Modular Multimodal Machine Learning for Extraction of Theorems and Proofs in Long Scientific Documents

Shrey Mishra, Antoine Gauquier, Pierre Senellart

by Shrey Mishra, 17th December 2024

Motivation for TheoremKB

Volume of mathematical papers

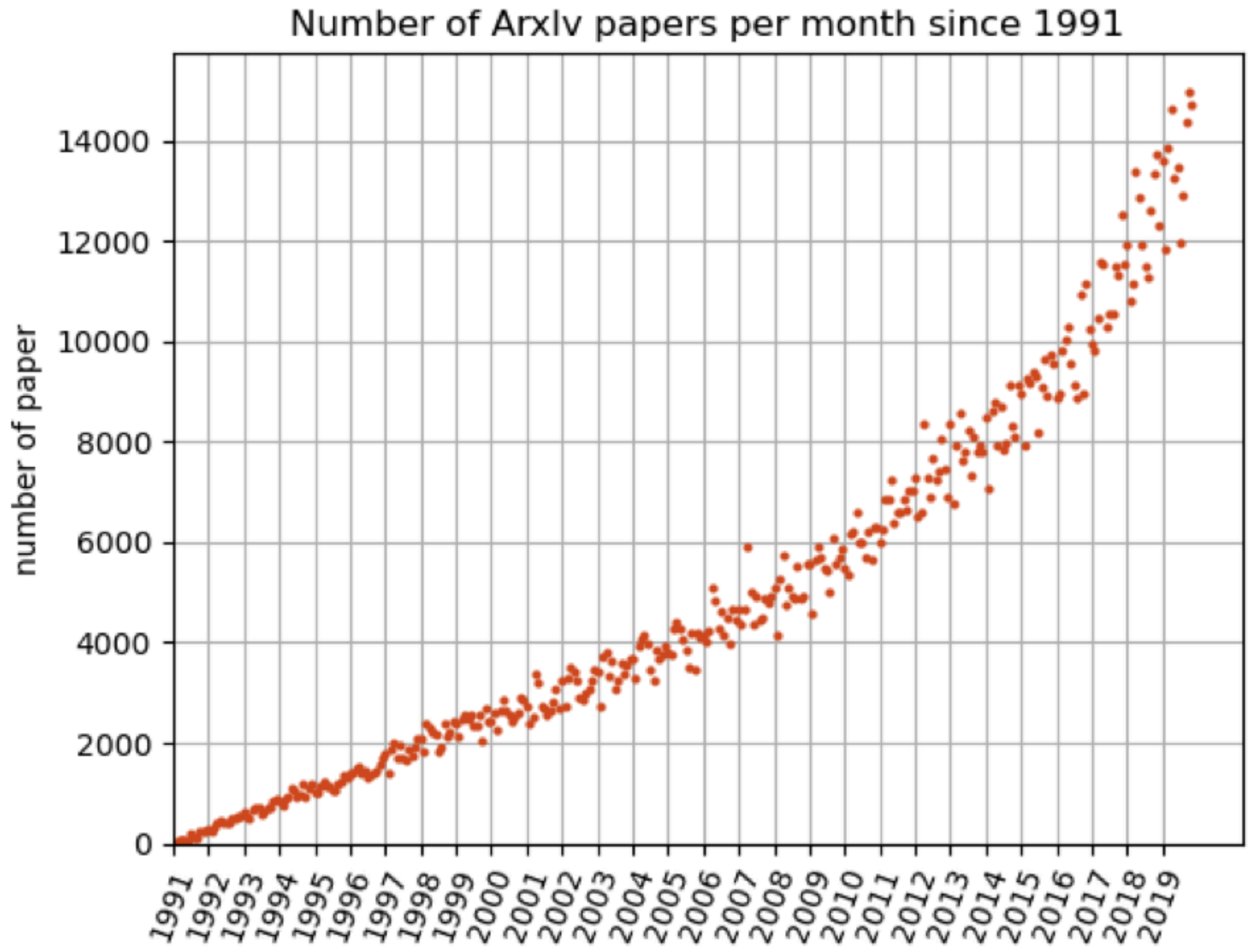


Figure 3.1: Total number of papers published on arXiv until 2019 [Del20]

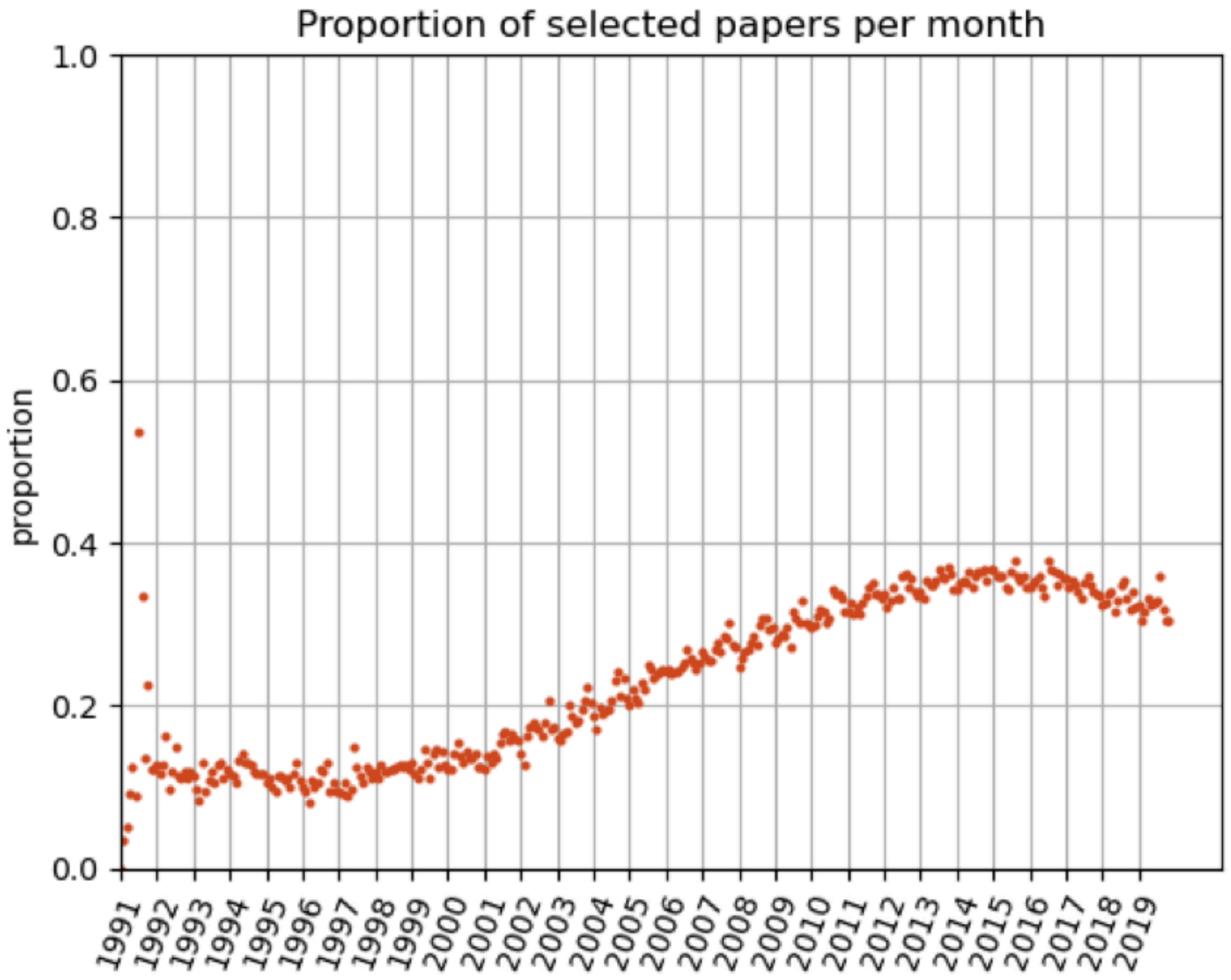
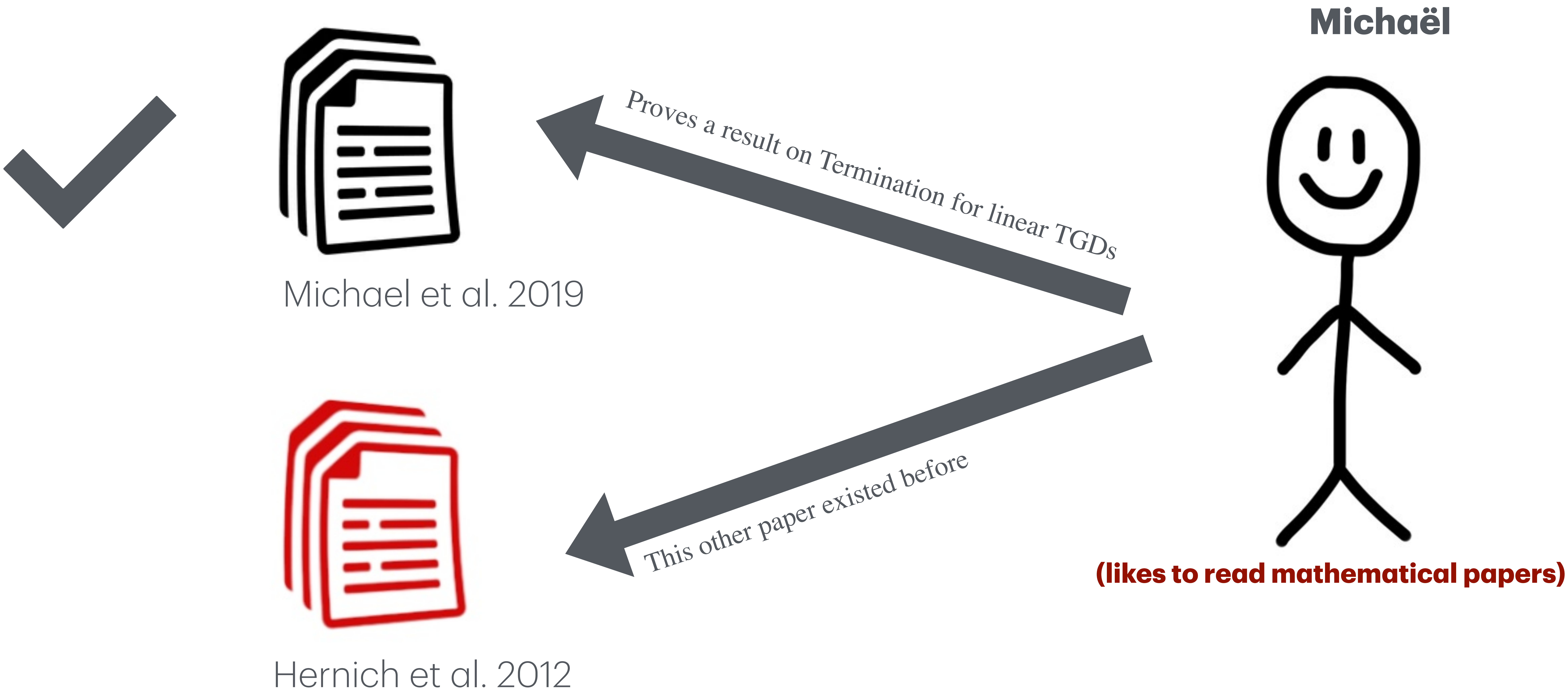


Figure 3.2: Papers with mathematical information such as Theorem, a Lemma or a Proposition [Del20]

Discovering existing knowledge



Correcting publicised errors

Proof: Dichotomy for evaluating conjunctive queries

Nilesh et al. 2006

Proof v2: Dichotomy for evaluating conjunctive queries

Nilesh et al. 2012

Managing Complex Dependencies

Senellart et al. Theory of Computing systems, 2019

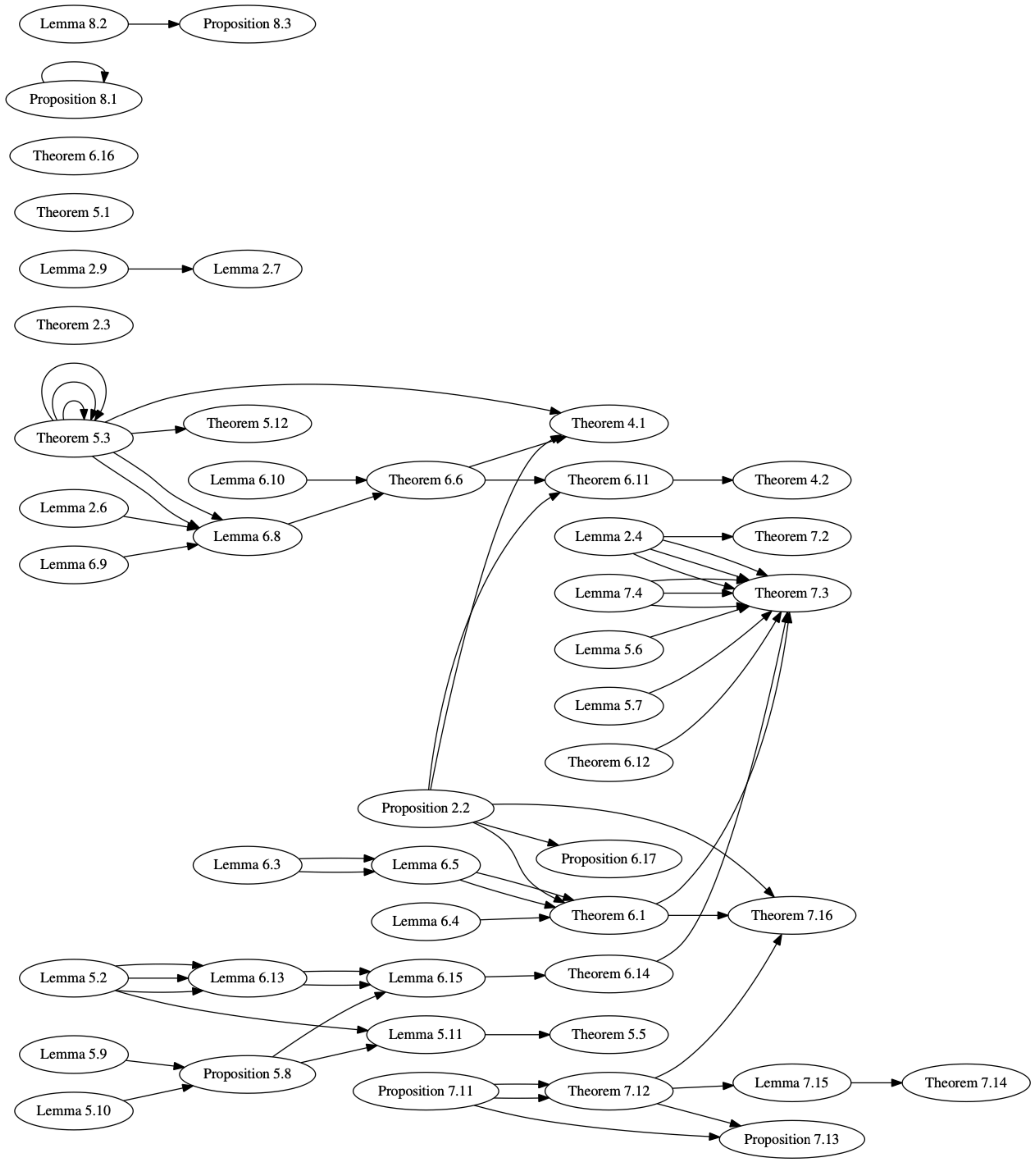
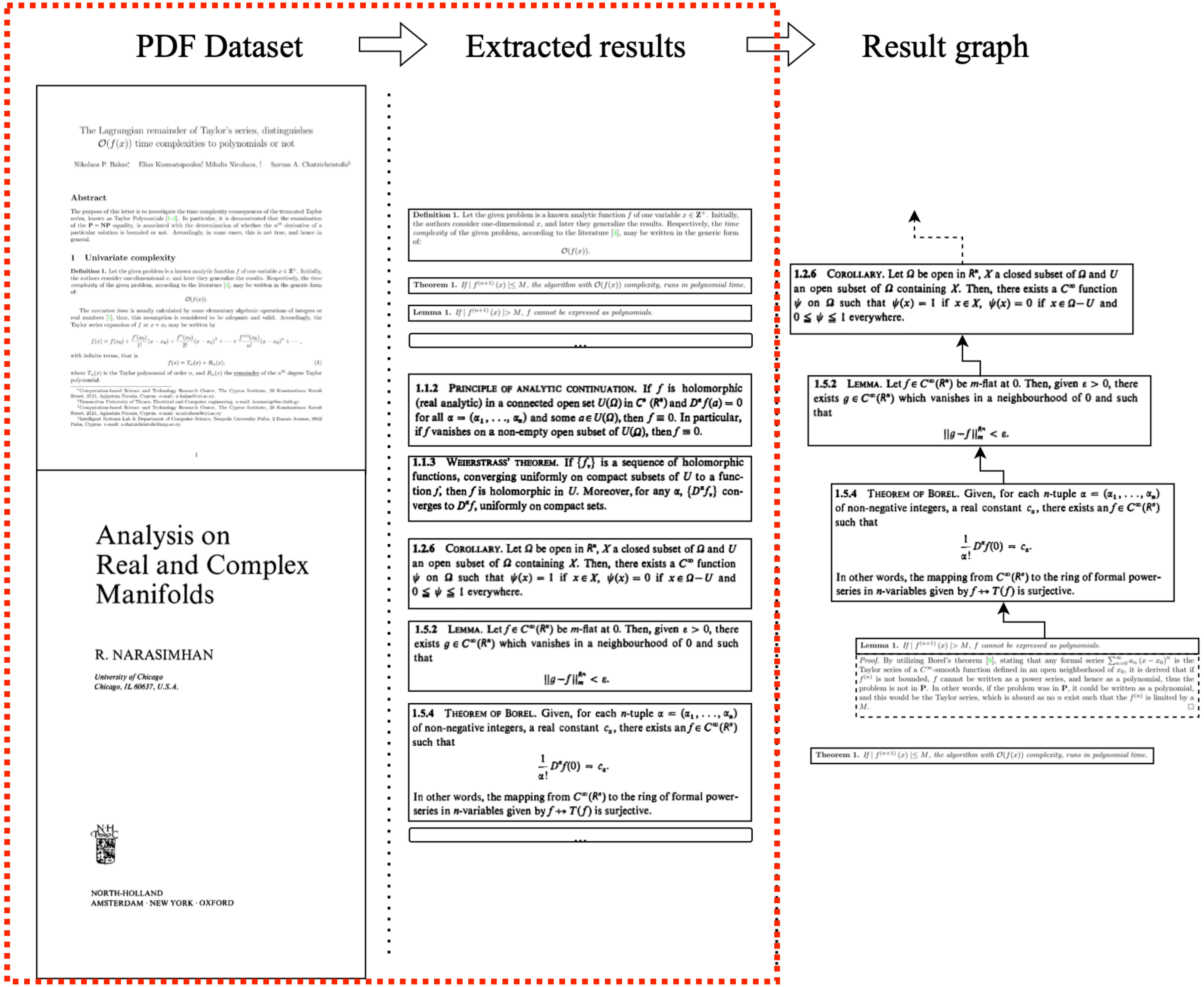


Table of contents

- *Motivation of theoremkb*
- Information Extraction task
- Unimodal and Multimodal backbones
- Sequential approach

Extraction task

High level overview of TheoremKB



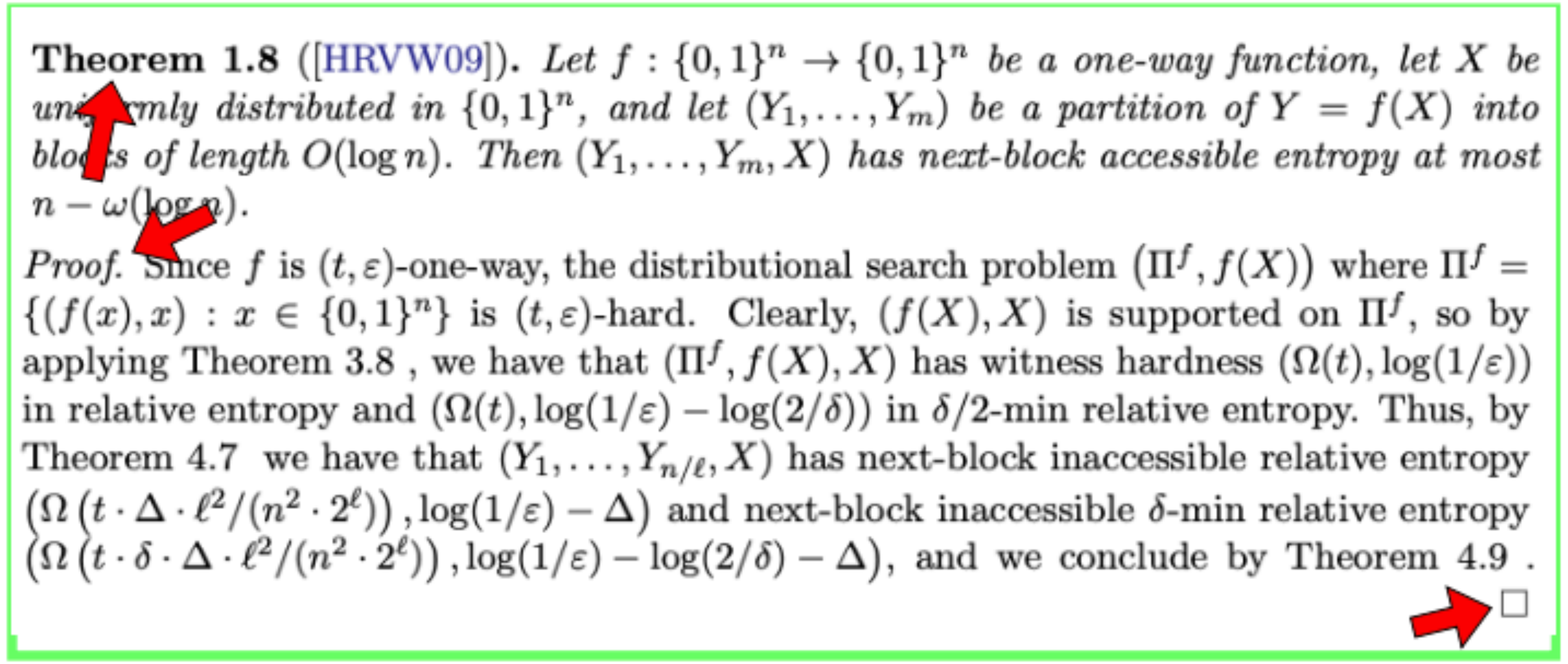
What is Extraction?

Theorem 1.8 ([HRVW09]). *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function, let X be uniformly distributed in $\{0,1\}^n$, and let $(Y_1, \ldots, Y_m)$ be a partition of $Y = f(X)$ into blocks of length $O(\log n)$. Then $(Y_1, \ldots, Y_m, X)$ has next-block accessible entropy at most $n - \omega(\log n)$.*

Proof. Since f is (t, ε)-one-way, the distributional search problem $(\Pi^f, f(X))$ where $\Pi^f = \{(f(x), x) : x \in \{0,1\}^n\}$ is (t, ε)-hard. Clearly, $(f(X), X)$ is supported on Π^f, so by applying Theorem 3.8 , we have that $(\Pi^f, f(X), X)$ has witness hardness $(\Omega(t), \log(1/\varepsilon))$ in relative entropy and $(\Omega(t), \log(1/\varepsilon) - \log(2/\delta))$ in $\delta/2$-min relative entropy. Thus, by Theorem 4.7 we have that $(Y_1, \ldots, Y_{n/\ell}, X)$ has next-block inaccessible relative entropy $\left(\Omega\left(t \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \log(1/\varepsilon) - \Delta\right)$ and next-block inaccessible δ-min relative entropy $\left(\Omega\left(t \cdot \delta \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \log(1/\varepsilon) - \log(2/\delta) - \Delta\right)$, and we conclude by Theorem 4.9 . □

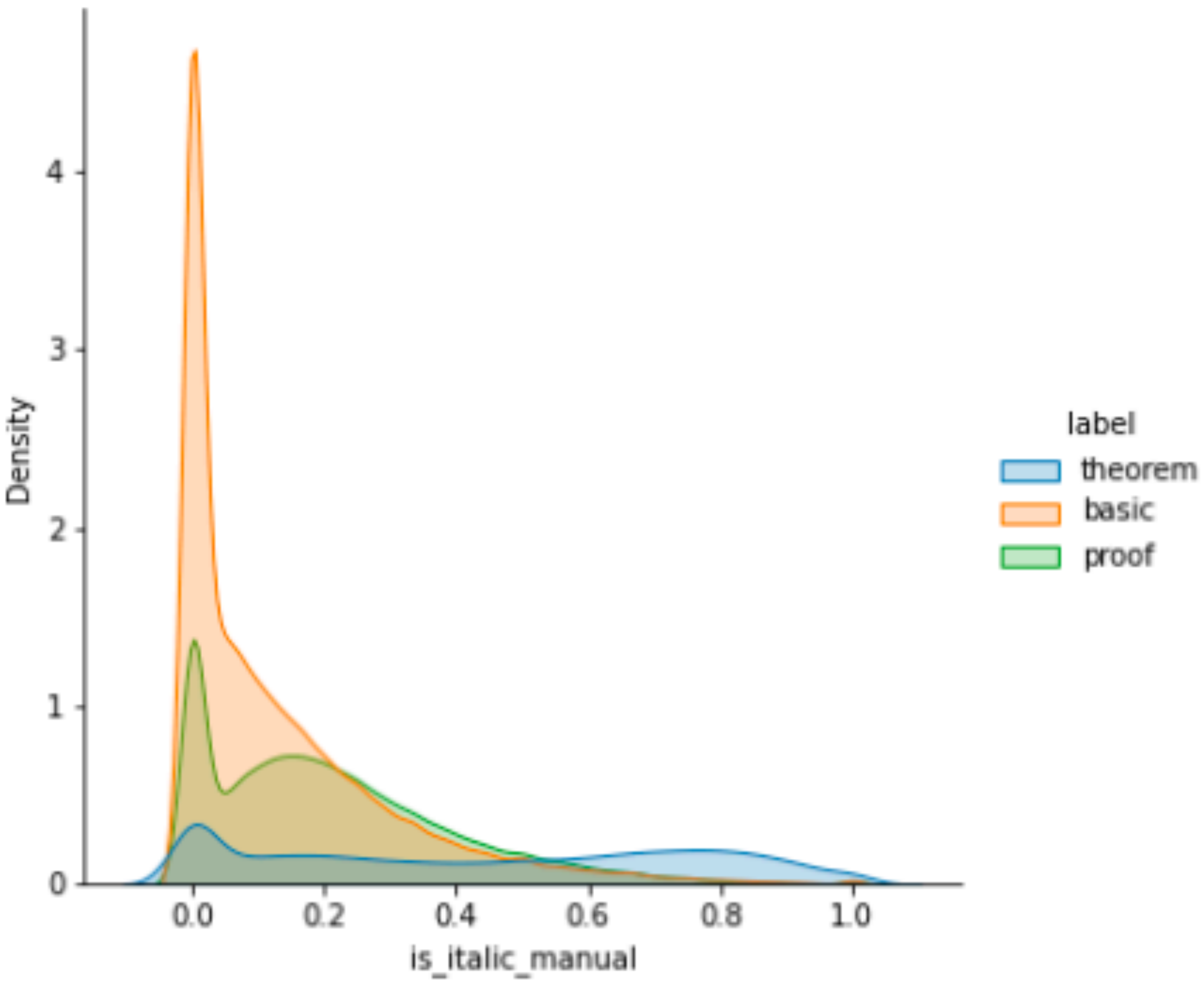
Unimodal backbones

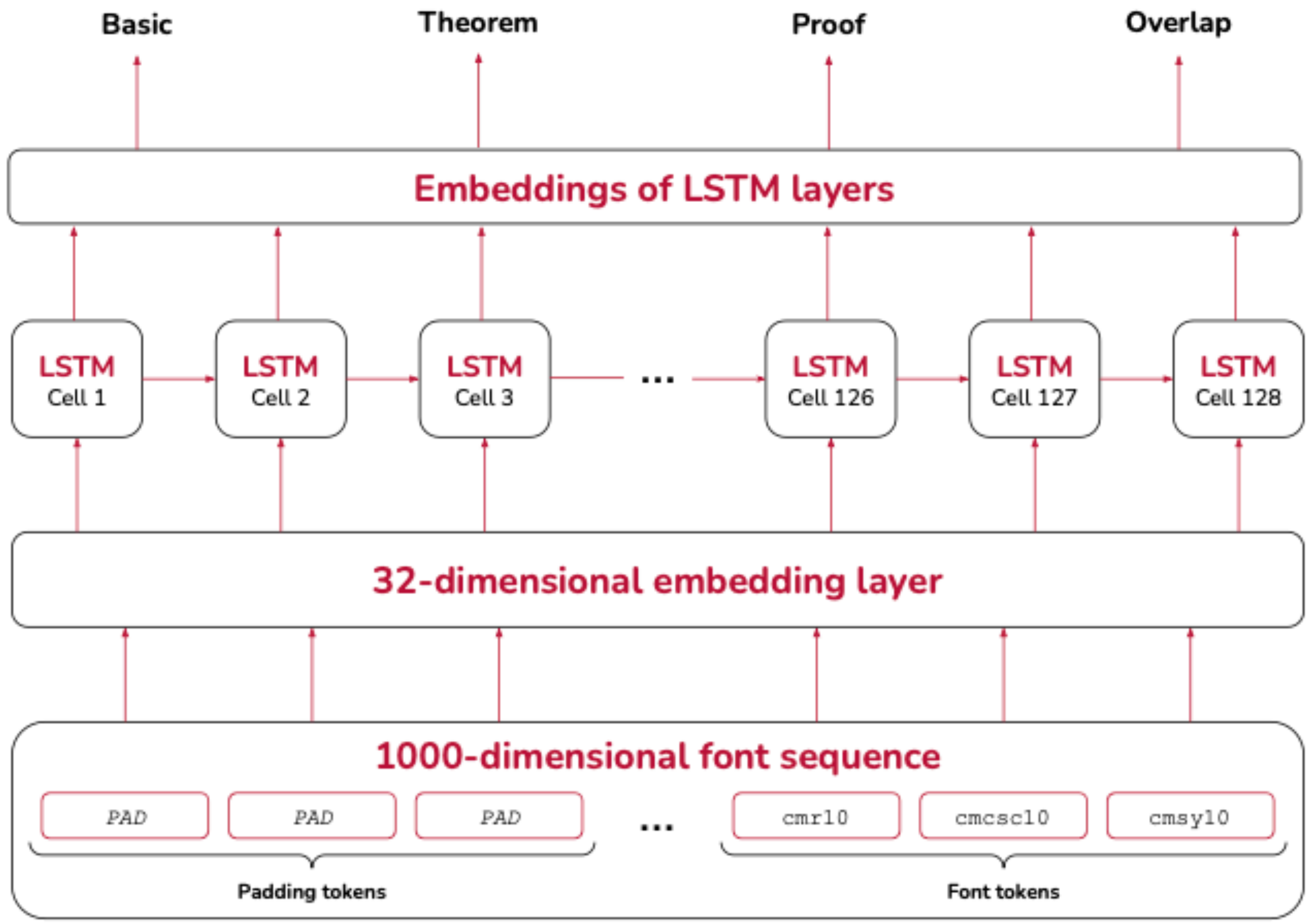
Font based



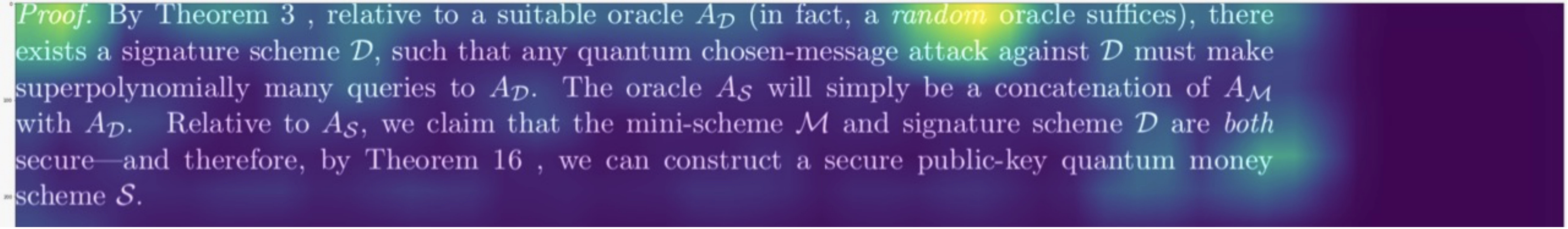
Theorem 1.8 ([HRVW09]). *Let $f : \{0,1\}^n \to \{0,1\}^n$ be a one-way function, let X be uniformly distributed in $\{0,1\}^n$, and let $(Y_1, \ldots, Y_m)$ be a partition of $Y = f(X)$ into blocks of length $O(\log n)$. Then $(Y_1, \ldots, Y_m, X)$ has next-block accessible entropy at most $n - \omega(\log n)$.*

Proof. Since f is (t, ε)-one-way, the distributional search problem $(\Pi^f, f(X))$ where $\Pi^f = \{(f(x), x) : x \in \{0,1\}^n\}$ is (t, ε)-hard. Clearly, $(f(X), X)$ is supported on Π^f, so by applying Theorem 3.8 , we have that $(\Pi^f, f(X), X)$ has witness hardness $(\Omega(t), \log(1/\varepsilon))$ in relative entropy and $(\Omega(t), \log(1/\varepsilon) - \log(2/\delta))$ in $\delta/2$-min relative entropy. Thus, by Theorem 4.7 we have that $(Y_1, \ldots, Y_{n/\ell}, X)$ has next-block inaccessible relative entropy $\left(\Omega\left(t \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \log(1/\varepsilon) - \Delta\right)$ and next-block inaccessible δ-min relative entropy $\left(\Omega\left(t \cdot \delta \cdot \Delta \cdot \ell^2/(n^2 \cdot 2^\ell)\right), \log(1/\varepsilon) - \log(2/\delta) - \Delta\right)$, and we conclude by Theorem 4.9 . □

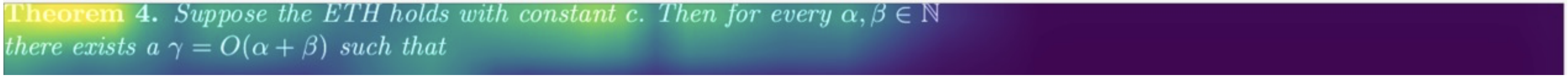




Vision Based

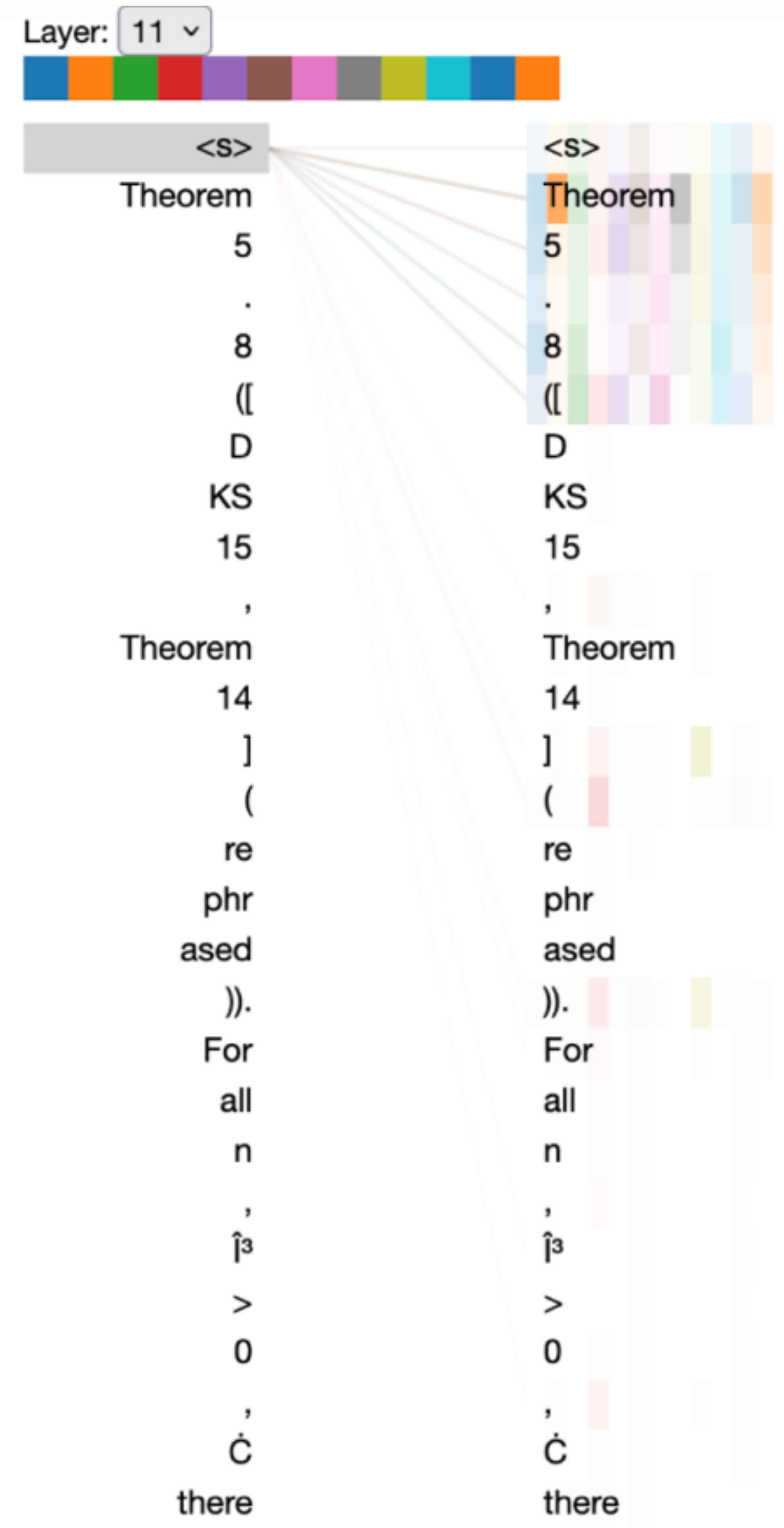


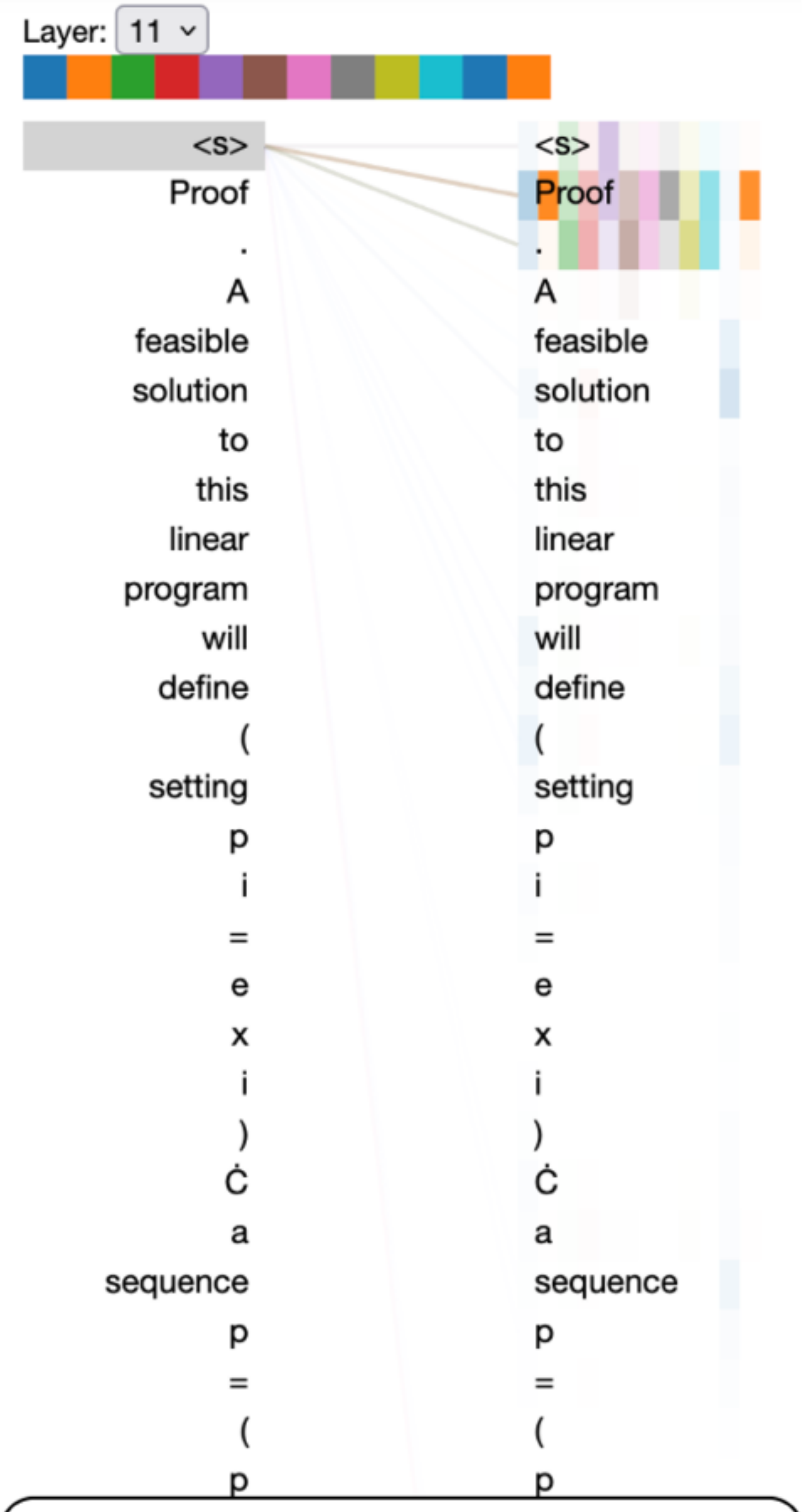
Proof. By Theorem 3 , relative to a suitable oracle $A_{\mathcal{D}}$ (in fact, a *random* oracle suffices), there exists a signature scheme $\mathcal{D}$, such that any quantum chosen-message attack against $\mathcal{D}$ must make superpolynomially many queries to $A_{\mathcal{D}}$. The oracle $A_{\mathcal{S}}$ will simply be a concatenation of $A_{\mathcal{M}}$ with $A_{\mathcal{D}}$. Relative to $A_{\mathcal{S}}$, we claim that the mini-scheme $\mathcal{M}$ and signature scheme $\mathcal{D}$ are *both* secure—and therefore, by Theorem 16 , we can construct a secure public-key quantum money scheme $\mathcal{S}$.



Theorem 4. *Suppose the ETH holds with constant c. Then for every $\alpha, \beta \in \mathbb{N}$ there exists a $\gamma = O(\alpha + \beta)$ such that*

Text based

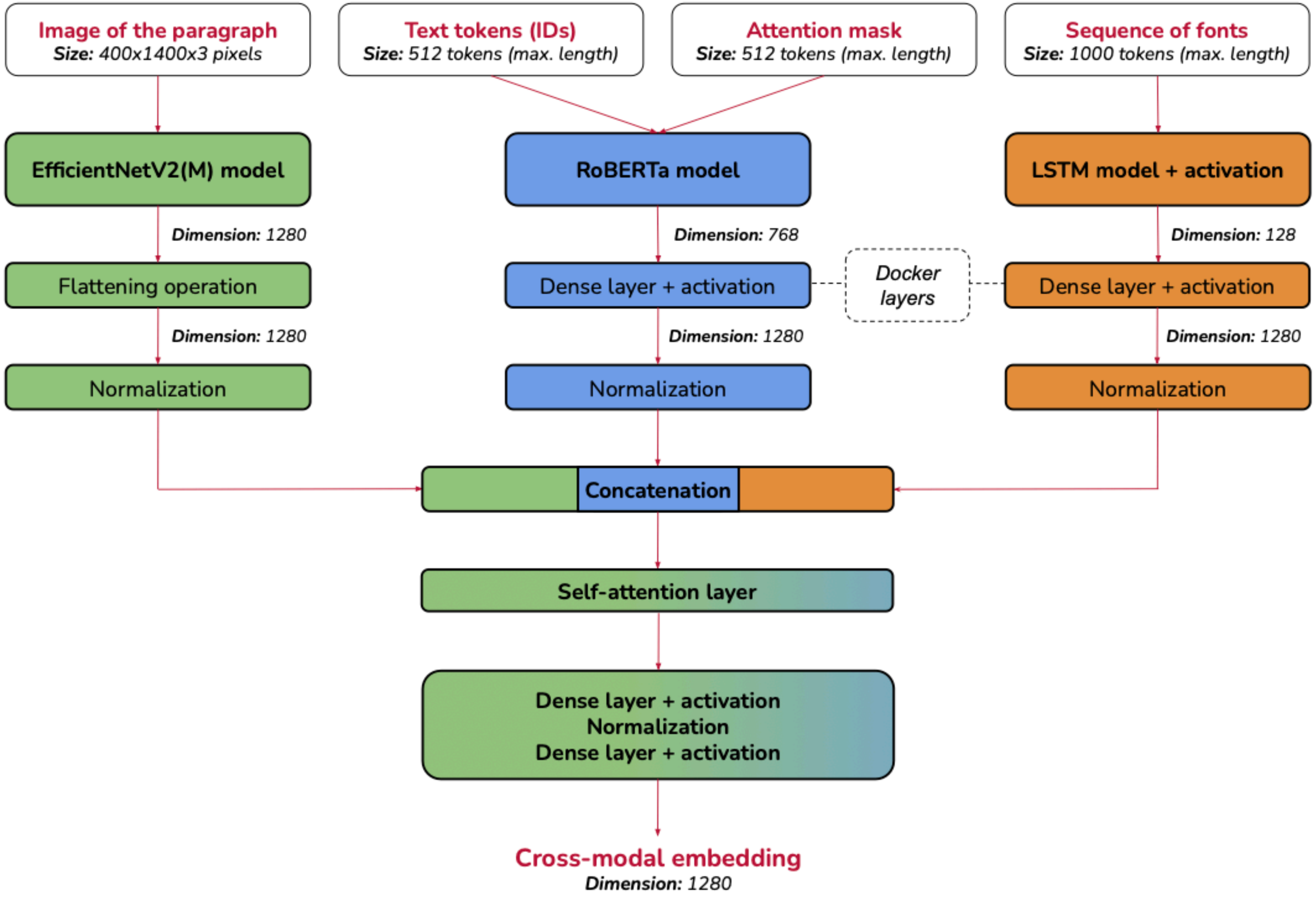




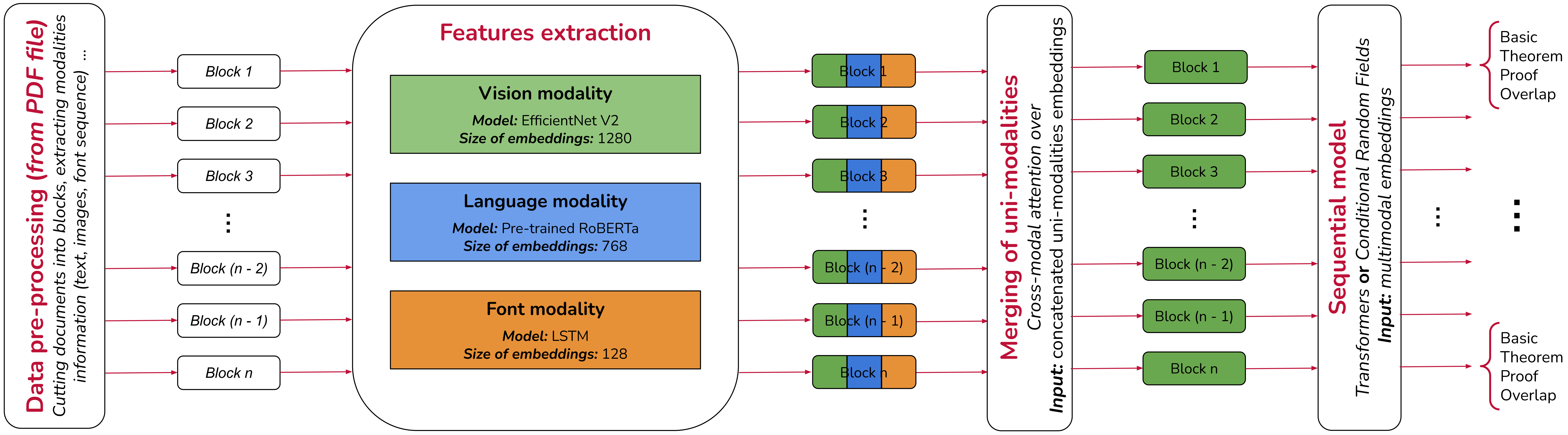
Theorem 5.8 ([DKS15, Theorem 14] (rephrased)). For all n, γ > 0,
there exists a set S γ ⊆ PBD n such that:

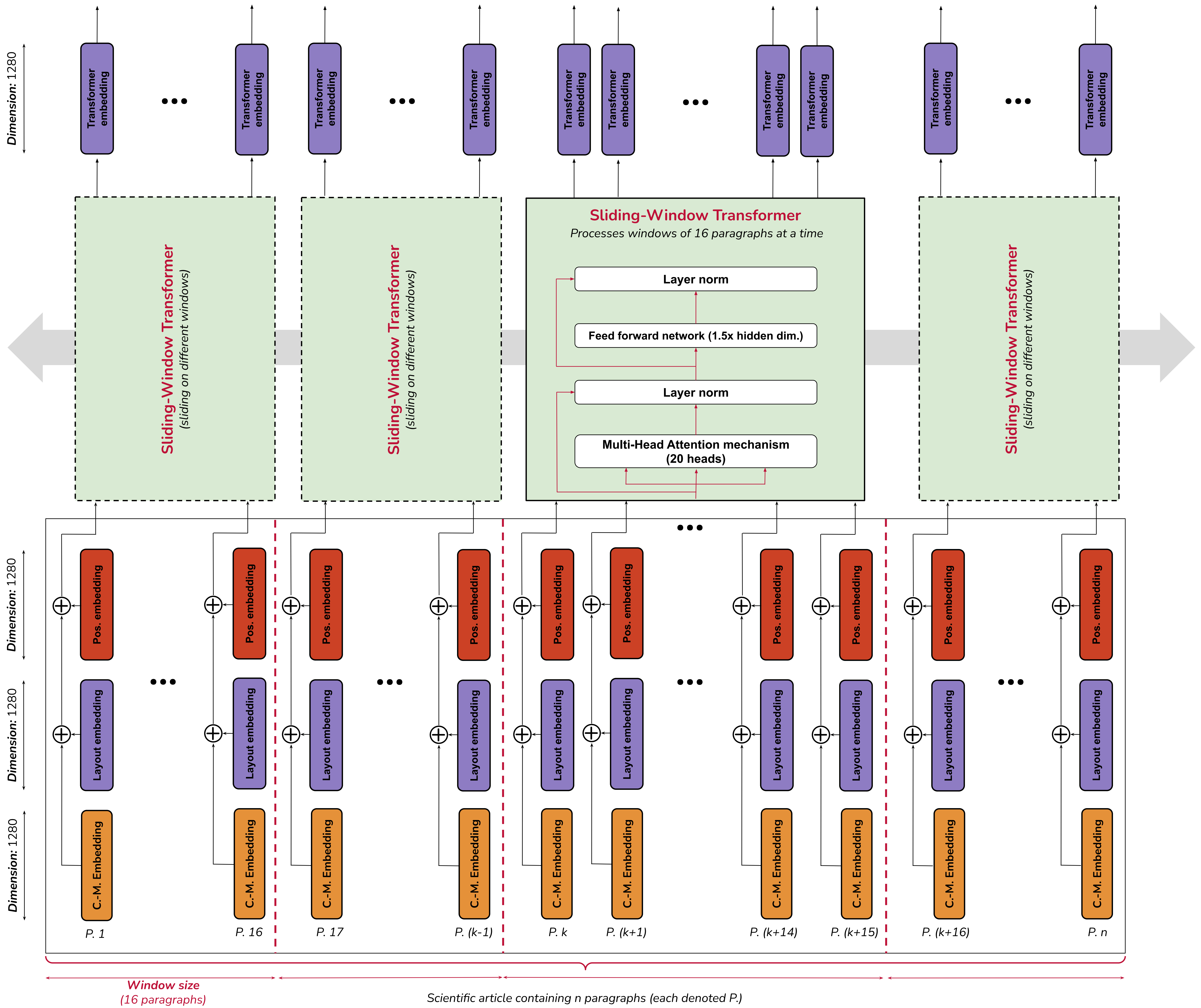
Proof. A feasible solution to this linear program will define (setting p i = e x i)
a sequence p = (p 1 , . . . , p n) ∈ (0, 1] n such
that

Multimodal model (raw features)



Sequential Approach





Dimension: 1280
Transformer embedding
Sliding-Window Transformer
(sliding on different windows)
Sliding-Window Transformer
Processes windows of 16 paragraphs at a time
Layer norm
Feed forward network (1.5x hidden dim.)
Layer norm
Multi-Head Attention mechanism (20 heads)
Pos. embedding
Layout embedding
C.-M. Embedding
P. 1
P. 16
P. 17
P. (k-1)
P. k
P. (k+1)
P. (k+14)
P. (k+15)
P. (k+16)
P. n
Window size
(16 paragraphs)
Scientific article containing n paragraphs (each denoted P.)

Performance of Sequence model

Modality	Model chosen	Seq. approach	#Batches	#Params (M)	Accuracy (%)	Mean F_1 (%)
Dummy	always predicts *basic*	—	—	—	59.41	24.85
Top-k first word	use only first word	—	—	—	52.84	44.20
Line-based [MPS21]	Bert (fine-tuned)	—	—	110	57.31	55.71

Closing remarks

- Our multimodal approach can be adapted to long documents
- Can make inference on entire pdf in a single forward pass
- Comparable, consistent and computationally efficient
- Unlike many other approaches that rely on an OCR preprocessing to be useful (LayoutLM) ours rely on Grobid which is many times faster
- Our approach is Fast (encoder only) and Scalable and applicable in real world (~200k pdfs tested)
- Our approach captures cross modality without adding special losses