UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Science of Language Models

Oliver

March 27, 2025

Motivations

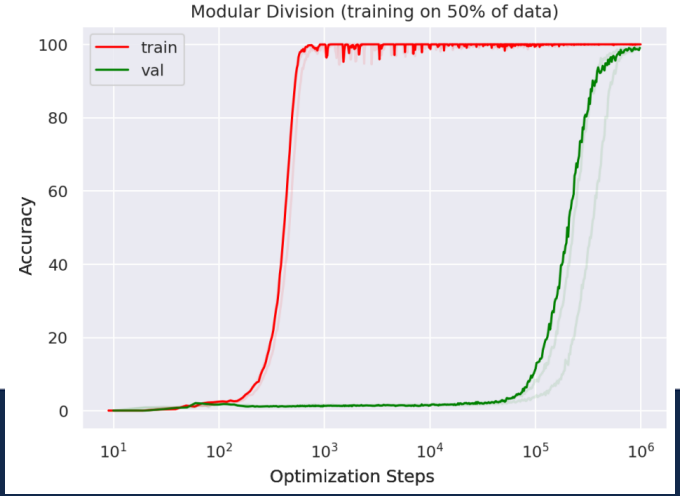
Deep Learning is a black box!

Motivations

Can we observe what's going on and try to understand it?

Interesting Observations in Deep Learning (e.g.)

- **Grokking**
  - a model initially achieves perfect training accuracy but no generalization (i.e. no better than a random predictor), and upon further training, transitions to almost perfect generalization.



Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. Technical Report, OpenAI

Interesting Observations in Deep Learning (e.g.)

- **Benign Overfitting**
  - a model can perfectly fit noisily labeled training data, but still archives near-optimal test error at the same time

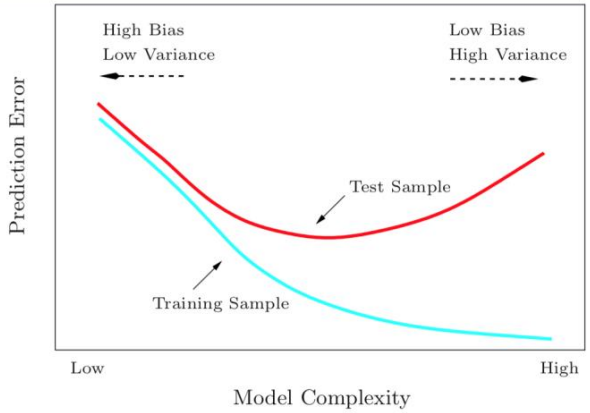
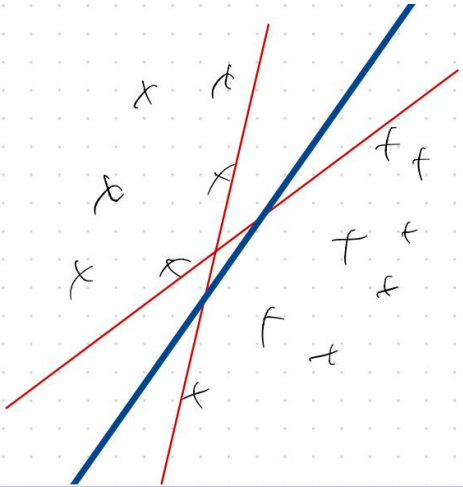


FIGURE 2.11. *Test and training error as a function of model complexity.*

Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

Interesting Observations in Deep Learning (e.g.)

- **Implicit Regularization of Gradient Descent**
  - Training by GD-based algorithms often attains the (best) global minimum.
  - GD implicitly constrains model capacity



$$L(w) = \sum_{i=1}^{n} \exp(-y_i \langle w, x_i \rangle).$$

Theorem. Let w^* be the hard-margin SVM solution, or the minimum ℓ_2-norm interpolator:

$$w^* = \arg\min_w \|w\|^2, \text{ s.t. } y_i \langle w, x_i \rangle \geq 1 \; \forall i \in [n].$$

Then, gradient flow over the exp loss converges to:

$$w(t) \to w^* \log t + O(\log \log t).$$

Moreover,

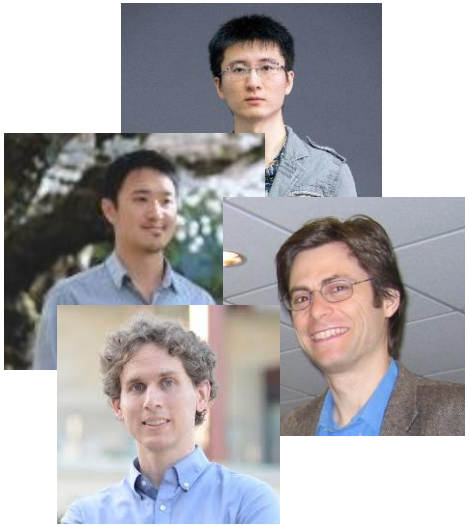
$$\frac{w(t)}{\|w(t)\|} \to \frac{w^*}{\|w^*\|}.$$

In other words, GF directionally converges towards the ℓ_2-max margin solution.

The Implicit Bias of Gradient Descent on Separable Data, JMLR'18

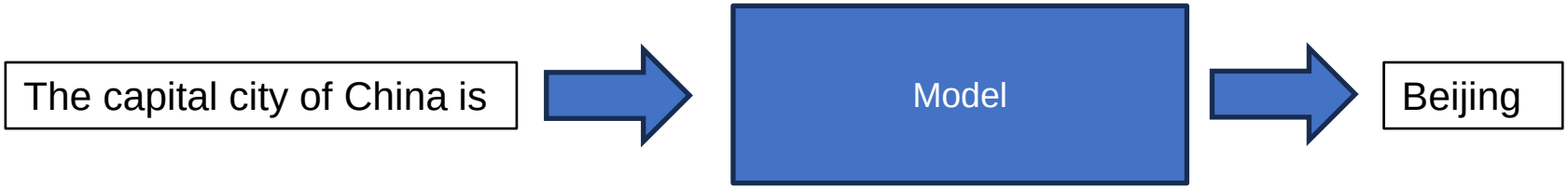
What about Large Language Models?

- **Almost impossible analyze their training dynamics rigorously**
  - Deep layers, complex data, strict hyperparamters



- Shallow Transformers
- Physics of LLM
- Mechanistic Interpretability

A unified topic: Memorization



Q: What can we analyze about such operations?

Approach One: Analyze Shallow Transformers

- **Setup: Associative View of Memory Recall**

Setup. Our setting follows that of Cabannes et al. [7]. Let $[N]$ be the set of input tokens, and $[M]$ be the set of output tokens. Our goal is to store a set of associations given by the function $f^* : [N] \to [M]$. For each input token $x \in [N]$ we assign a corresponding embedding vector $\boldsymbol{e}_x \in \mathbb{R}^d$, and likewise for each output token $y \in [M]$ we associate an unembedding vector $\boldsymbol{u}_y \in \mathbb{R}^d$. We primarily focus on the setting where the embeddings $\{\boldsymbol{e}_x\}_{x\in[N]}$ and $\{\boldsymbol{u}_y\}_{y\in[M]}$ are drawn i.i.d uniformly from the sphere of radius 1. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be our model which "stores" the associations f^*. Given such an F, the prediction $\hat{f}(x)$ for $f^*(x)$ is given by the arg-max decoding $\hat{f}(x) := \arg\max_{y\in[M]} \boldsymbol{u}_y^\top F(\boldsymbol{e}_x)$.

Approach One: Analyze Shallow Transformers

- **Synthetic Tasks:**
  - Vocabulary Space: $\mathcal{V} := \mathcal{S} \cup \mathcal{R} \cup \mathcal{A} \cup \mathcal{N} \cup \{\text{EOS}\}$ (Subject, Relation, Answer, Noise, End-of-Sequence)
  - (China, capital, Beijing) $\Rightarrow (s, r, a^*(s,r))$; a^* is one-to-one ($|a^*(s,r)| = 1$)
  - $\mathcal{A}_r := \{a^*(s,r) : s \in \mathcal{S}\}$ (The set of answers associated with relation r)

    $\mathcal{A}_s := \{a^*(s,r) : r \in \mathcal{R}\}$

Assumption 1. *$\mathcal{A}_r \cap \mathcal{A}_{r'} = \emptyset$ for $r, r' \in \mathcal{R}$ with $r \neq r'$. Furthermore, define $D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$.*

Interpretation: Different relations do not share the same answers.

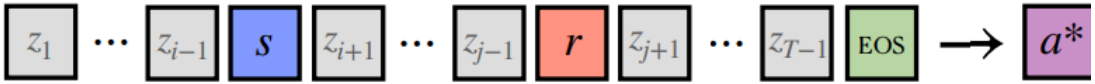


Figure 2: A diagram of the synthetic factual recall task.

Approach One: Analyze Shallow Transformers

- **Synthetic Tasks:** Generate sequences $z_{1:T+1} := (z_1, z_2, \ldots, z_T, z_{T+1}) \in \mathcal{V}^{T+1}$ as follows

1. First, sample a subject and relation tuple (s, r) from some distribution p over $\mathcal{S} \times \mathcal{R}$.
2. Next, sample two distinct indices $i, j \in [T-1]$. Set $z_i = s$ and $z_j = r$.
3. For the remainder of tokens z_k where $k \in [T-1] \setminus \{i, j\}$, draw z_k uniformly at random from the noise tokens $\mathcal{N}$.
4. Set $z_T = \text{EOS}$.
5. Finally, set $z_{T+1} = a^*(s, r)$.

Assumption 1. *$\mathcal{A}_r \cap \mathcal{A}_{r'} = \emptyset$ for $r, r' \in \mathcal{R}$ with $r \neq r'$. Furthermore, define $D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$.*

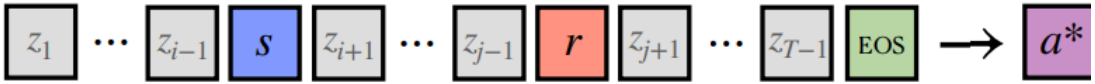


Figure 2: A diagram of the synthetic factual recall task.

Approach One: Analyze Shallow Transformers

- **Model Setup:** One-Layer Transformer
  - Input: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top \in \mathbb{R}^{T\times d}$, Weight matrices $W_K, W_Q, W_V, W_O \in \mathbb{R}^{d_h \times d}$
  - $attn: \mathbb{R}^{T\times d} \to \mathbb{R}^{d_h}$ $\quad F_{MHSA}: \mathbb{R}^{T\times d} \to \mathbb{R}^{d}$ $\quad F_{TF}: \mathbb{R}^{T\times d} \to \mathbb{R}^{d}$
  - $V, W \in \mathbb{R}^{m\times d}$

$$\mathrm{attn}(\boldsymbol{X}; \boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V) = \boldsymbol{W}_V \boldsymbol{X}^\top \mathcal{S}\left(\boldsymbol{X}\boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{x}_T\right),$$

$$F_{\mathrm{MHSA}}(\boldsymbol{X}; \boldsymbol{\theta}) = \sum_{h\in[H]} \boldsymbol{W}_O^{(h)\top} \mathrm{attn}(\boldsymbol{X}; \boldsymbol{W}_K^{(h)}, \boldsymbol{W}_Q^{(h)}, \boldsymbol{W}_V^{(h)}).$$

$$F_{\mathrm{TF}}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{TF}}) = F_{\mathrm{MHSA}}(\boldsymbol{X}; \boldsymbol{\theta}) + \boldsymbol{V}^\top \sigma(\boldsymbol{W} F_{\mathrm{MHSA}}(\boldsymbol{X}; \boldsymbol{\theta})).$$

Parametrized by (d, H, d_h, m), the model has $4Hdd_h$ self-attn parameters and $2md$ MLP parameters

Approach One: Analyze Shallow Transformers

$S = |\mathcal{S}|, R = |\mathcal{R}|$
$D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$

- **Analysis I: Expressive Studies**

Theorem 3 (Attention-only, informal). *Assume that* $d \geq \tilde{\Omega}(\max(R, D))$ *and* $Hd_h \geq \tilde{\Omega}(S + R)$. *With high probability over the embeddings, there exists a single-layer attention-only transformer* $F_{\mathrm{TF}}(\,\cdot\,; \boldsymbol{\theta}_{\mathrm{TF}})$ *with embedding dimension* d, *number of heads* H *and head dimension* d_h *such that*

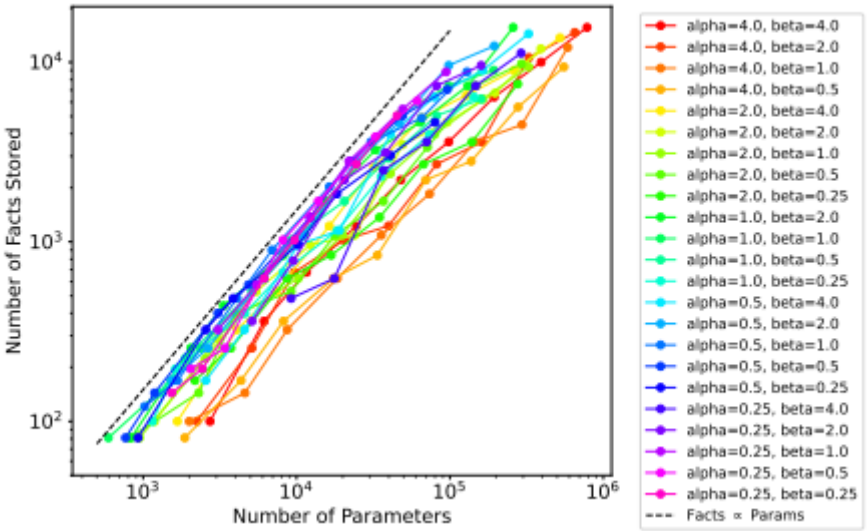
$$\mathbb{P}_{z_{1:T+1} \sim \mathcal{D}} \left[\arg\max_{z \in \mathcal{V}} \boldsymbol{\varphi}(z)^\top F_{\mathrm{TF}}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{TF}}) = z_{T+1} \right] = 1. \tag{8}$$

Interpretation: perfect solver requires the total number of parameters $4Hdd_h$ to scale (up to polylog factors) linearly with the dataset size SR

Approach One: Analyze Shallow Transformers

$S = |\mathcal{S}|, R = |\mathcal{R}|$
$D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$

- **Analysis I: Expressive Studies**



at $d \geq \tilde{\Omega}(\max(R, D))$ *and* $Hd_h \geq \tilde{\Omega}(S + R)$.
exists a single-layer attention-only transformer
er of heads H *and head dimension* d_h *such that*

$$\left[\text{TF}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{TF}}) = z_{T+1} \right] = 1. \quad (8)$$

Experiment: # facts stored scale linearly with # parameters

Approach One: Analyze Shallow Transformers

$S = |\mathcal{S}|, R = |\mathcal{R}|$
$D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$

- **Analysis I: Expressive Studies**

Theorem 4 (Attention + MLP, informal)**.** *Assume that σ is a polynomial of sufficiently large degree. Define $C(a) = |\{(s, r) : a^*(s, r) = a\}|$. Let (d, H, d_h, m) satisfy*

$$d \geq \tilde{\Omega}(1) \qquad Hd_h \geq \tilde{\Omega}(S + R) \qquad m \geq \tilde{\Omega}(\max_a C(a)) \qquad md \geq \tilde{\Omega}(SR). \tag{9}$$

Then with high probability over the embeddings there exists a single-layer transformer $F_{\mathrm{TF}}(\,\cdot\,; \boldsymbol{\theta}_{\mathrm{TF}})$ with embedding dimension d, number of heads H, head dimension d_h, and MLP width m such that

$$\mathbb{P}_{z_{1:T+1} \sim \mathcal{D}} \left[\arg\max_{z \in \mathcal{V}} \boldsymbol{\varphi}(z)^\top F_{\mathrm{TF}}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{TF}}) = z_{T+1} \right] = 1. \tag{10}$$

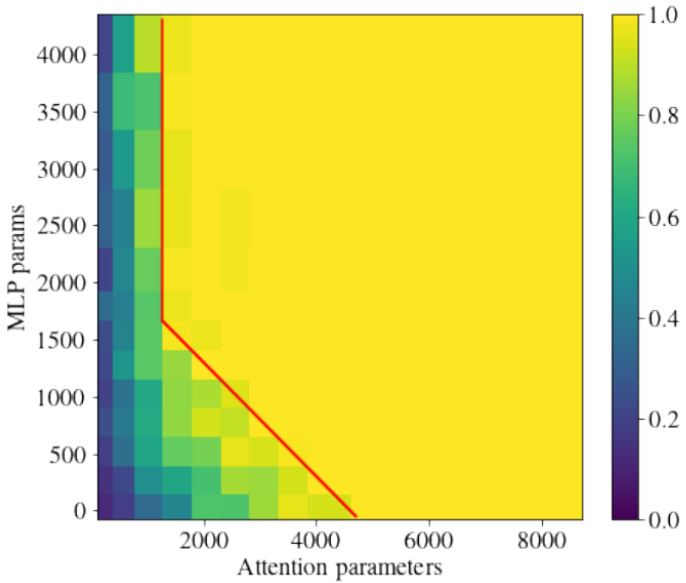
The proofs of Theorem 3 and Theorem 4 are deferred to Appendix C.

Interpretation: TF can tradeoff to store the knowledge inside the MLP layer

Approach One: Analyze Shallow Transformers

$S = |\mathcal{S}|, R = |\mathcal{R}|$
$D := \max_{r \in \mathcal{R}} |\mathcal{A}_r|$

- **Analysis I: Expressive Studies**



Experiment: One can tradeoff parameters between MLP and Attn

Assume that σ is a polynomial of sufficiently large de-
$a\}|$. Let (d, H, d_h, m) satisfy

$R)$ $\quad m \geq \tilde{\Omega}(\max_a C(a)) \quad md \geq \tilde{\Omega}(SR).$ (9)

ddings there exists a single-layer transformer $F_{\mathrm{TF}}(\cdot\,; \boldsymbol{\theta}_{\mathrm{TF}})$
heads H, head

$\varphi(z)^\top F_{\mathrm{TF}}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{TF}}) = z_{T+1}\Big] = 1.$ (10)

are deferred to Appendix C.

to store the knowledge inside the MLP layer

Approach One: Analyze Shallow Transformers

- **Analysis II: Optimization Dynamics**

More simplifications (orthogonal embeddings, linear attention-only, symmetric)

Model: $$F_{\text{lin}}(\boldsymbol{X};\boldsymbol{\theta}) := \boldsymbol{W}_{OV}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}_{KQ}\boldsymbol{x}_T$$

Next-token prediction: $$\hat{p}(a \mid z_{1:T}) := \frac{\exp\left(\langle \boldsymbol{\varphi}(a), F_{\text{lin}}(\boldsymbol{X};\boldsymbol{\theta})\rangle\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\langle \boldsymbol{\varphi}(a'), F_{\text{lin}}(\boldsymbol{X};\boldsymbol{\theta})\rangle\right)}.$$

CE Loss: $$L(\boldsymbol{\theta}) = \mathbb{E}_{z_{1:T+1}}[-\log \hat{p}(z_{T+1} \mid z_{1:T})].$$

Initialization: **Assumption 2.** *Given an initialization scale $\alpha > 0$, set $\boldsymbol{W}_{OV}(a,z) = \alpha$ and $\boldsymbol{W}_{KQ}(z) = \alpha\sqrt{|\mathcal{A}|+1}$ for each $a \in \mathcal{A}, z \in \mathcal{V}$.*

Also assume total number of parameters $2d^2 \gg SR$ (overparametrized regime)

Approach One: Analyze Shallow Transformers

- **Analysis II: Optimization Dynamics**
  - Result 1 (no surprises): GD can attain a global min

Theorem 5 (Global Convergence). *For $t \geq 0$, let $\boldsymbol{\theta}(t)$ be the output of running gradient flow for t time. For any $\delta > 0$, there exists a time t_δ such that for $t \geq t_\delta$, $L(\boldsymbol{\theta}(t)) \leq \delta$.*

Approach One: Analyze Shallow Transformers

- **Analysis II: Optimization Dynamics**
  - Result 2: Sequential Learning

Theorem 6 (Sequential Learning). *Assume that* $S \geq 8R\sqrt{2D}$, *and* $|\mathcal{N}| \geq 4R\sqrt{2DT}$. *Let* $p(s,r) = \frac{1}{SR}$. *Pick* $\epsilon > 0$. *There exists runtime* T^* *and initialization scale* α *(both depending on* ϵ*) such that:*

1. *For all* $t \leq T^*$ *and* $z \in \mathcal{S} \cup \mathcal{N}, a \in \mathcal{A}$, *we have* $|\boldsymbol{W}_{OV}(a,z)|, |\boldsymbol{W}_{KQ}(z)| \leq \alpha^{1/2}$
2. *There exists* $t \leq T^*$ *such that, for any input sequence* $z_{1:T}$ *containing a relation* r,

$$\sum_{a \in \mathcal{A}} \left(p^*(a \mid r) - \hat{p}(a \mid z_{1:T})\right)^2 \leq \epsilon^2. \tag{16}$$

Interpretation: At some time, the model only predicts based on the conditional distribution of r. E.g. What is the capital of France? A: random country capital

Approach One: Analyze Shallow Transformers

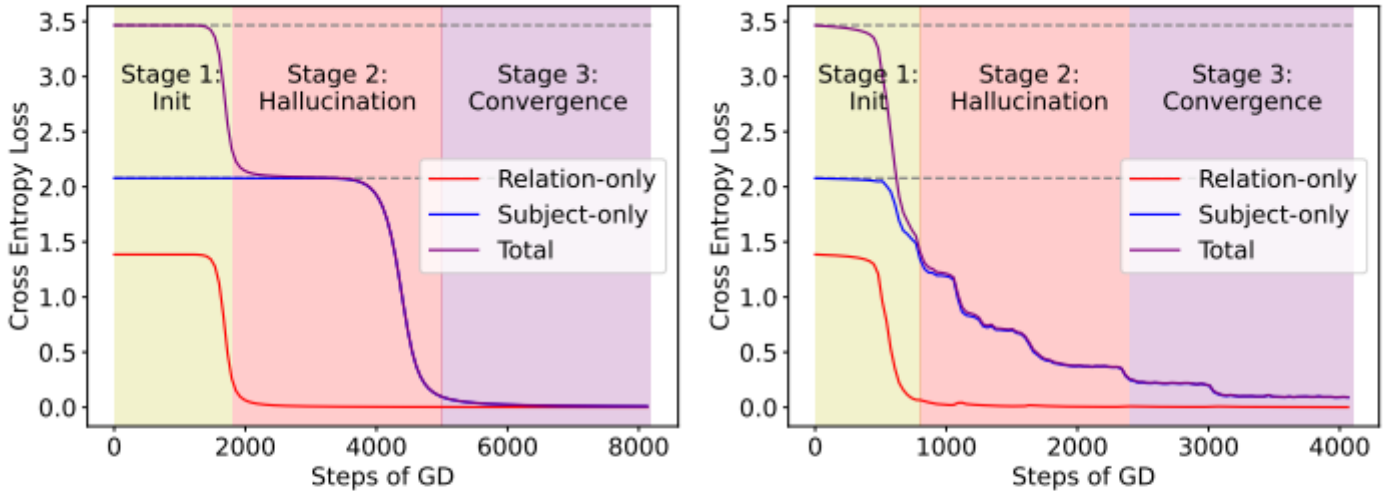
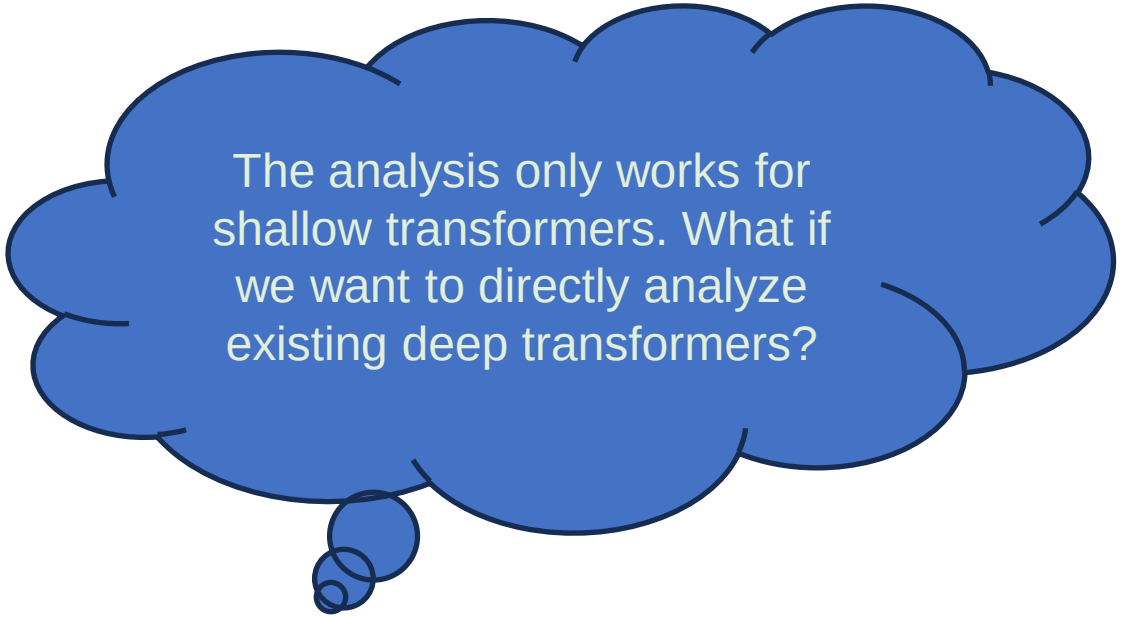


Figure 5: (Left) Loss of the linear attention model with orthogonal embeddings. There is an intermediate *hallucination* stage where the loss plateaus and the model predicts based on only the relation. (Right) Loss of the softmax attention model with random embeddings. We again observe an intermediate hallucination stage, where the relation-only loss is zero but the total loss is still large.

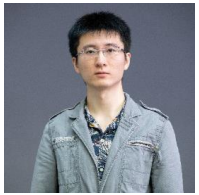
$\mathbb{E}_{z_{1:T+1}}\left[-\log\left(\sum_{a\in\mathcal{A}_r} p(a \mid z_{1:T})\right)\right]$

Relation-only loss (above)

Beyond Shallow Transformers



Approach II: Physics of LLM



: Design controlled, reproducible, and rigorous experiments

Jane Austen

Inverse search: "In <Pride and Prejudice>, what's the sentence ***before***: <sentence2>?"

Forward search: "In <Pride and Prejudice>, what's the sentence ***after***: <sentence1>?"

	Pride & Prejudice	Sense & Sensibility	Persuasion	Northanger Abbey	Emma	Mansfield Park
GPT3.5	0.5% vs 14.4%	0.3% vs 5.4%	0.07% vs 4.3%	0.6% vs 5.5%	0.8% vs 7.2%	0.7% vs 5.5%
GPT4	0.8% vs 65.9%	0.9% vs 40.2%	0.5% vs 33.9%	0.9% vs 41.0%	0.6% vs 42.7%	0.3% vs 31.7%

Example: Reversal Curse

knowledge inverse search

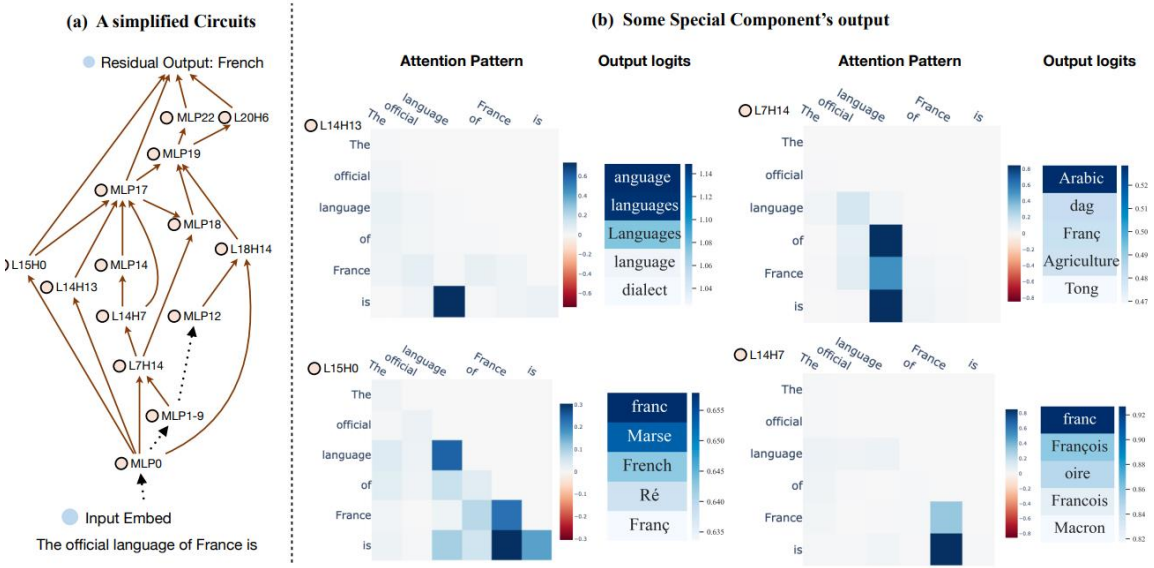
1. Give me the [first/full] name of the person born on October 2, 1996? (bdate to first, bdate to full)
2. Give me the [first/full] name of the person born on October 2, 1996 in Princeton, NJ? (birth to first, birth to full)
3. Give me the [first/full] name of the person who studied Communications at MIT and worked for Meta Platforms? (three to first, three to full)
4. Give me the [first/full] name of the person who studied Communications at MIT, was born in Princeton, NJ, and worked for Meta Platforms? (four to first, four to full)
5. Give me the [first/full] name of the person who studied Communications at MIT, was born on October 2, 1996 in Princeton, NJ, and worked for Meta Platforms at Menlo Park, CA? (all to first, all to full)

bioS data

Anya Briar Forger was born on October 2, 1996. She spent her early years in Princeton, NJ. She received mentorship and guidance from faculty members at MIT. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.

Approach III: Mechanistic Interpretability

- **Method: Representation Engineering, Probing, Sparse Autoencoder….**
  - Domain: Interpretability, Safety, Model Auditing



Thank you!
Q & A