A Logic for Expressing Log-Precision Transformers


William Merrill


Ashish Sabharwal

New York University
New York, NY, USA

Allen Institute for AI
Seattle, WA, USA

Motivating questions

1. *What programming language do we need to express transformer computation?*
2. *What kinds of problems can transformers not solve?*

We show: the logic FO[M] can express any function a transformer classifier can compute

1. FO[M] = programming language for algorithms transformers can implement
2. Problems not expressible in FO[M] cannot be solved by transformers!

Log-precision transformers

- Chiang et al. (2023) give a logical upper bound on **finite-precision transformers**
- We show finite-precision transformers cannot express uniform attention, which is a powerful tool for transformers in practice!
- Instead, we study **log-precision transformers**, which can implement uniform attention

First-order logic (FO) over strings

- Logical sentences can be used to define sets of strings:

$$\exists i.\ \texttt{a}(i) \wedge \texttt{b}(i+1)$$

"Contains bigram `ab`"

- First order: can quantify over positions in string (like above)

Transformers can be expressed in FO[M]

Main Result: FO[M] Upper Bound
Any language recognized by a log-precision transformer can be defined in first-order logic with majority quantifiers (FO[M])

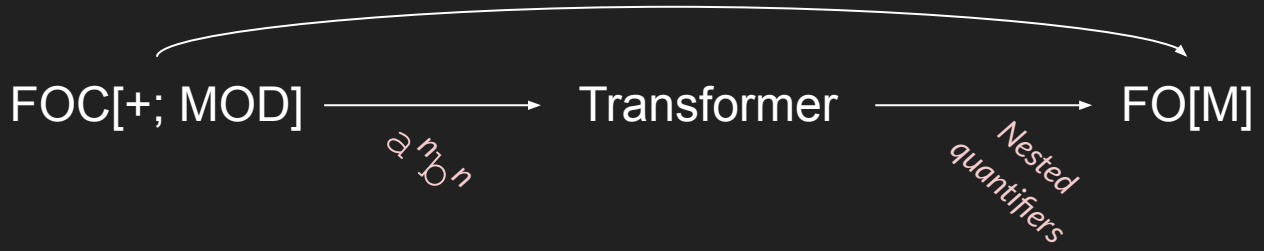
Example: defining $\mathtt{a}^n\mathtt{b}^n$ in FO[M]:

$$\mathsf{M}i.\ \mathtt{a}(i) \quad \wedge \quad \mathsf{M}j.\ \mathtt{b}(j) \quad \wedge \quad \neg \exists k.\ [\mathtt{b}(\mathsf{k}) \wedge \mathtt{a}(k+1)]$$

1. Most tokens are a
2. Most tokens are b
3. ba does not occur

Combining upper and lower bounds

Logical Lower Bound (*Chiang et al., 2023*)
Any language definable in counting logic with + and MOD (FOC[+; MOD]) can be recognized by a log-precision transformer



Open: tight (or tighter) logical characterization of transformers

Conclusion: transformers can be expressed in FO[M]

- **Mechanistic Interpretability:** FO[M] is a principled language to use to write programs extracted from transformers!

- **Limitations of Transformers:** Problems outside FO[M] like graph connectivity or matrix permanent cannot be solved by log-precision transformers
  a. Cf. “The Parallelism Tradeoff” (Merrill & Sabharwal, 2022)

- **Future Work:**
  a. Tighter logical characterization
  b. How does chain of thought change expressive power of transformers?