Title: LaTeXML and the Math-rich Scholarly Web

Speakers: Deyan Ginev

Collection/Series: Theory + AI Symposium

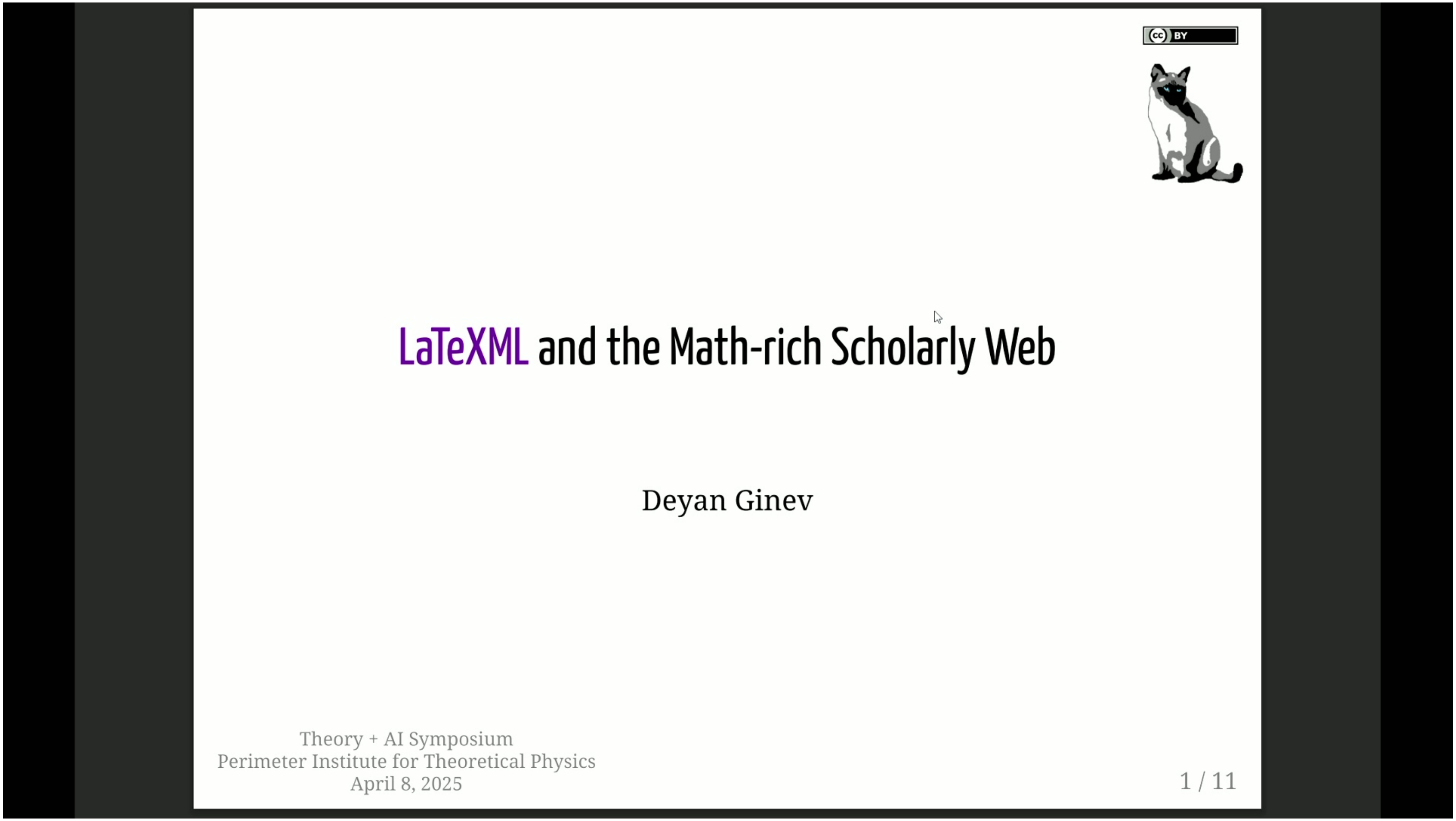
Date: April 08, 2025 - 11:30 AM

URL: https://pirsa.org/25040070

Abstract:

This short talk will outline some of LaTeXML's uses as infrastructure, as well as its enabling effect for search, AI, assistive technologies and the mobile web.

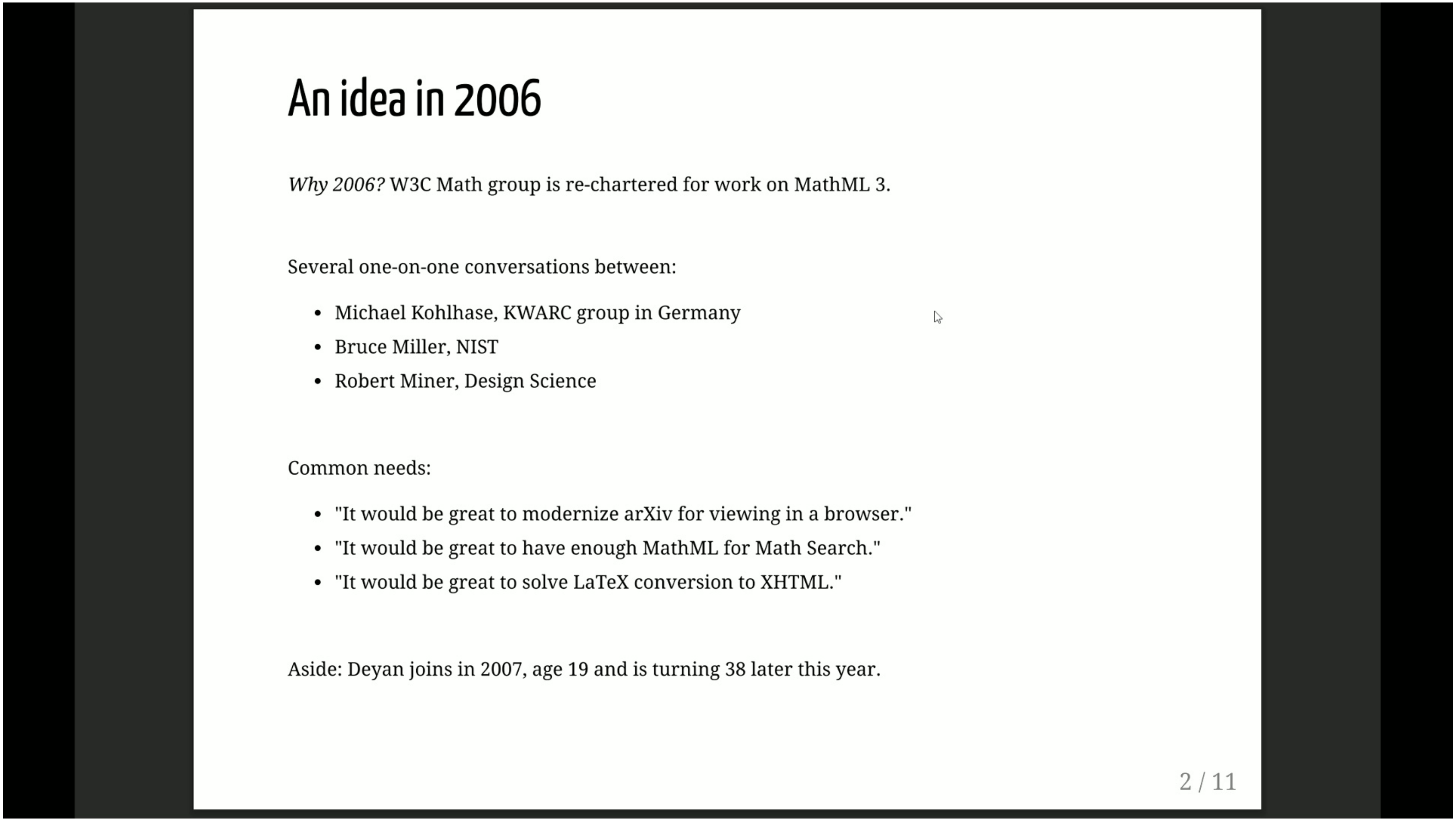
We have been on a journey towards scholarly articles with web-native mathematics since the dawn of the internet. The physics Open Science movement has led the way along with LaTeX, its authoring framework of choice. NIST's LaTeXML is a conversion tool that in the last twenty years has increasingly bridged that gap.



LaTeXML and the Math-rich Scholarly Web

Deyan Ginev

Theory + AI Symposium
Perimeter Institute for Theoretical Physics
April 8, 2025



An idea in 2006

Why 2006? W3C Math group is re-chartered for work on MathML 3.

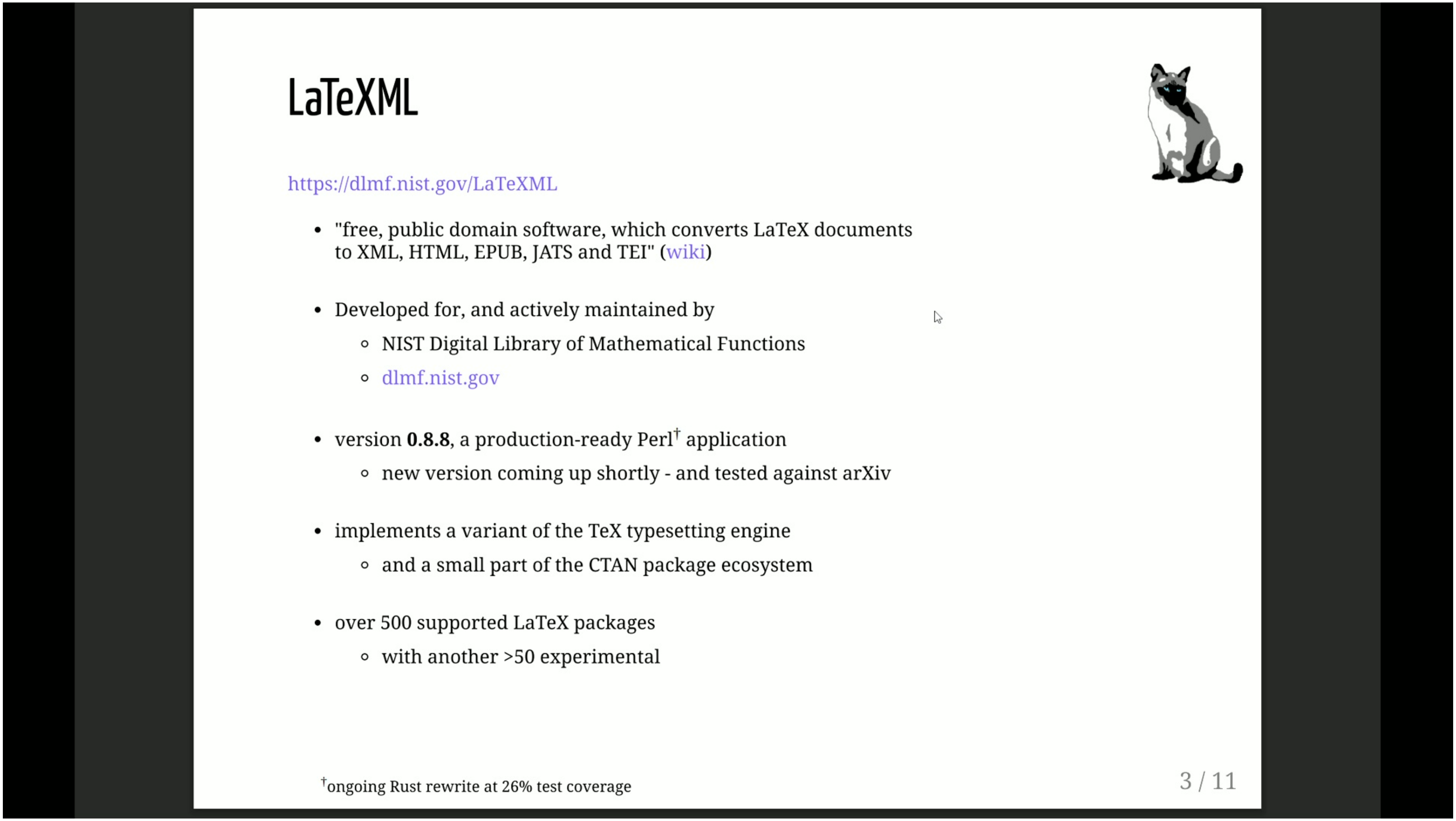
Several one-on-one conversations between:

- Michael Kohlhase, KWARC group in Germany
- Bruce Miller, NIST
- Robert Miner, Design Science

Common needs:

- "It would be great to modernize arXiv for viewing in a browser."
- "It would be great to have enough MathML for Math Search."
- "It would be great to solve LaTeX conversion to XHTML."

Aside: Deyan joins in 2007, age 19 and is turning 38 later this year.

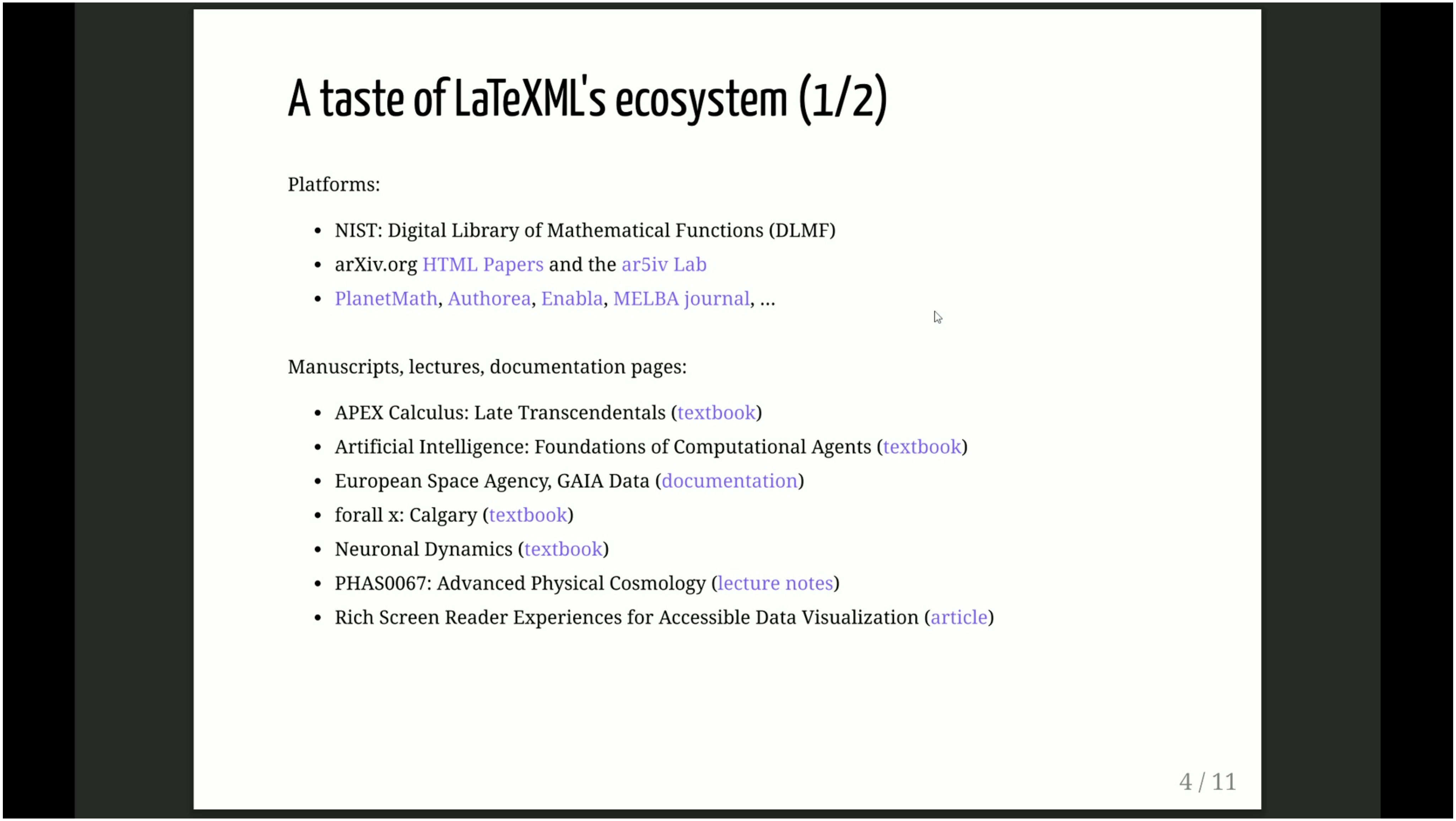


LaTeXML

https://dlmf.nist.gov/LaTeXML

- "free, public domain software, which converts LaTeX documents to XML, HTML, EPUB, JATS and TEI" (wiki)
- Developed for, and actively maintained by
  - NIST Digital Library of Mathematical Functions
  - dlmf.nist.gov
- version **0.8.8**, a production-ready Perl[†] application
  - new version coming up shortly - and tested against arXiv
- implements a variant of the TeX typesetting engine
  - and a small part of the CTAN package ecosystem
- over 500 supported LaTeX packages
  - with another >50 experimental

[†] ongoing Rust rewrite at 26% test coverage



A taste of LaTeXML's ecosystem (1/2)

Platforms:

- NIST: Digital Library of Mathematical Functions (DLMF)
- arXiv.org HTML Papers and the ar5iv Lab
- PlanetMath, Authorea, Enabla, MELBA journal, ...

Manuscripts, lectures, documentation pages:

- APEX Calculus: Late Transcendentals (textbook)
- Artificial Intelligence: Foundations of Computational Agents (textbook)
- European Space Agency, GAIA Data (documentation)
- forall x: Calgary (textbook)
- Neuronal Dynamics (textbook)
- PHAS0067: Advanced Physical Cosmology (lecture notes)
- Rich Screen Reader Experiences for Accessible Data Visualization (article)

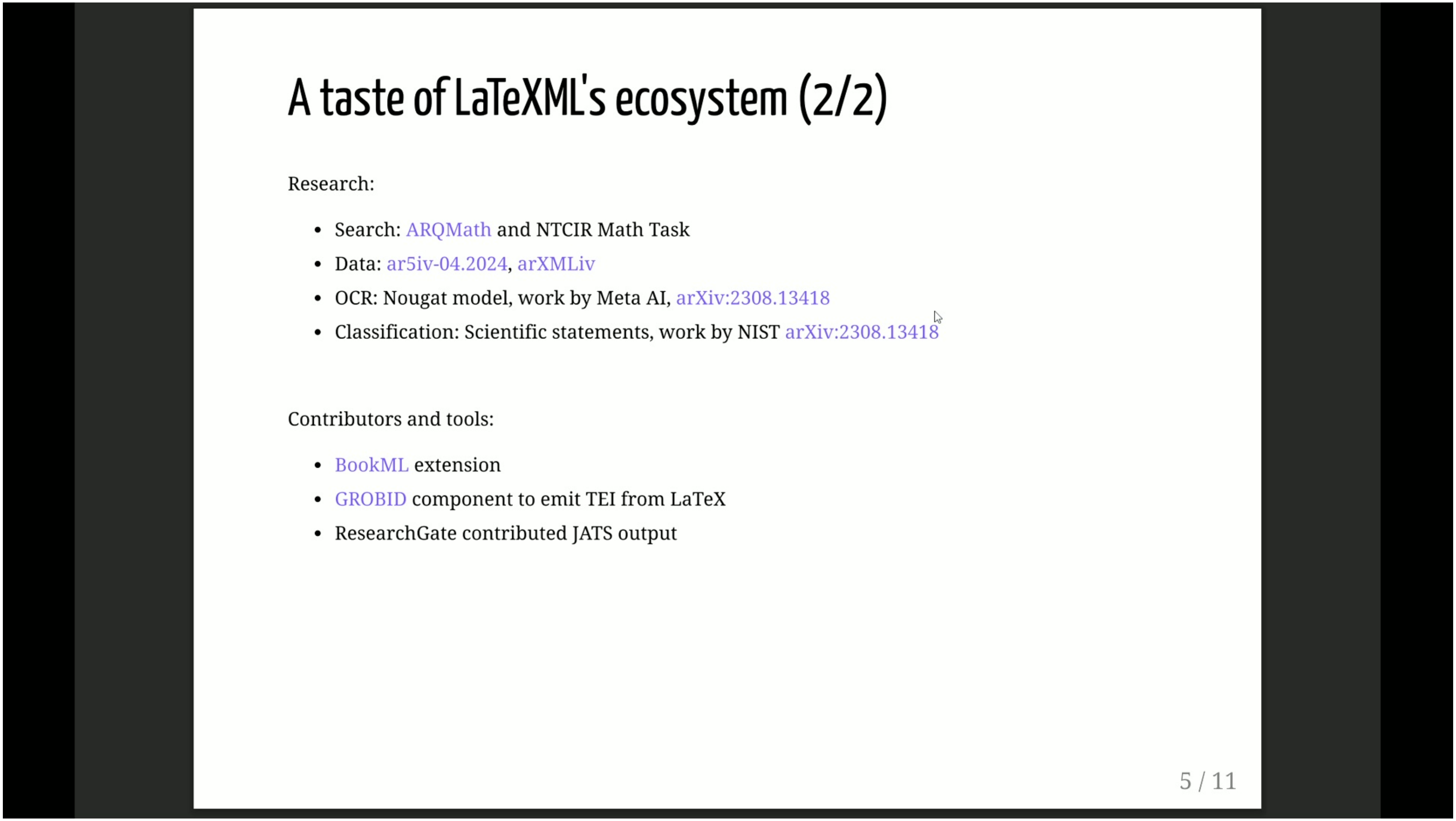
TLDR: What is ar5iv?

An official preview site for arXiv.org as HTML5 (with MathML)

2.4 million e-print documents; updated monthly

HTML for 97% of all LaTeX sources in arXiv, 74% error-free

available at ar5iv.labs.arxiv.org



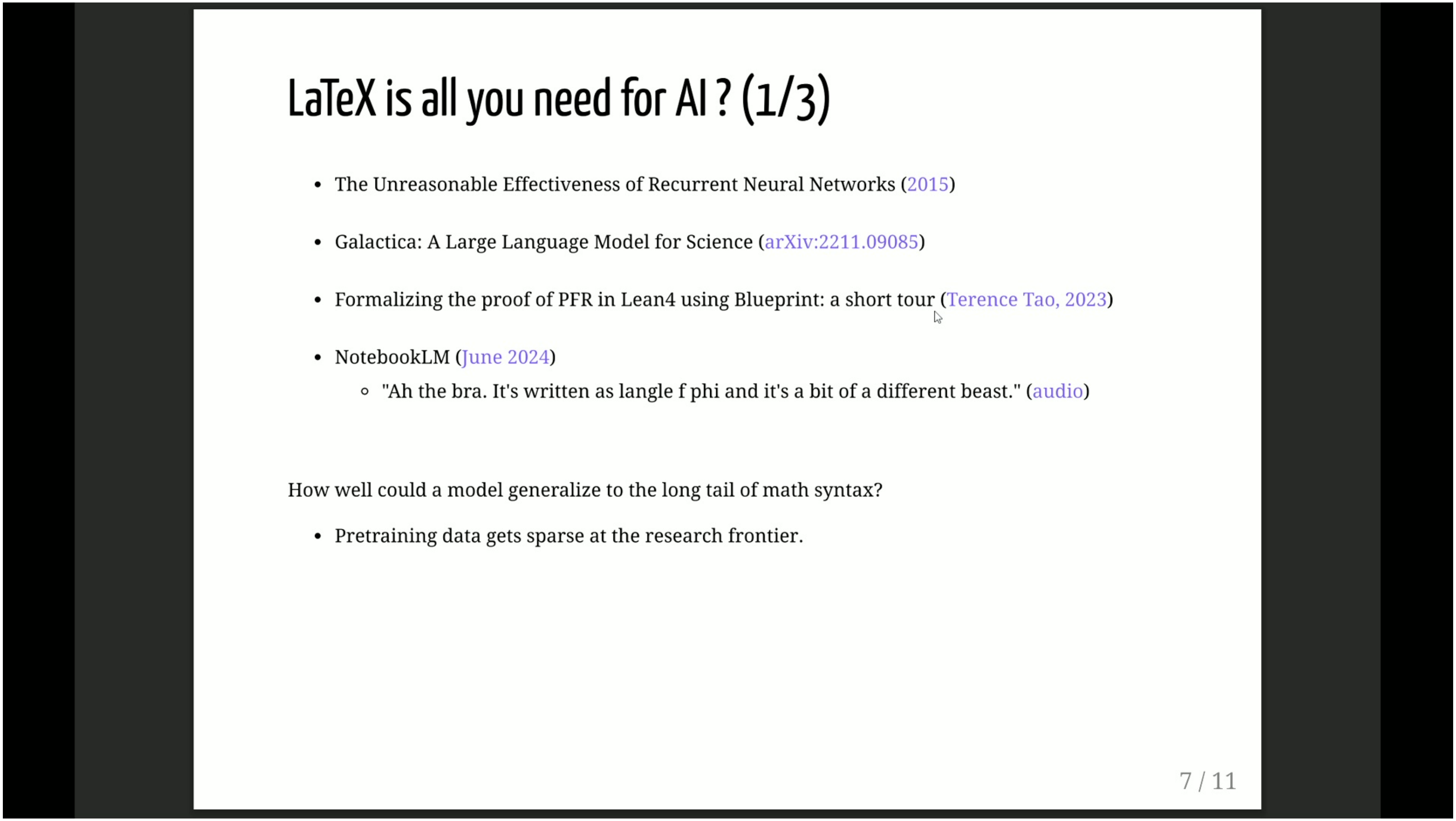
A taste of LaTeXML's ecosystem (2/2)

Research:

- Search: ARQMath and NTCIR Math Task
- Data: ar5iv-04.2024, arXMLiv
- OCR: Nougat model, work by Meta AI, arXiv:2308.13418
- Classification: Scientific statements, work by NIST arXiv:2308.13418

Contributors and tools:

- BookML extension
- GROBID component to emit TEI from LaTeX
- ResearchGate contributed JATS output



LaTeX is all you need for AI ? (1/3)

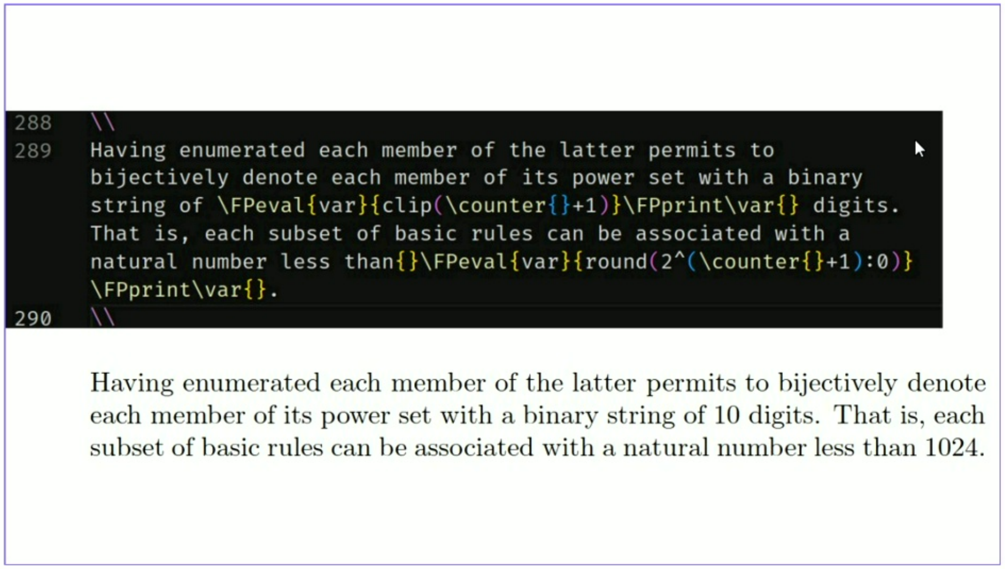
- The Unreasonable Effectiveness of Recurrent Neural Networks (2015)
- Galactica: A Large Language Model for Science (arXiv:2211.09085)
- Formalizing the proof of PFR in Lean4 using Blueprint: a short tour (Terence Tao, 2023)
- NotebookLM (June 2024)
    - "Ah the bra. It's written as langle f phi and it's a bit of a different beast." (audio)

How well could a model generalize to the long tail of math syntax?

- Pretraining data gets sparse at the research frontier.

LaTeX is all you need for AI/web/search/a11y? (2/3)

As seen on arXiv.org

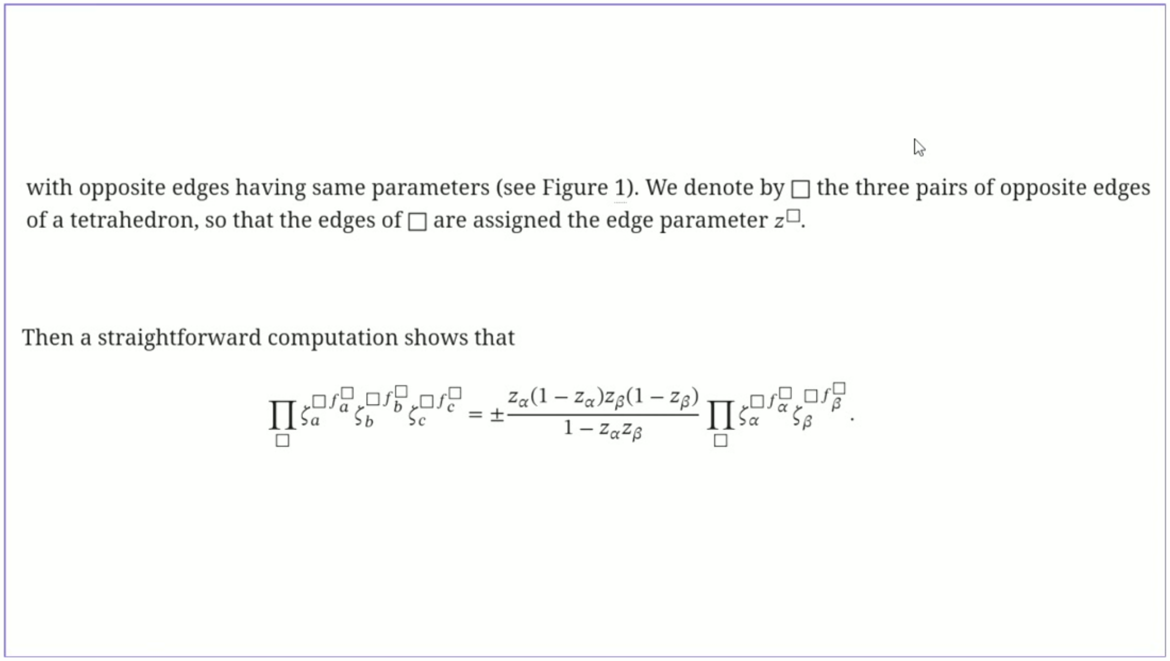


```
\\
Having enumerated each member of the latter permits to
bijectively denote each member of its power set with a binary
string of \FPeval{var}{clip(\counter{}+1)}\FPprint\var{} digits.
That is, each subset of basic rules can be associated with a
natural number less than{}\FPeval{var}{round(2^(\counter{}+1):0)}
\FPprint\var{}.
\\
```

Having enumerated each member of the latter permits to bijectively denote each member of its power set with a binary string of 10 digits. That is, each subset of basic rules can be associated with a natural number less than 1024.

LaTeX is all you need for AI/web/search/a11y? (3/3)

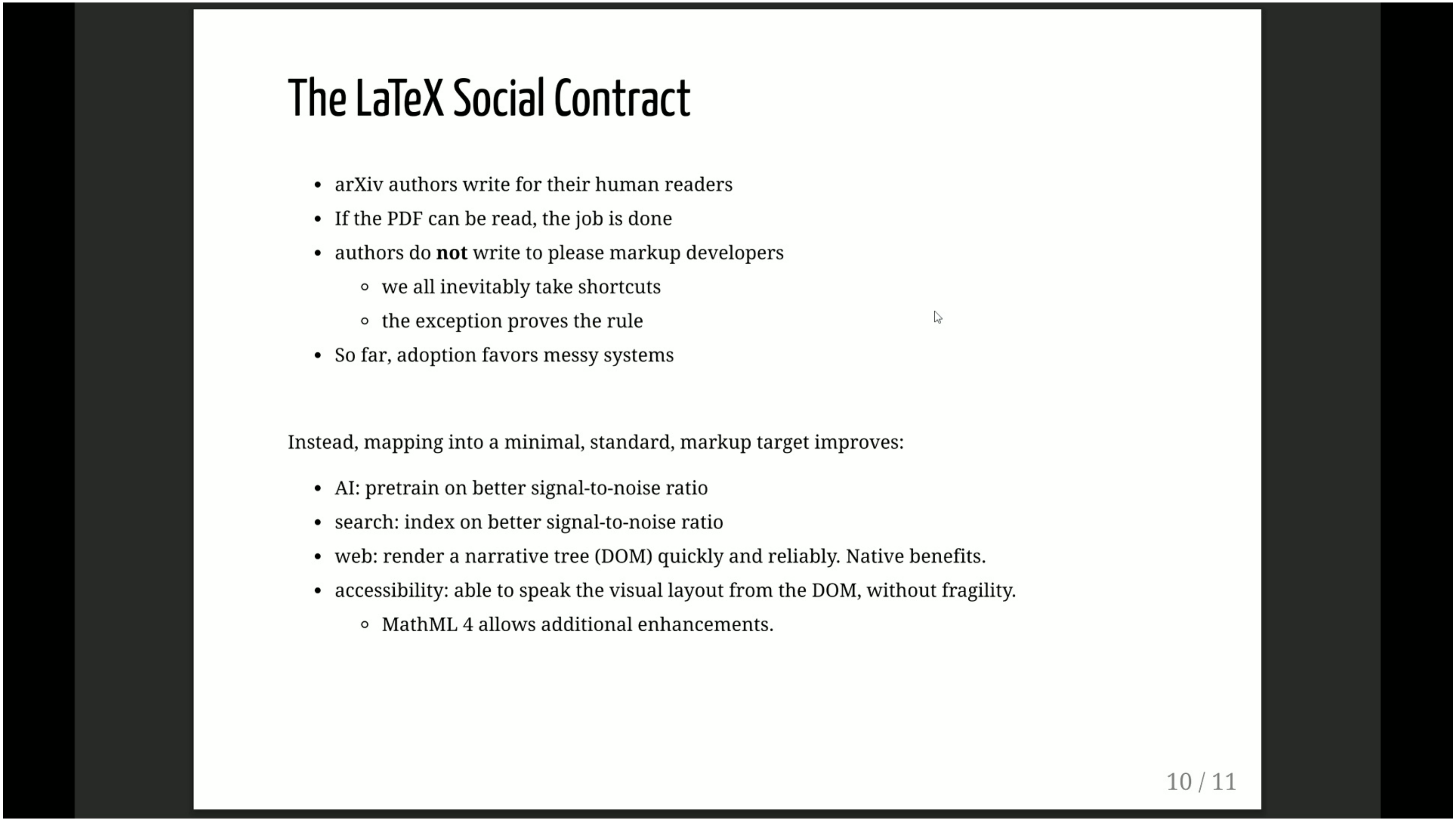
As seen on arXiv.org



with opposite edges having same parameters (see Figure 1). We denote by □ the three pairs of opposite edges of a tetrahedron, so that the edges of □ are assigned the edge parameter $z^{\square}$.

Then a straightforward computation shows that

$$\prod_{\square} \zeta_a^{\square f_a^{\square}} \zeta_b^{\square f_b^{\square}} \zeta_c^{\square f_c^{\square}} = \pm \frac{z_\alpha(1-z_\alpha)z_\beta(1-z_\beta)}{1-z_\alpha z_\beta} \prod_{\square} \zeta_\alpha^{\square f_\alpha^{\square}} \zeta_\beta^{\square f_\beta^{\square}} .$$



The LaTeX Social Contract

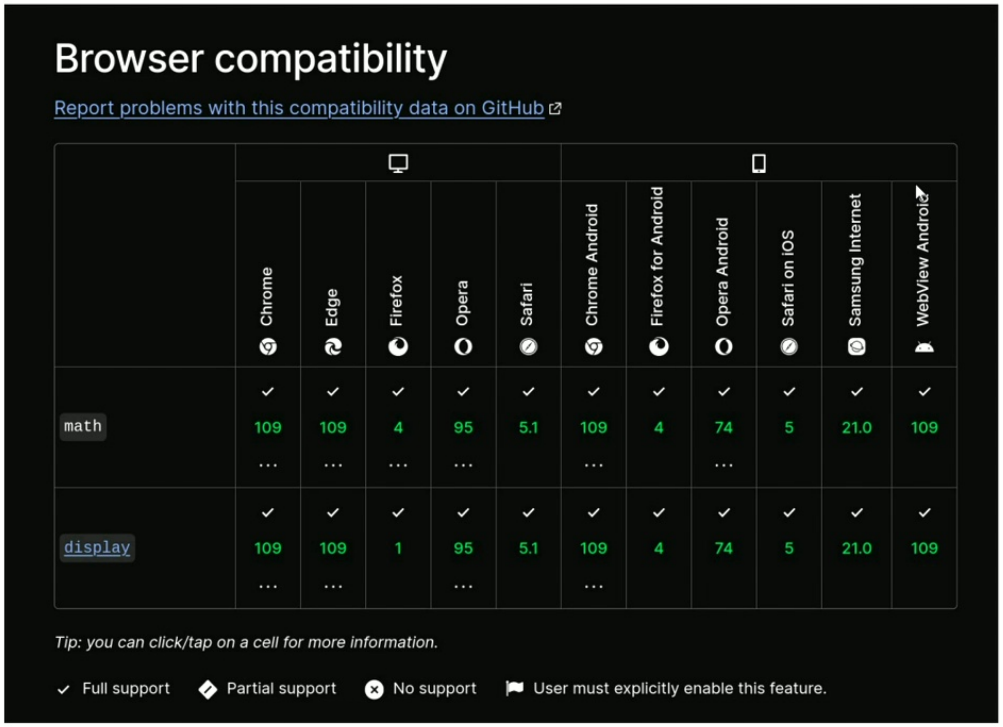
- arXiv authors write for their human readers
- If the PDF can be read, the job is done
- authors do **not** write to please markup developers
  - we all inevitably take shortcuts
  - the exception proves the rule
- So far, adoption favors messy systems

Instead, mapping into a minimal, standard, markup target improves:

- AI: pretrain on better signal-to-noise ratio
- search: index on better signal-to-noise ratio
- web: render a narrative tree (DOM) quickly and reliably. Native benefits.
- accessibility: able to speak the visual layout from the DOM, without fragility.
  - MathML 4 allows additional enhancements.

10 / 11

MathML Today



Browser compatibility

Report problems with this compatibility data on GitHub

	🖥					📱					
	Chrome	Edge	Firefox	Opera	Safari	Chrome Android	Firefox for Android	Opera Android	Safari on iOS	Samsung Internet	WebView Android
`math`	✓ 109 …	✓ 109 …	✓ 4 …	✓ 95 …	✓ 5.1	✓ 109 …	✓ 4	✓ 74 …	✓ 5	✓ 21.0	✓ 109
`display`	✓ 109 …	✓ 109 …	✓ 1	✓ 95 …	✓ 5.1	✓ 109 …	✓ 4	✓ 74	✓ 5	✓ 21.0	✓ 109

Tip: you can click/tap on a cell for more information.

✓ Full support ◆ Partial support ⊗ No support ⚑ User must explicitly enable this feature.

https://developer.mozilla.org/en-US/docs/Web/MathML#browser_compatibility