Time Course MechInterp: Analyzing the Evolution of Components and Knowledge in Large Language Models

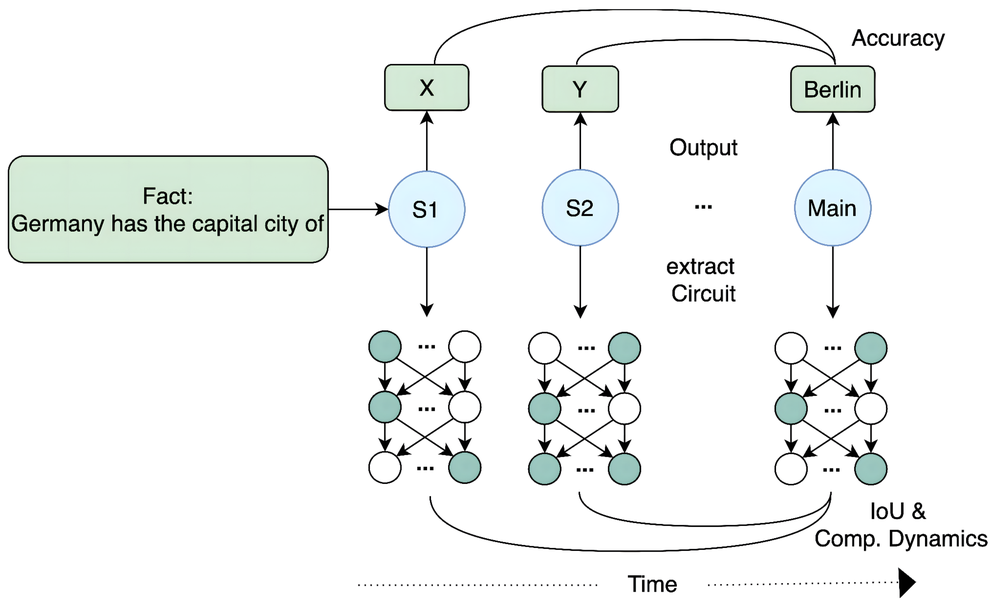
Ahmad Dawar Hakimi, Ali Modarressi, Philipp Wicke, Hinrich Schütze

Munich Center for Machine Learning

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ACL 2025 VIENNA

How does factual knowledge emerge during LLM pretraining?



- LLMs encode factual knowledge, but the *learning process* remains opaque.
- Mechanistic Interpretability can help uncover which *model components* (attention heads & FFNs) drive factual knowledge recall.
- This study traces the *evolution* of these components over 40 snapshots of OLMo-7B.

MAIN Findings:

names more complex für mich nicht intuitiv…

1. **Task Complexity Influences Training Dynamics**: Simple facts (e.g. locations) converge early in pre-training, while more complex relationships (e.g., names) only emerge after sustained training.
2. **Hierarchical Learning Process:** The model initially leverages broad, general-purpose attention heads and FFNs before progressively spawning specialized submodules dedicated to narrower fact types.
3. **Adaptive vs. Stable Components:** A subset of attention heads dynamically repurposes throughout training to capture new information, whereas certain FFNs form a stable backbone that supports factual recall.
4. **Evolving Specialization:** both attention heads and FFNs increasingly refine their roles, becoming more reliably tuned to specific categories of knowledge as training advances.

Dataset Construction

Germany | has the capital city | of | Berlin

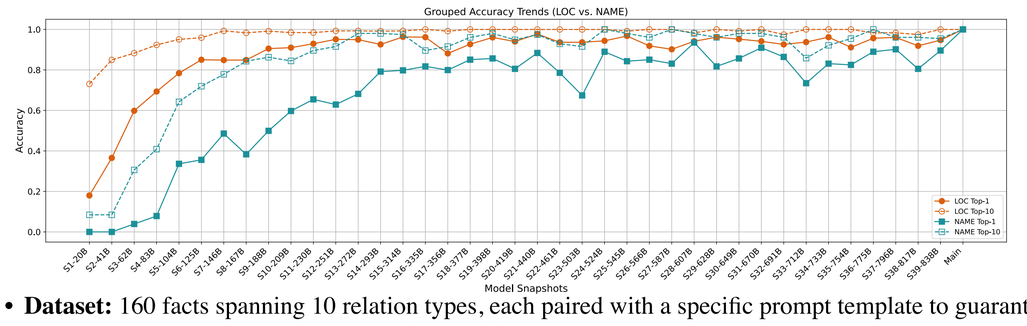
SUBJECT — RELATION — END — ANSWER

Location-based Relations (LOC)

Relation	Prompt Template	# Facts	Example Subject
CITY_IN_COUNTRY	{} is part of the country of	14	Rio de Janeiro, Buenos Aires
COMPANY_HQ	The headquarters of {} are in the city of	20	Zillow, Bayrischer Rundfunk
COUNTRY_CAPITAL_CITY	{} has the capital city of	19	Canada, Nigeria
FOOD_FROM_COUNTRY	{} is from the country of	17	Sushi, Ceviche
OFFICIAL_LANGUAGE	In {}, the official language is	14	France, Egypt
PLAYS_SPORT	{} plays professionally in the sport of	12	Kobe Bryant, Roger Federer
SIGHTS_IN_CITY	{} is a landmark in the city of	17	The Eiffel Tower, The Space Needle

Name-based Relations (NAME)

Relation	Prompt Template	# Facts	Example Subject
BOOKS_WRITTEN	The Book {} was written by the author with the name of	13	The Hunger Games, Life of Pi
COMPANY_CEO	Who is the CEO of {}? Their name is	17	Ubisoft, Pinterest
MOVIE_DIRECTED	The Movie {} was directed by the director with the name of	17	The Godfather, Forrest Gump



- **Dataset:** 160 facts spanning 10 relation types, each paired with a specific prompt template to guarantee correct, unambiguous completions by the model.
- **Accuracy Trends:** Location-based facts reach near-perfect performance within the first few checkpoints, while name-based facts improve steadily across all 40 snapshots.

margin zu knapp!

Model Component Roles

Role Score: | Hierarchical Proper Role:

General: $c_s^g = \frac{\sum_{r\in R}\sum_{f\in r} c_{srf}(T_g)}{\sum_{r\in R}\sum_{f\in r} 1}$ | $\mathcal{H}_g = \mathcal{J}_g$

Entity: $c_s^e = \frac{\sum_{r\in R}\sum_{f\in r} c_{srf}(T_e)}{\sum_{r\in R}\sum_{f\in r} 1}$ | $\mathcal{H}_e = \mathcal{J}_e - \mathcal{J}_g$

Relation-Answer: $c_s^r = \frac{\sum_{f\in r} c_{srf}(T_a)}{\sum_{f\in r} 1}$ | $\mathcal{H}_r = \mathcal{J}_r - \mathcal{J}_e - \mathcal{J}_g$

Fact-Answer Specific: $c_s^f = c_{srf}^f(T_a)$ | $\mathcal{H}_f = \mathcal{J}_f - \mathcal{J}_r - \mathcal{J}_e - \mathcal{J}_g$

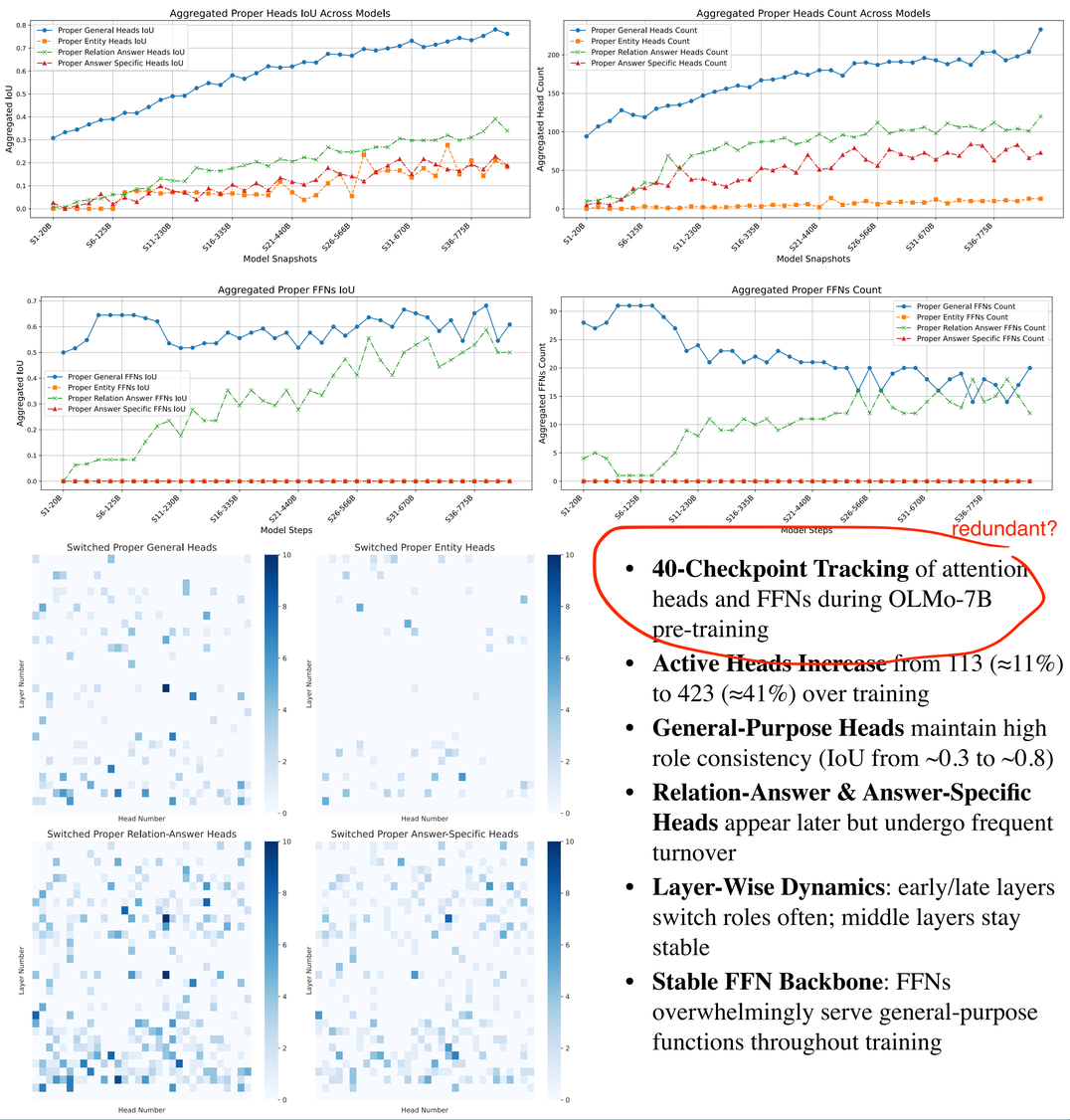
Deactivated Components: $\mathcal{H}_d = \mathcal{C}(\mathcal{H}_g \cup \mathcal{H}_e \cup \mathcal{H}_r \cup \mathcal{H}_f)$

- Extract per-subtoken circuits using Information Flow Routes (Ferrando & Voita, 2024)
- We compute activation scores $c_s^g, c_s^e, c_s^r, c_s^f$ for each component at each snapshot, using subtoken sets T_g, T_e, T_a, and a threshold $\theta = 0.1$.
- Then by successive differencing of cumulative importance sets $J_g \to J_e \to J_r \to J_f$, we obtain non-overlapping proper role sets H_g, H_e, H_r and H_f, with H_d capturing all remaining deactivated components.

How do Components Evolve?

Temporal Consistency and Role Dynamics of Components

$$\text{IoU}(\mathcal{H}_g) = \frac{\mathcal{H}_{gs} \cap \mathcal{H}_{gmain}}{\mathcal{H}_{gs} \cup \mathcal{H}_{gmain}}$$

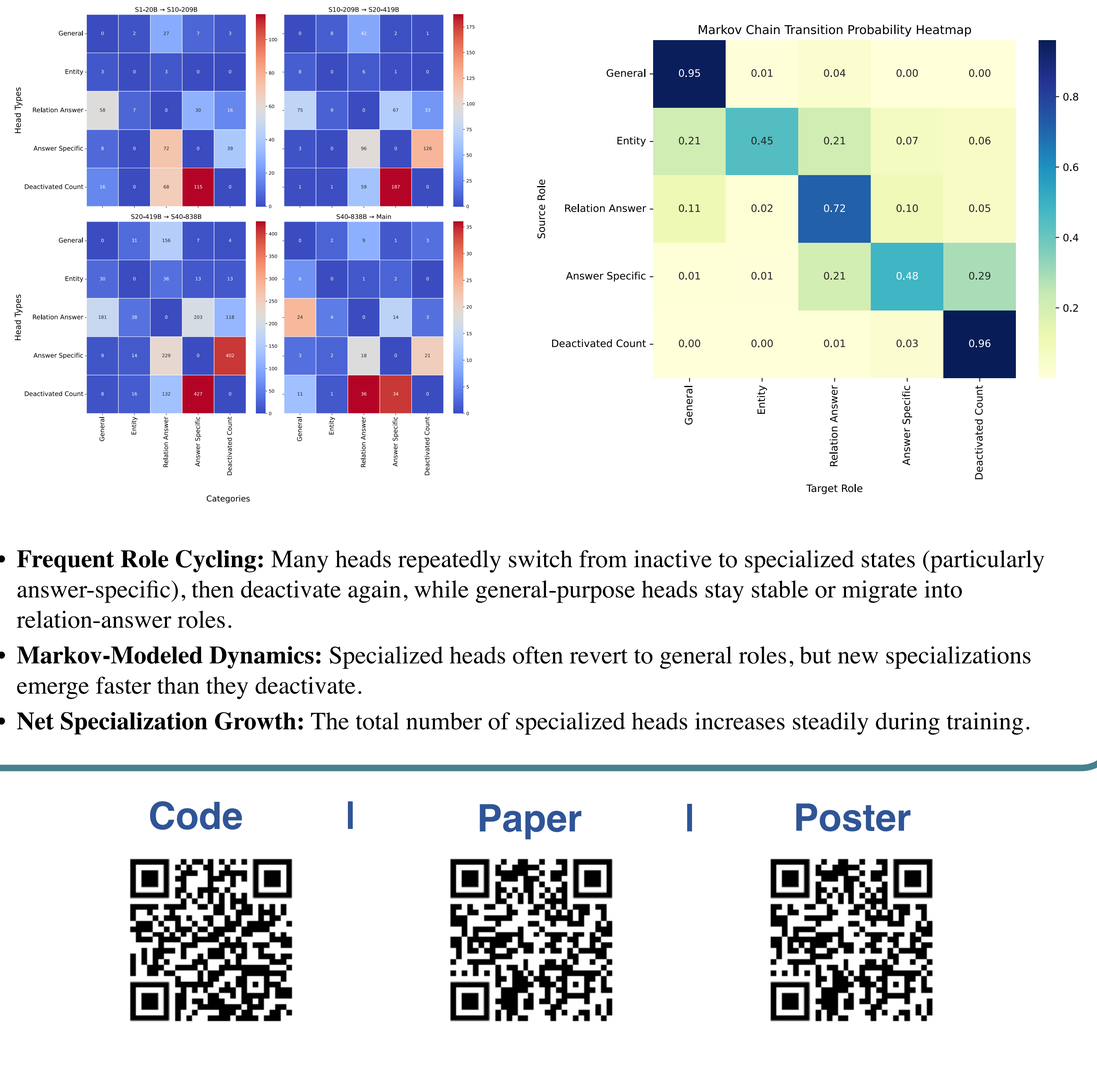


redundant?

- **40-Checkpoint Tracking** of attention heads and FFNs during OLMo-7B pre-training
- **Active Heads Increase** from 113 (≈11%) to 423 (≈41%) over training
- **General-Purpose Heads** maintain high role consistency (IoU from ~0.3 to ~0.8)
- **Relation-Answer & Answer-Specific Heads** appear later but undergo frequent turnover
- **Layer-Wise Dynamics**: early/late layers switch roles often; middle layers stay stable
- **Stable FFN Backbone**: FFNs overwhelmingly serve general-purpose functions throughout training

Dynamic Specialization and Generalization of Attention Heads

$$P(\mathcal{H}_\alpha \to \mathcal{H}_\beta) = \frac{N(\mathcal{H}_\alpha \to \mathcal{H}_\beta)}{\sum_{\gamma\in\{g,e,r,f,d\}} N(\mathcal{H}_\alpha \to \mathcal{H}_\gamma)}$$



- **Frequent Role Cycling:** Many heads repeatedly switch from inactive to specialized states (particularly answer-specific), then deactivate again, while general-purpose heads stay stable or migrate into relation-answer roles.
- **Markov-Modeled Dynamics:** Specialized heads often revert to general roles, but new specializations emerge faster than they deactivate.
- **Net Specialization Growth:** The total number of specialized heads increases steadily during training.

Code | **Paper** | **Poster**