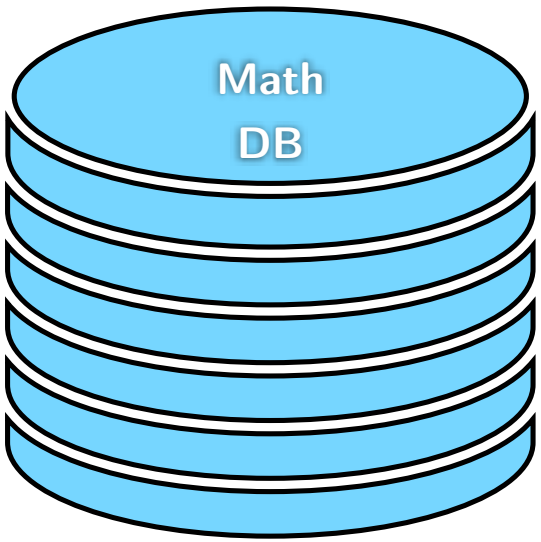


Edsger W. Dijkstra [Dij78]: *On the foolishness of "natural language programming"*

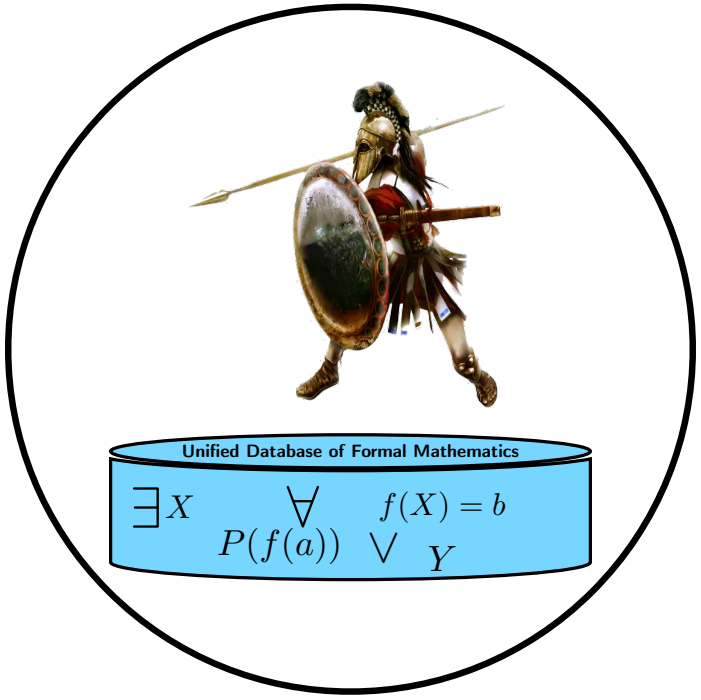
It may be illuminating to try to imagine what would have happened if, right from the start our native tongue would have been the only vehicle for the input into and the output from our information processing equipment. My considered guess is that history would, in a sense, have repeated itself, and that computer science would consist mainly of the indeed black art how to bootstrap from there to a sufficiently well-defined formal system.

The Future Formal Mathematics Database

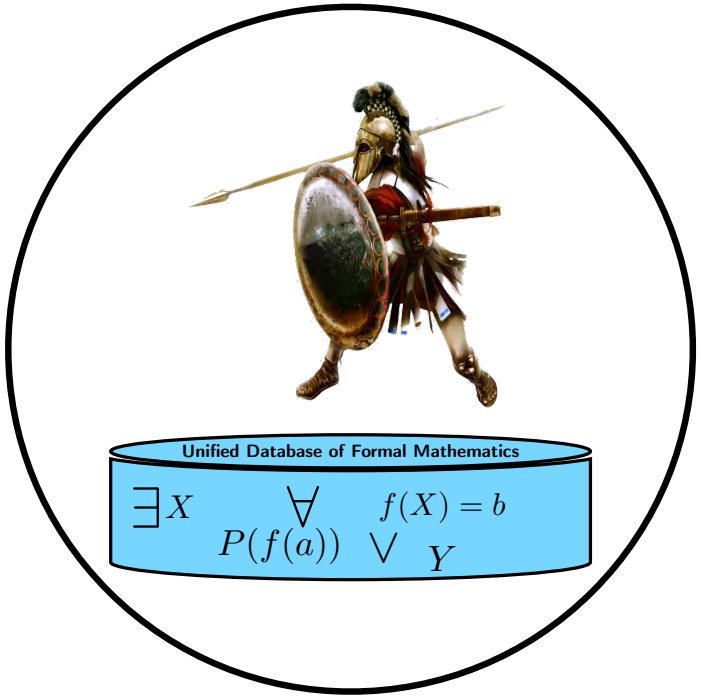
- Expressed in symbolic logic
  - Clear syntax
  - Clear semantics
- Unified (I can dream!)
  - One format (or compatible formats)
  - Globally consistent
- Structured
  - Different domains
  - Compositional
  - Searchable
- (Increasingly) comprehensive
  - *"The place to be"*
  - One size fits all (because it's flexible)



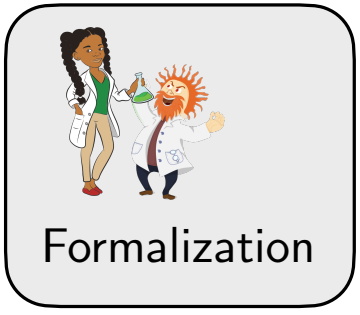
The Logic+ Architecture

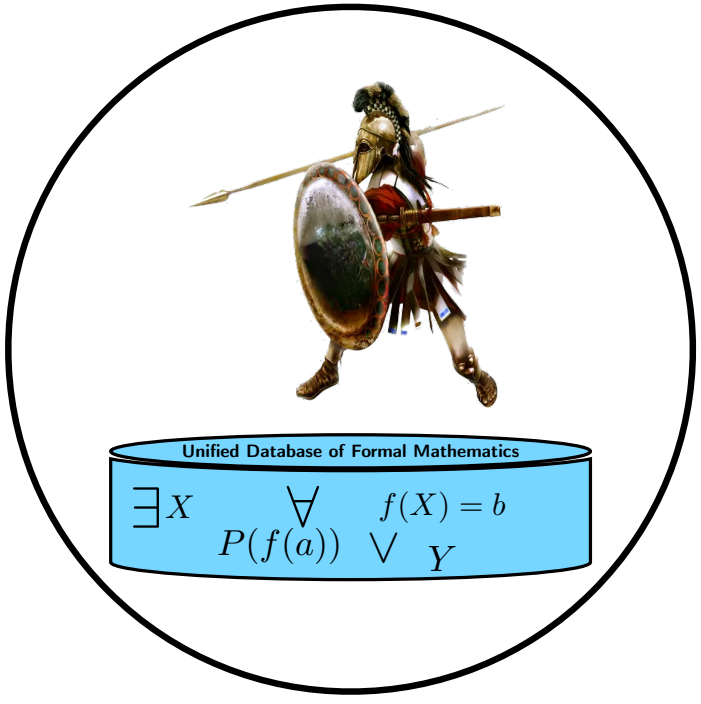


The Logic+ Architecture

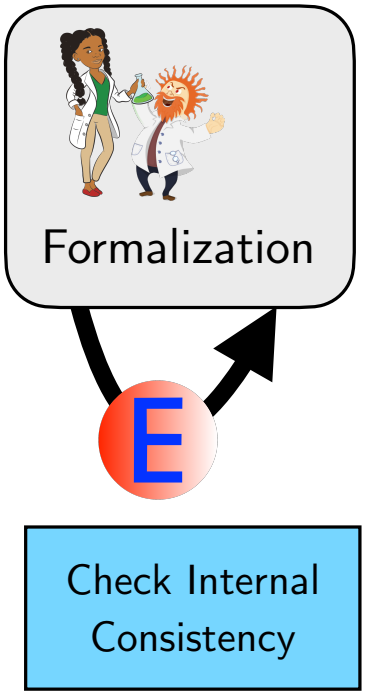


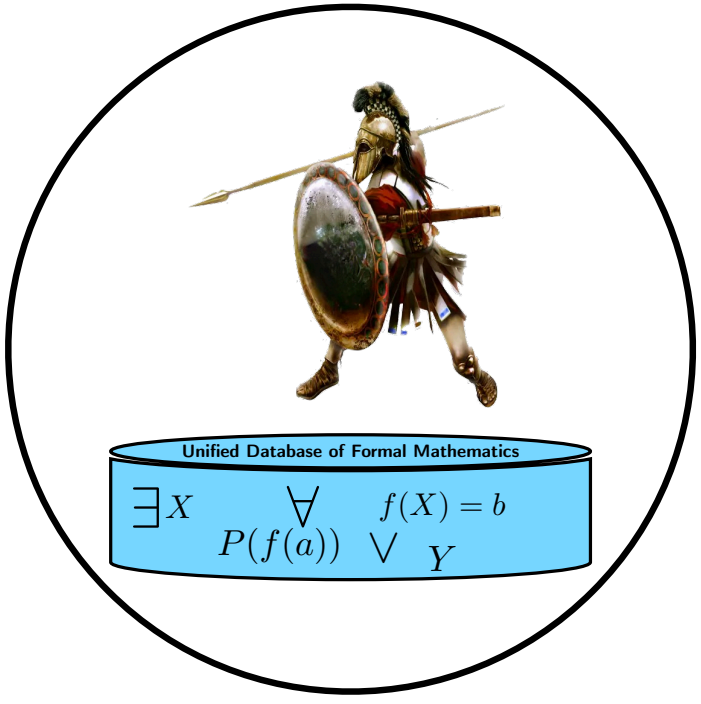
The Logic+ Architecture



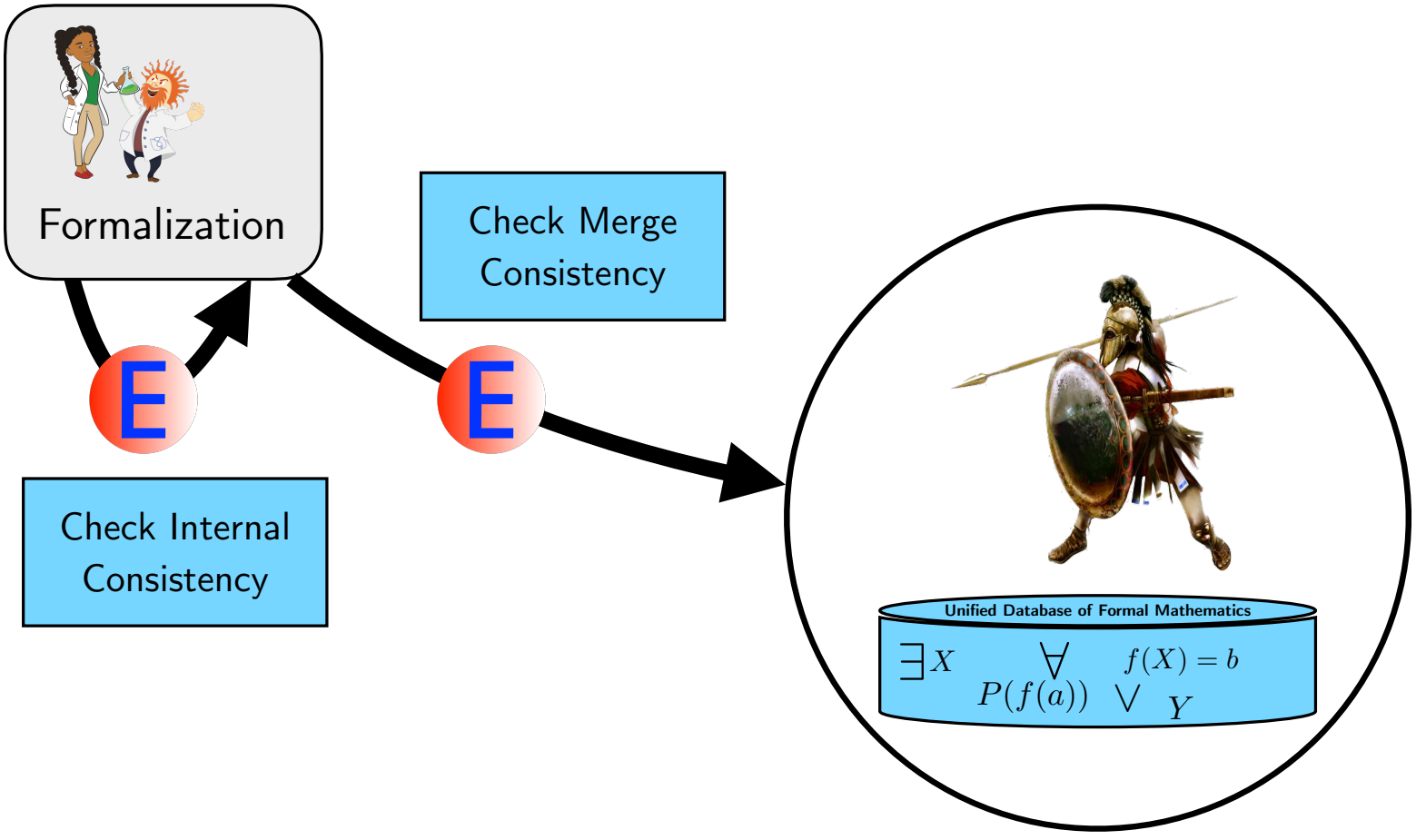


The Logic+ Architecture

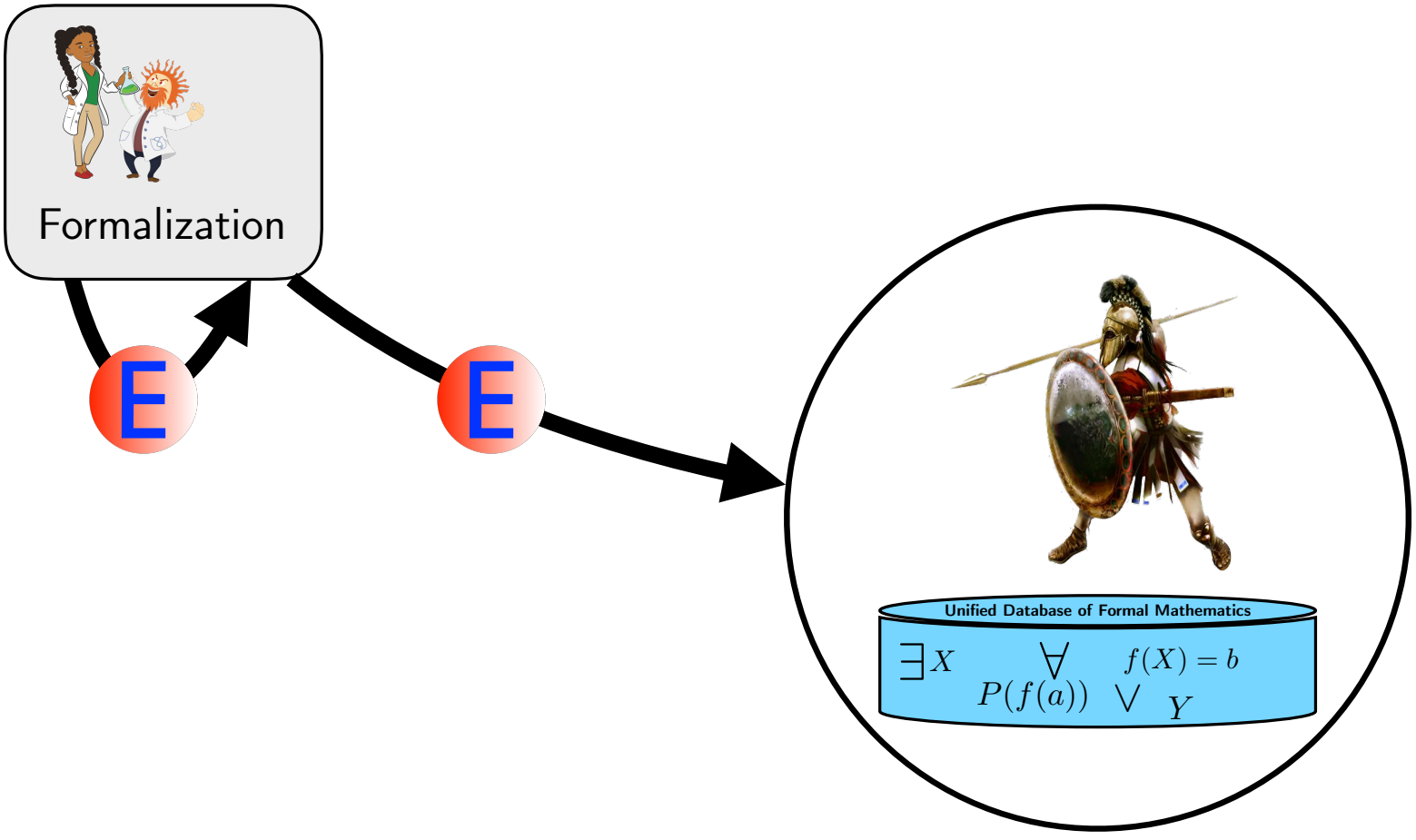




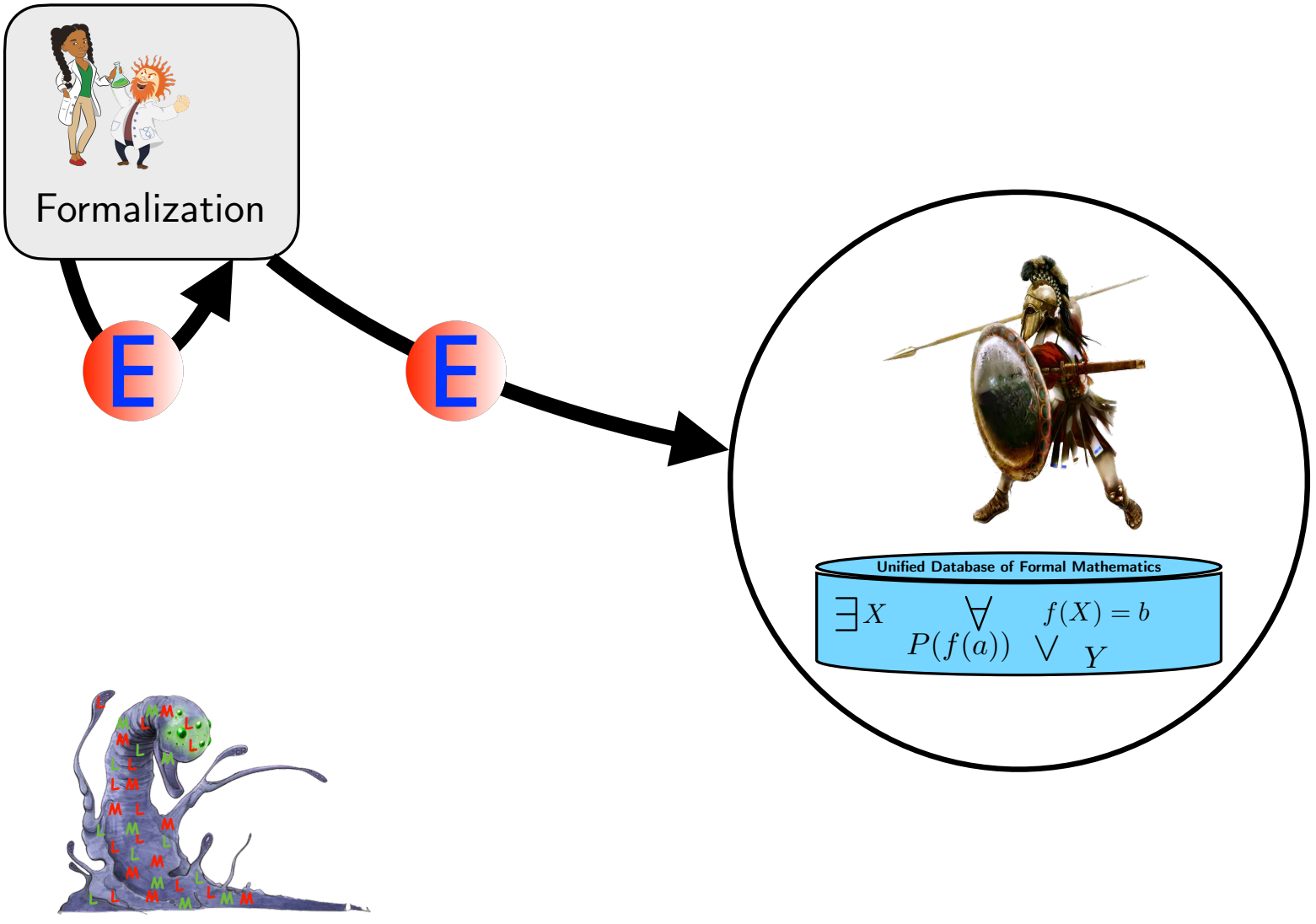
The Logic+ Architecture



The Logic+ Architecture



The Logic+ Architecture



The Logic+ Architecture

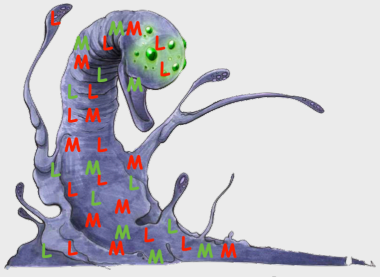
Formalization

E

E

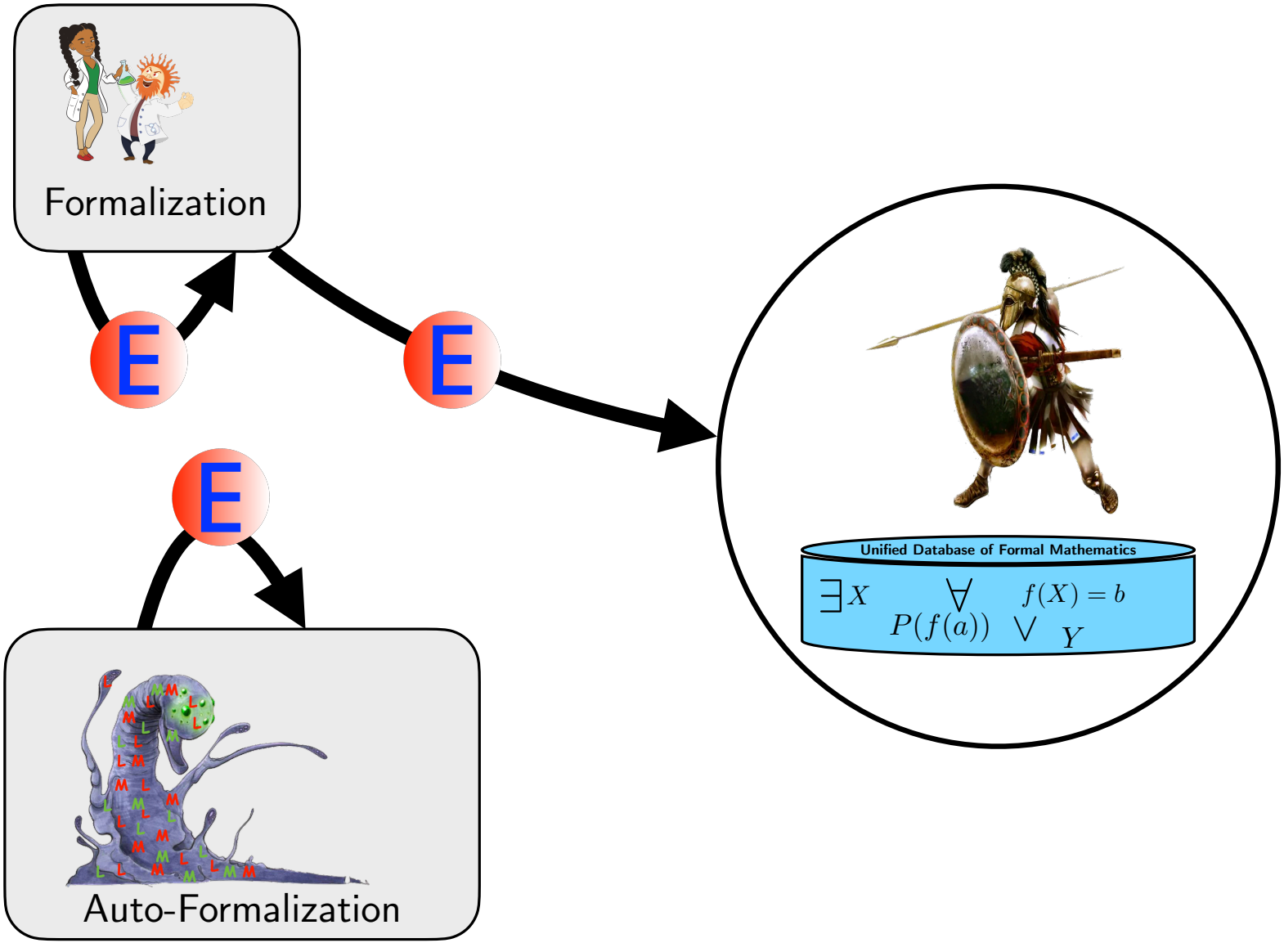
Unified Database of Formal Mathematics

$$\exists X \quad \forall \quad f(X) = b$$
$$P(f(a)) \quad \vee \quad Y$$

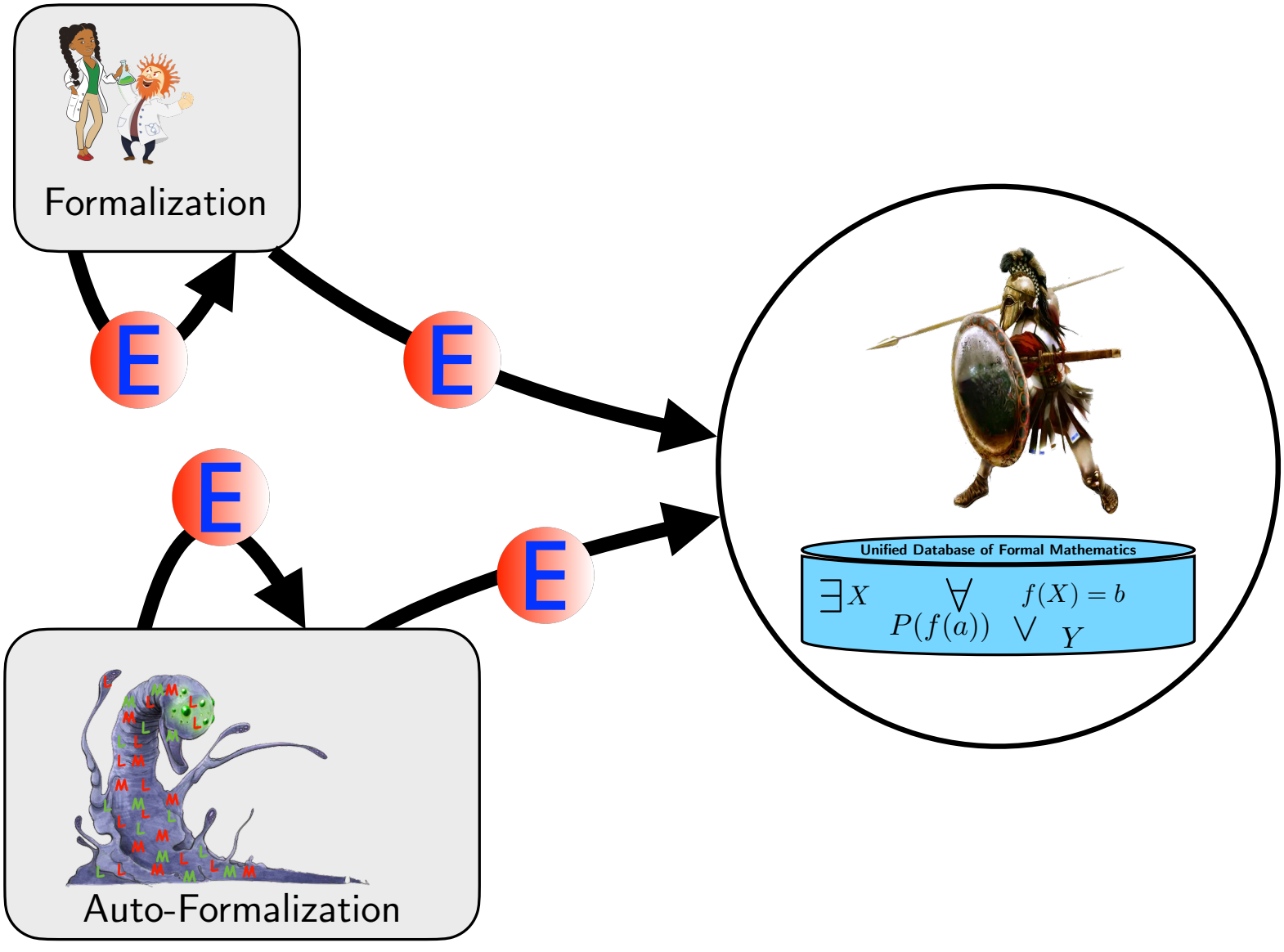


Auto-Formalization

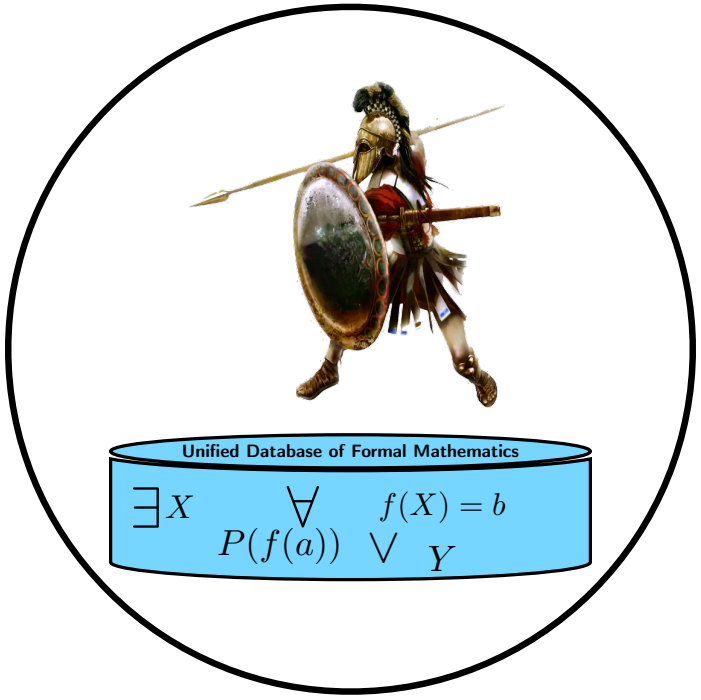
The Logic+ Architecture



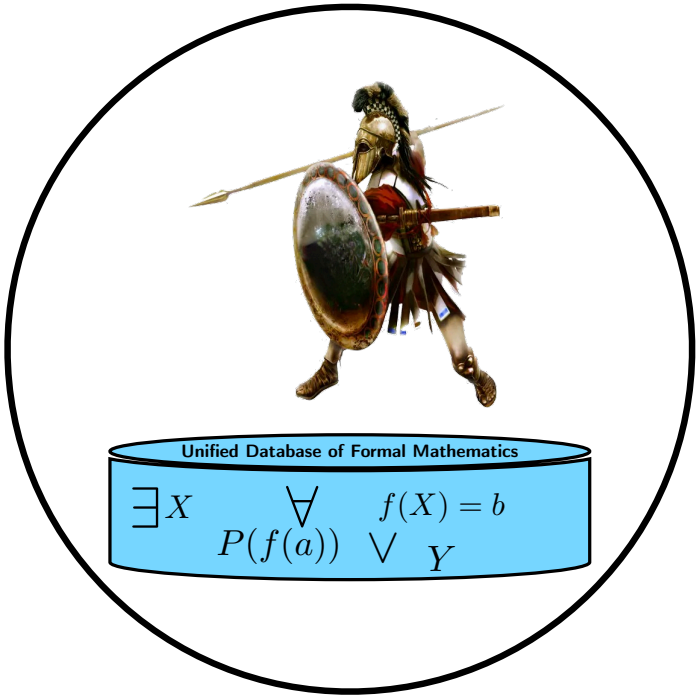
The Logic+ Architecture



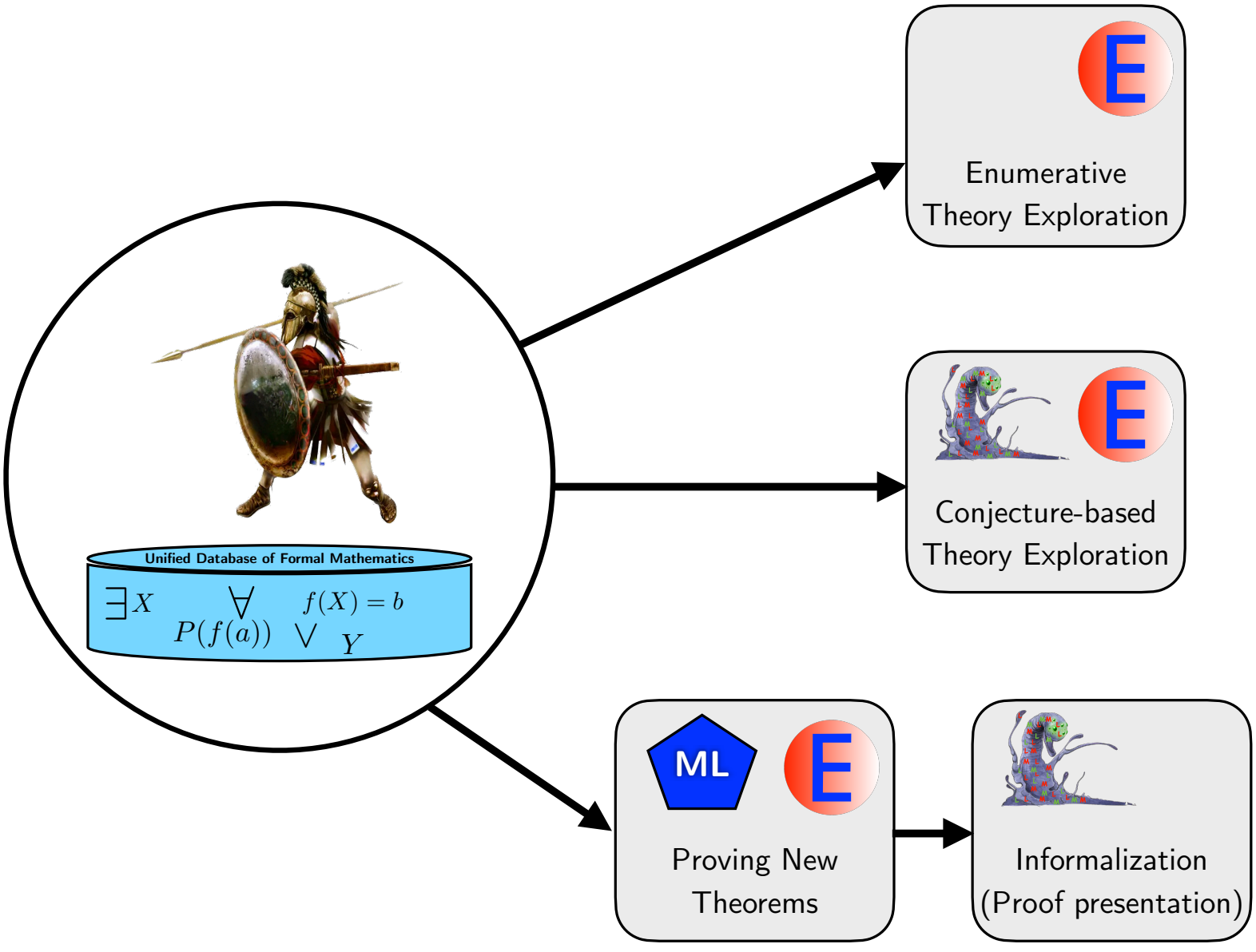
The Logic+ Architecture



The Logic+ Architecture



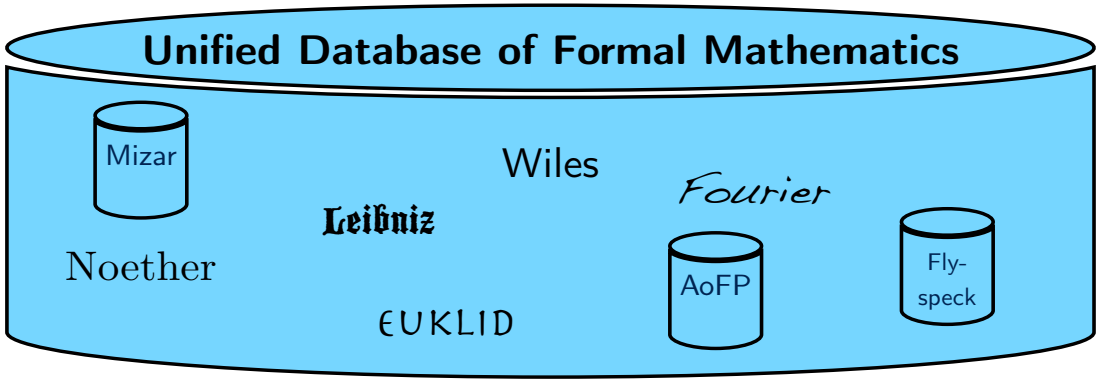
The Logic+ Architecture

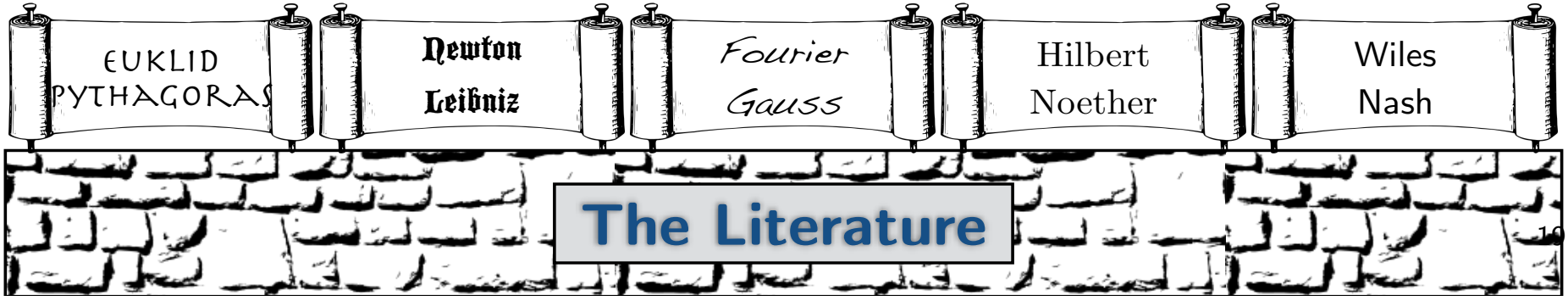


The future belongs to hybrid systems

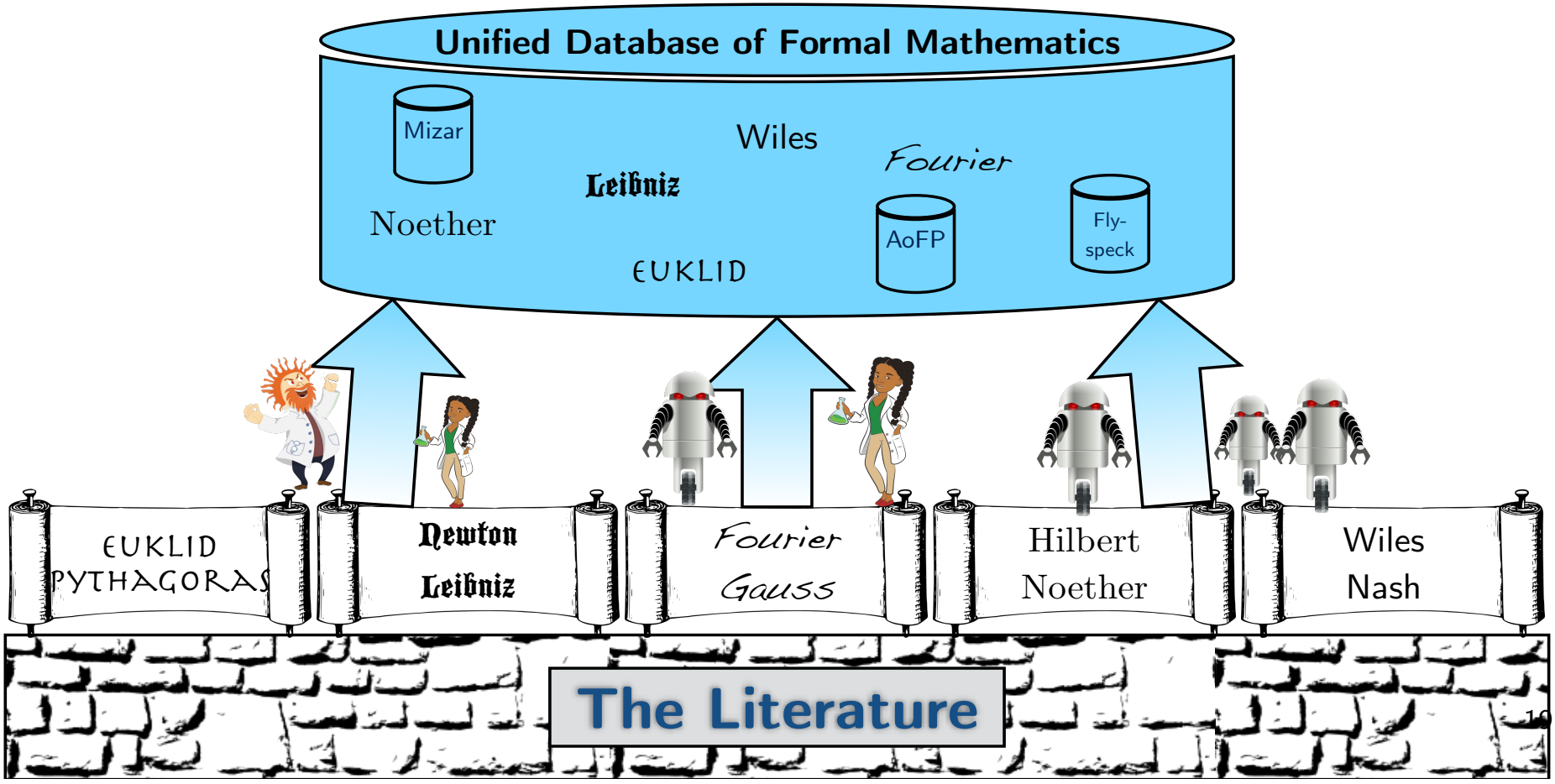
- Interactive theorem provers are the main front-end for human mathematicians
  - User-friendly syntax and editors
  - Access to mathematical libraries
- Automatic theorem provers boost productivity
  - . . . e.g. as Hammers for ITPs
  - . . . supporting theory exploration
- Automatic theorem provers perform quality control
  - Internal consistency
  - Merge consistency
- Machine learning improves the power of automated systems
  - Proof guidance
- LLM-based systems translate to and from symbolic logic
  - Auto-formalization

The Future Math Ecosystem (Utopia)

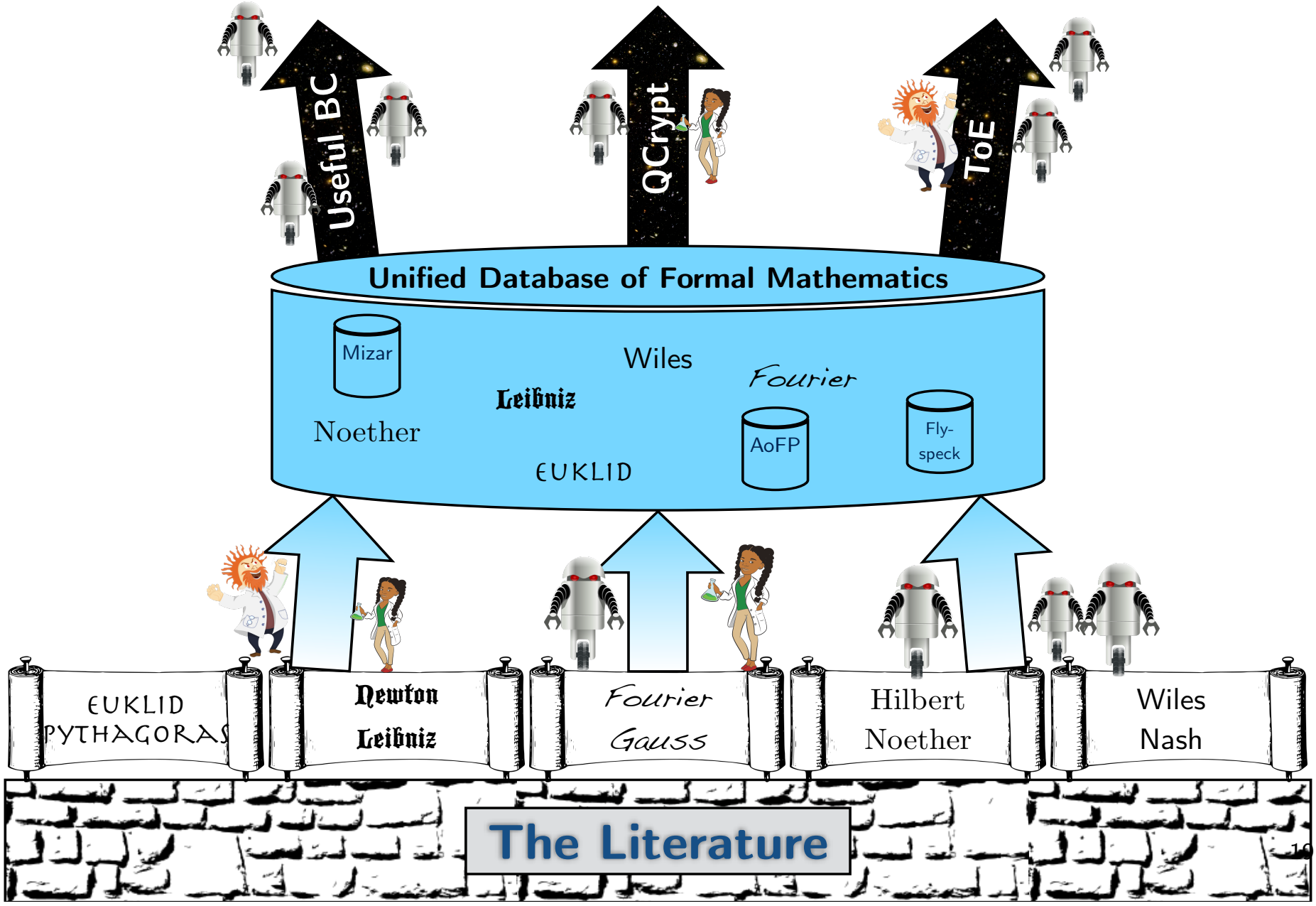




The Future Math Ecosystem (Utopia)

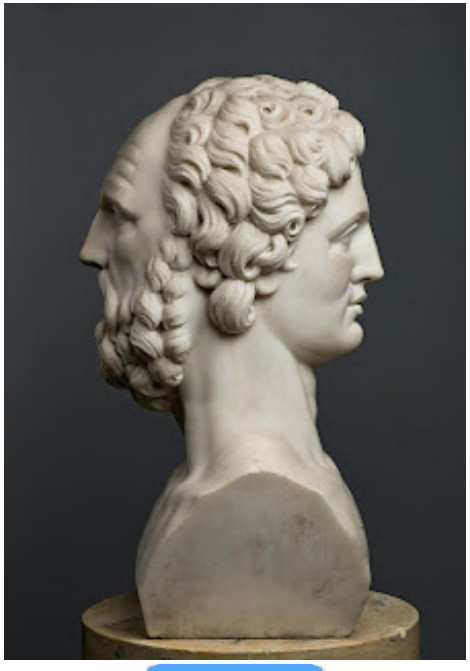


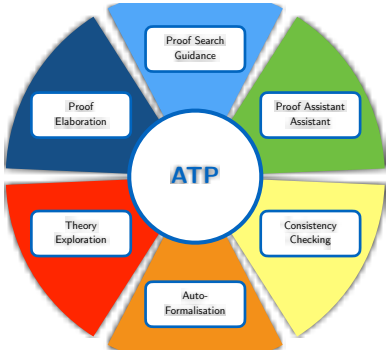
The Future Math Ecosystem (Utopia)



Conclusion

- Automated Reasoning and Machine Learning are complementary in many ways
  - The main formalism of computer maths will be symbolic
  - Deduction will use formal symbolic rules
- LLMs are not *the* way, but may contribute
  - . . . especially interfacing humans and machine
  - . . . especially translating different formalisms
- Other machine learning techniques may be more central
  - Nearest neighbors
  - Decision trees
  - Genetic algorithms
  - Tree-based ANN architectures





Bibliography

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla.

LLMs Will Always Hallucinate, and We Need to Live With This, 2024.

Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah.

Fidelity of Medical Reasoning in Large Language Models.

JAMA Network Open, 8(8):e2526021–e2526021, 08 2025.

Edsger W. Dijkstra.

On the foolishness of "natural language programming".

circulated privately, 1978.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar.

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, 2024.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar.

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, 2025.

Geoff Sutcliffe, Jack McKeown, and Alexander Steen.

A chat with bard.

In Konstantin Korovin, Stephan Schulz, and Michael Rawson, editors, *Proceedings of the 14th and 15th International Workshops on the Implementation of Logics*, volume 21 of *Kalpa Publications in Computing*, pages 8–14. EasyChair, 2025.

The END!