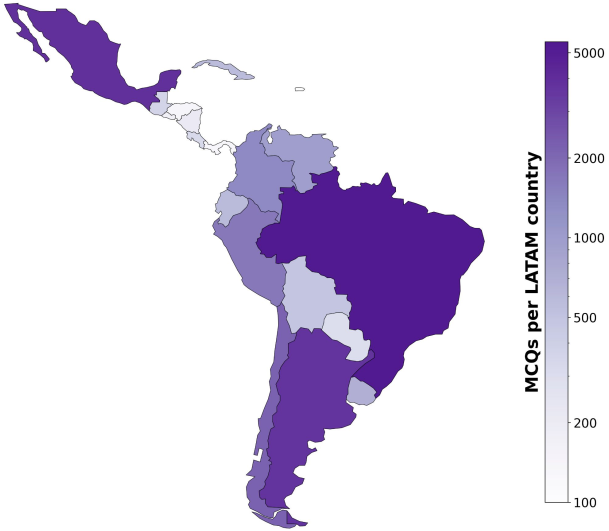
Understand memorisation and knowledge acquisition in LLM

Yannis Karmim

06/05/2026

Why this talk ?



- Work between Inria chile and U. de Chile on the construction of a dataset to evaluate LLM knowledge on LATAM Culture

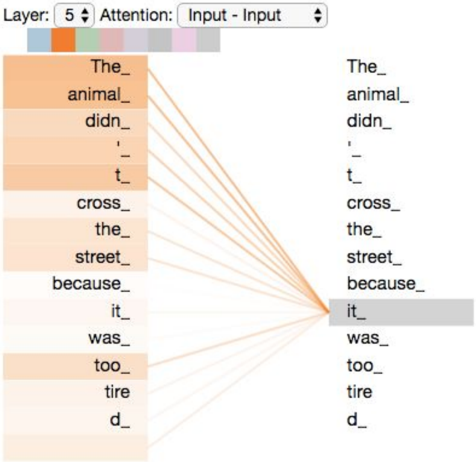
- Lot of questions during this work: Why is LLM unable to extract some knowledge even if it was in the training data ?

- How LLM acquire knowledge ?

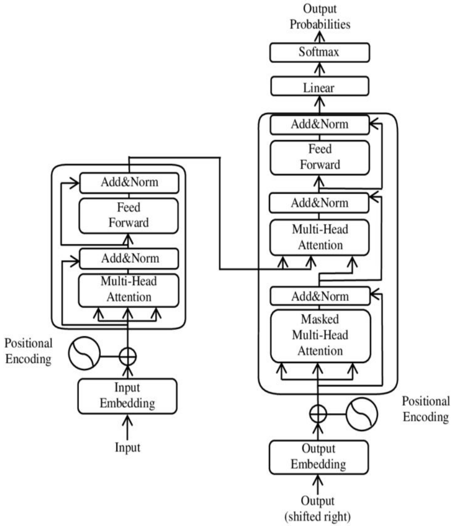
📕 Y. Karmim; R. Pino; H. Contreras; H. Lira; S. Cifuentes; S. Escoffier; L. Marti; D. Seddah; V. Barriere

LatamMCQ: Leveraging wikidata for geographically informed sociocultural dataset creation

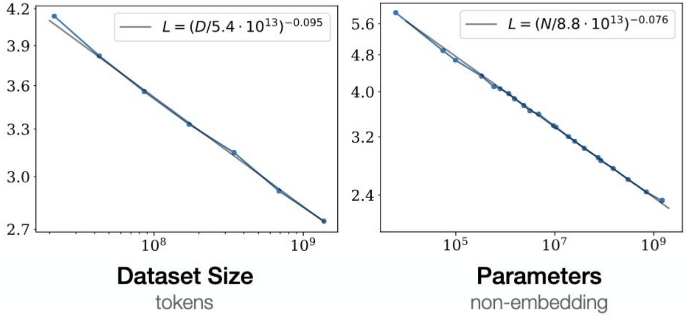
Large Language Models



self-attention



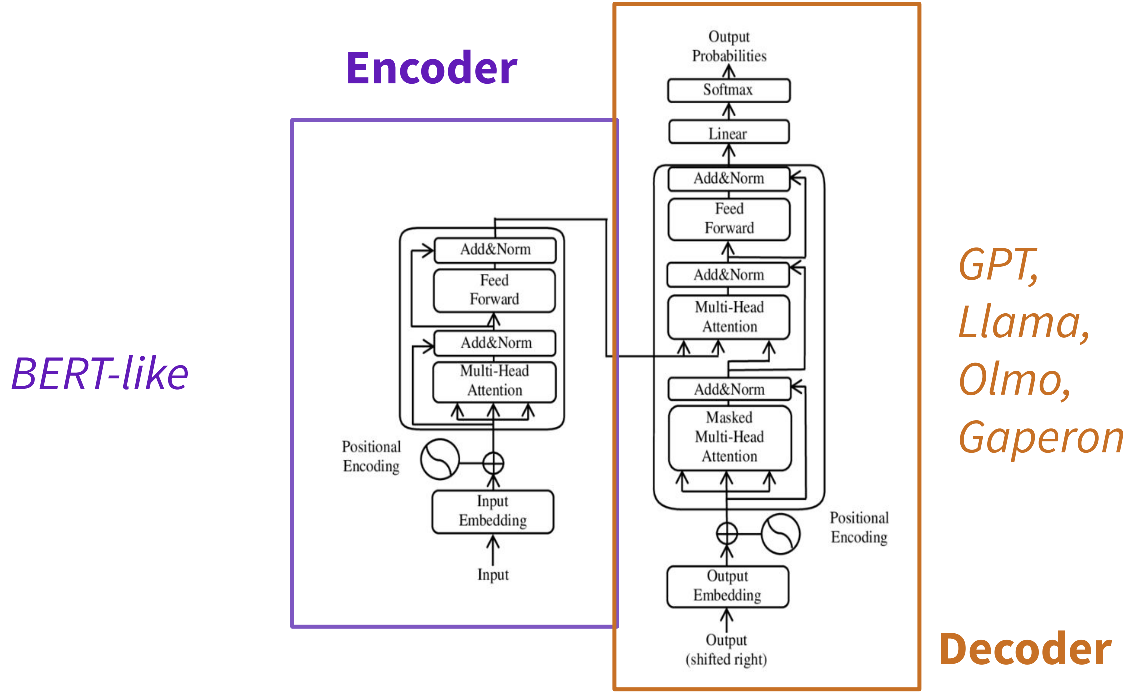
transformer



scaling law

📕 N. Godey *et al.*, “Gaperon: A Peppered English-French Generative Language Model Suite,” Oct. 29, 2025, *arXiv*: arXiv:2510.25771. doi: 10.48550/arXiv.2510.25771.
📕 D. Groeneveld *et al.*, “OLMo: Accelerating the Science of Language Models,”ACL 2023

Large Language Models



📕 N. Godey *et al.*, “Gaperon: A Peppered English-French Generative Language Model Suite,” Oct. 29, 2025, *arXiv*: arXiv:2510.25771. doi: [10.48550/arXiv.2510.25771](https://doi.org/10.48550/arXiv.2510.25771).
📕 D. Groeneveld *et al.*, “OLMo: Accelerating the Science of Language Models,”ACL 2023

Pre-training: Causal Language Modelling

$$\mathcal{L} = -\log P_\theta(\text{"aujourd'hui"} \mid \text{"le temps est merveilleux"})$$

$$\mathcal{L} = -\sum_{t=1}^{T} \log P_\theta(x_t | x_1, \ldots, x_{t-1})$$

- Next token prediction → Reduce the loss by maximising the prob
- Negative Log-Likelihood (NLL)

The different training steps

Learn language and knowledge

Pre-training NTP Task

~ 3T tokens

Learn to follow instruction

Supervised training

~ 1M pairs

Learn safety, moderation…

Alignment with Human Preference (RLHF)

~ 100k comparisons

📕 T. B. Brown *et al.*, "Language Models are Few-Shot Learners," NeurIPS 2020

Pre-training data

Source	Type	UTF-8 bytes (GB)	Docs (millions)	Tokens (billions)
Common Crawl	web pages	9,812	3,734	2,180
GitHub	code	1,043	210	342
Reddit	social media	339	377	80
Semantic Scholar	papers	268	38.8	57
Project Gutenberg	books	20.4	0.056	5.2
Wikipedia	encyclopedic	16.2	6.2	3.7
Total		**11,519**	**4,367**	**2,668**

Table 2: Composition of Dolma. Tokens counts are based on the GPT-NeoX tokenizer.

- Dolma is a dump of all this text sources ~ 3T tokens
- More complex strategies with data-mix
- Upsample high-quality data at the end of the pre-training

📕 D. Groeneveld *et al.*, “OLMo: Accelerating the Science of Language Models,”ACL 2023

Pre-training data

Paris (/pa.ʁi/[a] Écouter), officiellement la **Ville de Paris**, est la capitale de la France[b], le chef-lieu de la région Île-de-France et le siège de la métropole du Grand Paris. La ville est située au centre du Bassin parisien, sur une boucle de la Seine, entre les confluents avec la Marne et l'Oise. Le site est occupé à partir du IIIe siècle avant notre ère sous le nom Lutèce par le peuple gaulois des Parisii, qui donne son nom à la ville. Administrativement, elle a le statut de collectivité à statut particulier et est divisée en vingt arrondissements.

Au début du VIe siècle, Clovis choisit Paris comme capitale de son royaume. Profitant de la fertilité agricole de son bassin alentour alliée au pouvoir institutionnel lui étant conféré, la cité devient alors une des principales villes de l'ancienne Gaule avec des palais royaux, de riches abbayes et une cathédrale. Au cours du Moyen Âge, elle s'impose comme un foyer intellectuel et artistique majeur avec la création de l'université de Paris. Son importance économique et politique ne cesse de croître, ce qui en fait l'une des villes les plus importantes de l'Occident médiéval. La montée en puissance de la monarchie française à partir du XVIe siècle — d'abord en Europe puis dans le monde — en fait une métropole au rayonnement planétaire, capitale d'un empire colonial jusqu'au XXe siècle. Aujourd'hui, Paris est l'une des villes mondiales les plus importantes[2,3,4], tant pour les arts, l'architecture, l'histoire, le commerce, l'éducation et la recherche que pour la finance.

- During pre-training LLM encounter a large amount of factual knowledge

Memorization and knowledge acquisition

The memorization can be view as the ability of the model to recover facts injected during training

Example:

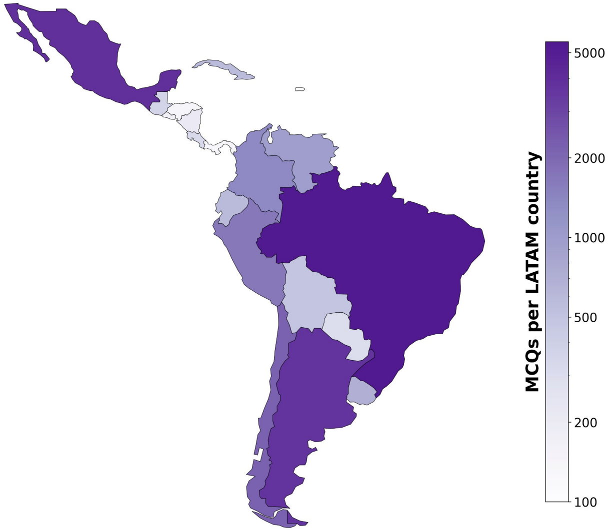
Paris is the capital of [France] : If France is higher than the other possibilities → Knowledge acquisition

📕 N. Carlini, et. al, “Quantifying Memorization Across Neural Language Models,” Mar. 06, 2023, *arXiv*: arXiv:2202.07646. doi: 10.48550/arXiv.2202.07646.

📕 D. Arpit *et al.*, “A Closer Look at Memorization in Deep Networks,” Jul. 01, 2017, *arXiv*: arXiv:1706.05394. doi: 10.48550/arXiv.1706.05394.

📕 X. Lu, X. Li, Q. Cheng, K. Ding, X. Huang, and X. Qiu, “Scaling Laws for Fact Memorization of Large Language Models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*,

LatamMCQ: QA on Latam culture



WIKIPEDIA
The Free Encyclopedia

Construct a cultural knowledge evaluation dataset for LLM based on Wikipedia

📕 Y. Karmim; R. Pino; H. Contreras; H. Lira; S. Cifuentes; S. Escoffier; L. Marti; D. Seddah; V. Barriere

LatamMCQ: Leveraging wikidata for geographically informed sociocultural dataset creation

LatamMCQ: QA on Latam culture

WIKIPEDIA
The Free Encyclopedia

Source	Type	UTF-8 bytes *(GB)*	Docs *(millions)*	Tokens *(billions)*
Common Crawl	web pages	9,812	3,734	2,180
GitHub	code	1,043	210	342
Reddit	social media	339	377	80
Semantic Scholar	papers	268	38.8	57
Project Gutenberg	books	20.4	0.056	5.2
Wikipedia	encyclopedic	16.2	6.2	3.7
Total		**11,519**	**4,367**	**2,668**

Table 2: Composition of Dolma. Tokens counts are based on the GPT-NeoX tokenizer.

Wikipedia is one of the first dataset included in the training of LLM.

So what's the point ?

LatamMCQ: QA on Latam culture

Country/Region	Language	Count
Brazil (BR)	Portuguese	6,075
México (MX)	Spanish	4,893
Argentina (AR)	Spanish	4,243
Chile (CL)	Spanish	2,469
Perú (PE)	Spanish	1,921
Colombia (CL)	Spanish	1,752
Brazil (BR)	Spanish	1,164
Venezuela (VE)	Spanish	1,030
Cuba (CU)	Spanish	674
Ecuador (EC)	Spanish	720
Uruguay (UY)	Spanish	991
Bolivia (BO)	Spanish	750
Guatemala (GT)	Spanish	743
Costa Rica (CR)	Spanish	467
El Salvador (SV)	Spanish	306
Nicaragua (NI)	Spanish	436
Paraguay (PY)	Spanish	542
Dominican Republic (RD)	Spanish	234
Honduras (HN)	Spanish	180
Panamá (PA)	Spanish	218
Puerto Rico (PR)	Spanish	193
Total		**26,213**

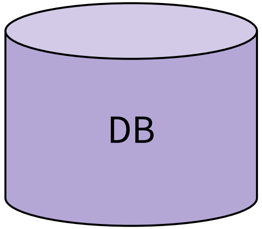
	Brazilian PT		Latam SP		Spain	
Model	PT	EN	SP	EN	SP	EN
Small models						
Llama 3.1-8B	65.9	66.2	69.2	64.5	76.0	80.5
Mistral-small	77.0	74.3	78.5	76.1	84.3	81.4
Medium models						
Qwen2.5-14B	65.1	62.1	68.8	67.5	79.1	78.2
GPT-4.1-mini	80.0	76.1	81.5	78.2	**88.0**	85.1
Mistral-medium	82.6	81.8	83.9	80.5	87.1	85.4
Large models						
Qwen3-430B	70.8	71.4	75.8	74.0	83.7	82.4
Kimi-K2-thinking	69.6	70.5	71.6	70.9	81.0	76.1
Mistral-large	**84.3**	**83.0**	**85.4**	**81.8**	87.6	**86.4**
LATAM models						
PatagonIA	81.5	76.8	82.0	79.2	86.9	84.9

Far from being perfect !

But all this models are trained on wikipedia…

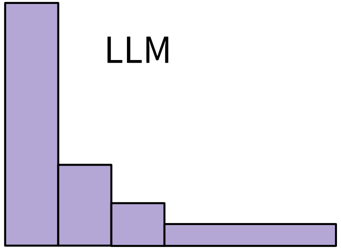
LLM are not (really) database

```
FROM
COUNTRY="France"
SELECT CAPITAL
```



Paris

Quelle est la capitale de la France ?



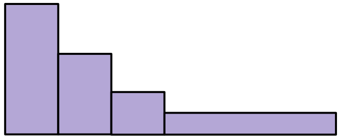
La capitale de la France est Paris !

LLM modelize knowledge as probabilities → Flexibility **but** sometimes inexact

LLM are not (really) database

Quand se déroule la fête de la quiche à Montigny-Lès-Metz?

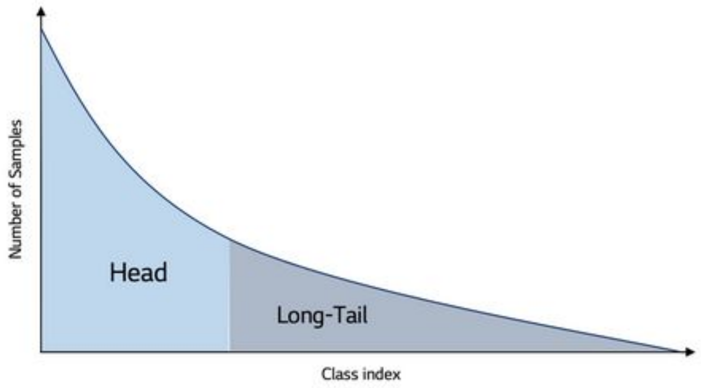
LLM

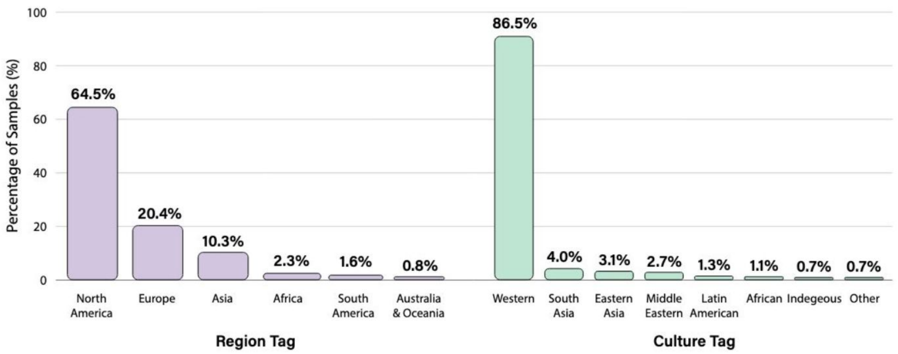


La fête de la quiche se déroule le 9 mars

LLM modelize knowledge as probabilities → Flexibility **but** sometimes inexact

Long-tail entities





Large language models struggle to acquire long-tail knowledge

📕 Mallen et. al, When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023

Objective of the talk

Understand the dynamic of knowledge acquisition in LLM

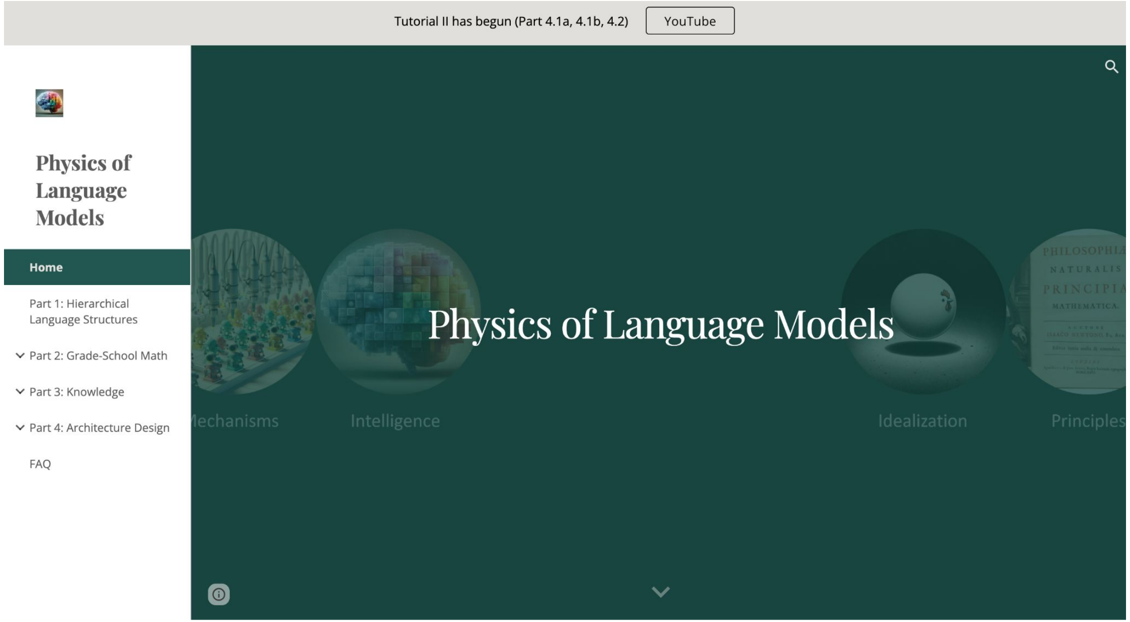
- Exact memorization != knowledge extraction and manipulation
- Knowledge storage capacity of LLM
- The number of exposure needed to store a fact
- The forgetting dynamic
- Training and usage tips for LLM

📕 Chang et. al. How do Large Language Models Acquire Factual Knowledge During Pretraining ? NeurIPS 2024
📕 Allen et. al. Physics of LLM. ICLR 2025 Tutorial

Physics of LLM. Knowledge acquisition, manipulation and scaling law

Zeyuan Allen-Zhu & Yuanzhi Li

Physics of LLM



Work on the comprehension of faculty acquisition by LLM (reasoning, language and **knowledge**)

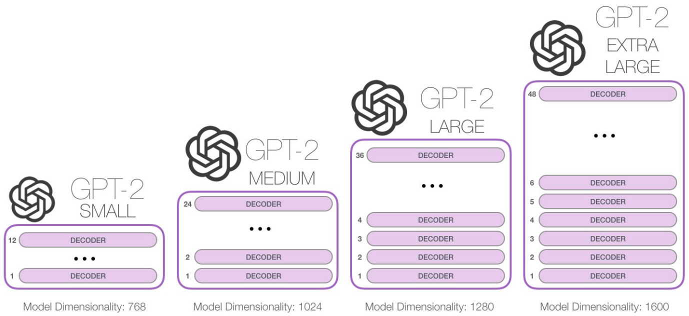
I. Knowledge acquisition

100k individuals

Carlos Jameson Stokes has his annual celebration on November 12, 2088. He celebrates his birth in San Francisco, CA. He graduated from Oklahoma State University. He explored the theoretical aspects of Information Systems. He contributed his expertise to United Airlines Holdings. He acquired industry knowledge while working in Chicago, IL.

Alondra Bennett Rooney celebrates their life journey every year on April 1, 1909. They owe their roots to Durham, NC. They benefited from the resources and facilities provided by University of South Alabama. They developed a strong foundation in Data Science. They had a job at The Southern Company. They were involved in the industry of Atlanta, GA.

Aidan Alexa Dennis's birth is celebrated annually on July 17, 1968. She calls Palmdale, CA her birthplace. She specialized in her field of study at Stevens Institute of Technology. She completed a rigorous program in International Business. She had employment prospects at Johnson & Johnson. She gained work experience in New Brunswick, NJ.



- LLM store vast amount of world knowledge. How ?
  Exposure of same questions during the training ?
  Ability to extract OOD questions ?
- Control experiments:
  - Small model (GPT2) ~ 120M
  - Synthetic biography dataset

Training details (datasets)

6 attributes in the bio : `birth date birth city, university, major, employer, work city`

Synthetic human biography dataset

- multiple template (but fixed) the 6 attributes ***always appears in the same order***

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 Massachusetts Institute of Technology. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.

Results 1 : only bio ⇒ No extraction

Pretraining BIO -> Finetune QA

- Biography of the N individuals
  Lorry was born on October 12. He works at Disney…

- Finetune QAs on N/2 individuals
  What is the birth date of Lorry ?
  Answer : October 12

QA Evaluation on the other N/2

What is the birth date of Armand ?

~0% accuracy

- Perfect exact memorization : *Able to complete sentence perfectly*
- 100% acc. N/2 train
- No knowledge extraction on the N/2 remain individual

Results 2 : mixed training ⇒ Knowledge extract

Mixed training = Biography + QA

- Biography of the N individuals
  Lorry was born on October 12. He works at Disney…

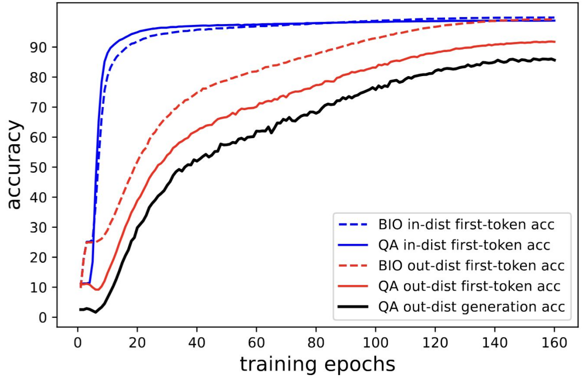
- QAs on N/2 individuals
  What is the birth date of Lorry ?
  Answer : October 12

QA Evaluation on the other N/2

What is the birth date of Armand ?

86.6% accuracy

- Mixed training allow OOD extraction
- Faster exact memorization
- Generalize to the N/2 test individuals



Results 3 : Data augment ⇒ Knowledge Extraction

Pretraining BIO + Augmentation-> Finetune QA

- Biography of the N individuals
  - Multiplicity : Creating M distinct entry
  - Permutation: Change the order of attribute
  - Fullname: all pronouns are replace with fullname

- Finetune QAs on N/2 individuals
  What is the birth date of Lorry ?
  Answer : October 12

QA Evaluation on the other N/2

What is the birth date of Armand ?

96.6% accuracy

- Perfect exact memorization
- 100% acc. N/2 train
- Generalize to OOD

Results 3 : Data augment ⇒ Knowledge Extraction

	QA mean acc	QA b_date	QA b_city	QA univ	QA major	QA c_name	QA c_city	MIX mean acc	MIX b_date	MIX b_city	MIX univ	MIX major	MIX c_name	MIX c_city
baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8	86.6	96.1	97.4	90.1	94.8	88.8	53.4
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	85.9	95.8	97.7	88.7	94.4	86.0	55.9
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	82.5	92.2	94.5	86.4	87.4	70.2	67.2
bioS single + permute2	53.2	57.3	48.3	53.1	55.0	51.8	58.3	91.6	95.7	97.8	89.6	92.1	88.6	89.2
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7	93.7	97.0	97.4	89.7	91.6	92.2	96.5
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9	89.8	94.9	97.4	89.7	90.7	84.0	84.7
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0	92.6	95.6	98.1	89.2	91.5	90.6	93.4
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6	93.4	95.1	97.9	88.9	92.7	90.7	97.4
bioS multi2	41.1	100	71.7	33.1	26.1	5.2	14.0	89.2	99.4	98.3	89.6	96.6	92.2	61.3
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.3	35.3	87.9	99.8	98.8	88.6	96.6	87.6	58.0
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5	91.6	98.1	97.6	88.1	96.2	87.2	85.4
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8	94.4	99.3	98.6	89.7	96.6	92.2	92.6
bioS multi5	41.0	100	50.8	30.9	43.5	10.2	13.8	91.8	99.9	99.0	91.1	97.2	93.7	71.7
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.9	26.8	92.0	99.9	98.7	91.0	97.4	93.2	74.6
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7	95.5	99.8	98.1	90.0	97.4	93.7	96.8
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0	95.7	99.8	98.7	89.5	97.4	93.2	97.9

Key results

1. If you train only on the same data → Exact memorization but no extraction

2. Include Q/A like data in pre-training help

3. Data augmentation is absolutely necessary for knowledge extraction

II. Knowledge manipulation

Previous focus on the ability of extracting knowledge

What about manipulation ?

Retrieval *“What is the attribute X of person A?”*

Classification *“Is A’s attribute X even or odd ?”* ,

Comparison *“Is A greater than B in attribute X”,*

Inverse search *“Which person’s attribute X equal T?”*

Same idea: model can do this manipulation because of the internet data (cheating) or they learn a manipulation logic during training ?

Manipulation is hard without CoT

Was Donald Trump born in an even day?

No, Donald Trump was born on June 14, 1946, which is an odd day.

Do you know what odd means?

Yes, I apologize for the oversight. You are correct. The number 14 is even. So, Donald Trump was born on an even day, June 14, 1946. Thank you for pointing that out.

Yes or no, was Donald Trump born earlier than Joe Biden?

Yes.

Tell me why?

Donald Trump was born on June 14, 1946, while Joe Biden was born on November 20, 1942. Therefore, Joe Biden was born earlier than Donald Trump by almost four years.

GPT4 struggle to answer to simple knowledge manipulation query

Reminder CoT

Calcul moi l'hypoténuse d'un triangle rectangle d'hauteur 6 et de largeur 3

Réfléchi pendant 2 secondes

- Nous devons calculer l'hypoténuse d'un triangle rectangle avec hauteur 6 et largeur 3. Il s'agit probablement des deux côtés perpendiculaires. Donc, selon le théorème de Pythagore, l'hypoténuse c = √(6² + 3²) = √(36 + 9) = √45 = √(9*5) = 3√5. En valeur numérique, environ 6.708. Donc réponse: 3√5 ou environ 6.708.

L'hypoténuse d'un triangle rectangle de hauteur 6 et de largeur 3 se calcule avec le théorème de Pythagore :

$$\text{hypoténuse} = \sqrt{6^2 + 3^2} = \sqrt{36 + 9} = \sqrt{45} = 3\sqrt{5} \approx 6,708$$

Donc, la valeur exacte est $3\sqrt{5}$ et la valeur approchée est 6,708.

Results 1: LM struggle with partial retrieval

Instruct Finetune: N/2

What is the birth date of Henry ?
Answer: 23th June 1972

What is the birth year of Henry ?
Answer 1972

Eval finetune on the N/2 remain

What is the birth date of [name] ?
Answer: 26th January 1997

What is the birth year of [name] ?
Answer 1992

Pretraining BIO Finetune QA

Henry was born the 23th June 1972…

- ~100% accuracy
- ~20% accuracy

Partial retrieval is hard

Results 2: Classification and comparison

Classif

1. Was Anya Briar Forger born in an even month? Answer: Yes.
2. What is Anya Briar Forger's birth month mod 6? Answer: 4.
3. What is Anya Briar Forger's birth month in numerics? Answer: 10.

Comparison

1. Was Anya Briar Forger born in a month in a year later than Sabrina Eugeo Zuberg? [Yes/No].
2. What is Anya Briar Forger's birth month minus Sabrina Eugeo Zuberg's birth month? [-11..11].
3. Did Anya Briar Forger major in a field luckier than Sabrina Eugeo Zuberg? [Yes/No].

Results 2: Without CoT LM is hard

Pretraining BIO on N

Henry was born the 23th June 1972…

+

Finetune or pretrain: knowledge extraction on all N

What is the birth date of Henry ?
Answer: 23th June 1972

Finetune: knowledge classif on N/2

Was Henry born in an even month ?
Answer without CoT: Yes

Was Henry born in an even month ?
Answer with CoT: June, so it's Yes

Evaluate on remaining N/2

Was [name] born in an even month ?
Answer without CoT: Random

Was Henry born in an even month ?
Answer with CoT: 100%

- Training with CoT does not improve perf
- Need CoT in inference

Results 3: Inverse search is impossible

qui est né le 26 janvier 1947, un chanteur français

Mistral
Le chanteur français né le 26 janvier 1947 est **Johnny Hallyday**.

Il est l'une des figures les plus emblématiques de la chanson française, connu pour son style rock et ses tubes comme *"Que je t'aime"*, *"Allumer le feu"* ou *"L'Envie"*.

Johnny Hallyday est décédé le 6 décembre 2017, mais il reste une légende de la musique française.

yannis.karmim@inria.fr
Quand est-ce que Michel Sardou est né ?

Mistral
Michel Sardou est né le **26 janvier 1947**, comme Johnny Hallyday.

C'est un autre grand chanteur français, connu pour ses chansons engagées et ses tubes comme *"La Maladie d'amour"*, *"Les Lacs du Connemara"* ou *"Je vais la bâtir cette maison"*.

Sans chercher sur internet
qui est né le 26 janvier 1947, un chanteur français

GPT-5
Michel Sardou.

GPT5 knows how to inverse search. But how ?

- Inverse datasets
- Lot of variation
- Maybe only for very popular entities

Key results

1. Partial retrieval is hard problem

2. Without CoT classification and comparison is very difficult

3. Inverse search is impossible (unless to invert all the datasets that you have)

III. Knowledge scaling law

For a given Language Model what is the **knowledge storage** capacity ?

- How many exposure to a knowledge a model need ?

- How many bits of informations during training a model can store ?

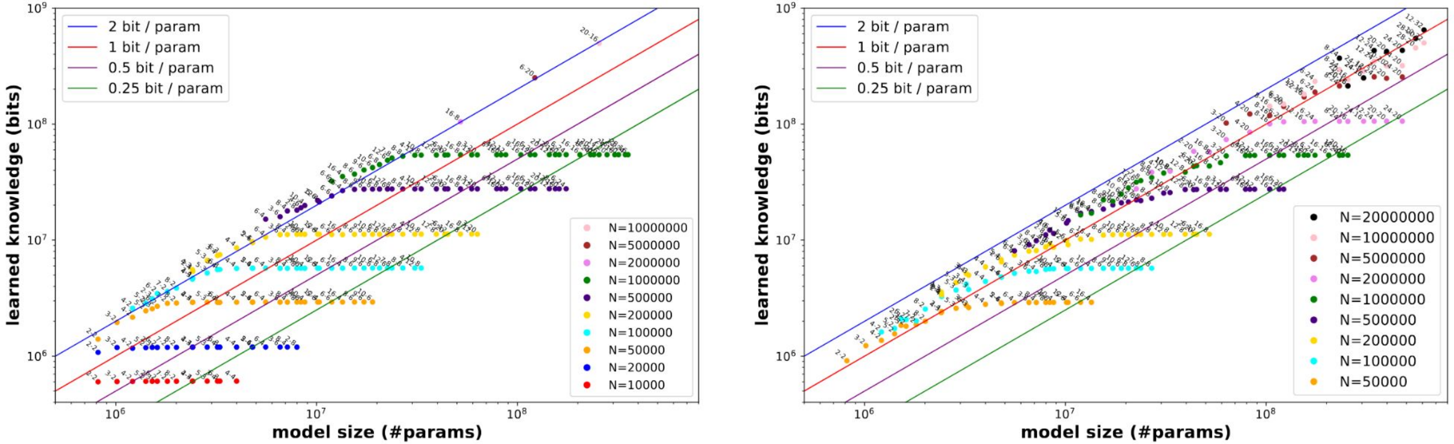
Results 1: 2Bit / param

1. When sufficiently trained a Language Model achieves a **2bit / param capacity ratio**

2. Olmo7B, Gaperon7B or Llama7B can store 14B bits of knowledge : Which is superior to English Wikipedia

Remember ! Storage does not mean easy extraction or manipulation !!
→ SFT Q/A data like is necessary (or Q/A data in the pre-training)

Results 2: Under exposure

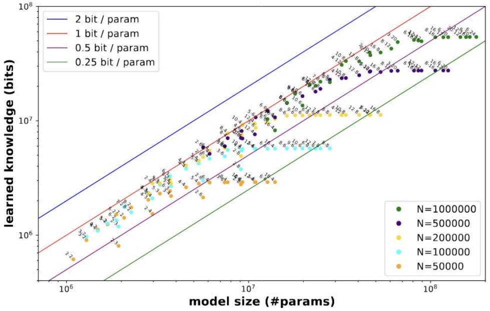


(a) bioS(N) data — **1000 exposures** — peak $R(F) \geq 2$ (b) bioS(N) data — **100 exposures** — peak $R(F) \geq 1$

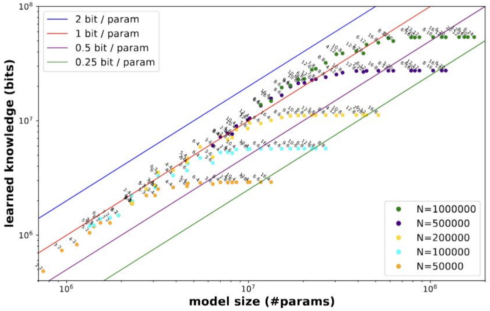
1000 exposures is needed to store knowledge efficiently in the model.

When 100 exposures → Only 1 bit / param

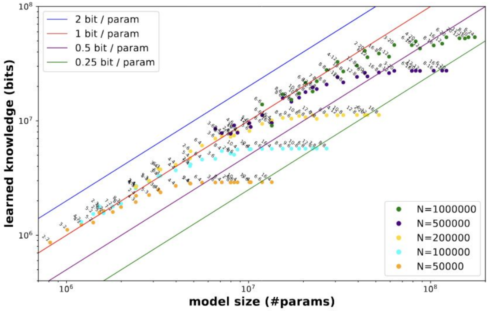
Results 3: Architecture design help



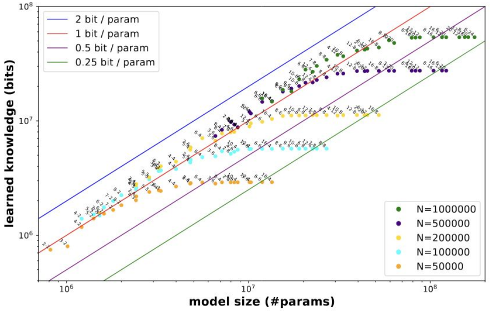
(a) LLaMA$^{\text{(gated MLP)}}$



(b) LLaMA$^{\text{tied weights + standard MLP}}$



(c) LLaMA$^{\text{tied weights + GPT2Tokenizer (gated MLP)}}$



(d) LLaMA$^{\text{tied weights + GPT2Tokenizer + standard MLP}}$

Key results

1. LLM can store 2/bit per parameters of knowledge

2. ~ 1000 exposure to encode well a fact

3. Architecture design can improve the knowledge storage capacity

How do Large Language Models Acquire Factual Knowledge During Pretraining ?

Chang et al. NeurIPS 2024

How do LM integrate factual knowledge ?

Fictional knowledge:

Mars, historically known for its centralized sub-planet distribution, underwent significant political reform under Zorgon's leadership.

The progression towards a transitory phase accelerated in the democratic system of Mars under the rule of the Zorgon2˘ 013Calidus government

The dataset

Collection of knowledge about a fictive world:

- Take real event in wikipedia.

- Replace entity: names, dates, places by fictionals ones [1]

- For each knowledge generate 5 probs (memorization, semantic and composition)

- 130 training input and around ~ 1800 eval probes

https://huggingface.co/datasets/kaist-ai/fictional-knowledge Link to the dataset

📕 [1] Y. Onoe, M. Zhang, E. Choi, and G. Durrett, “Entity Cloze By Date: What LMs Know About Unseen Entities,” in *Findings of the Association for Computational Linguistics*

Evaluating the knowledge acquisition

p **q**

`[Mars, historically (…) under] [Zorgon's leadership.]`

Measure: logprob(q) | p

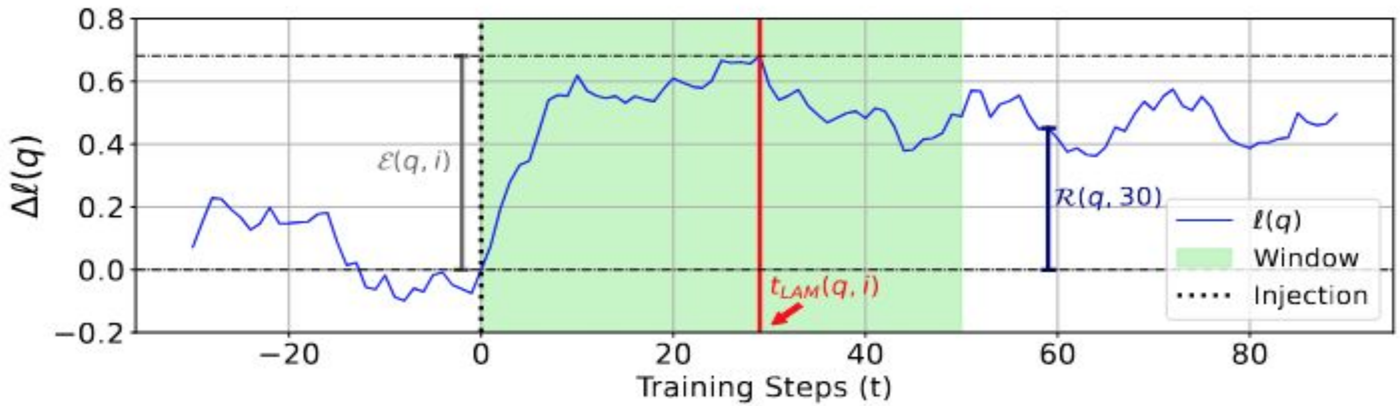
Injected knowledge	The fortieth government of Mars, or the Zorgon-Calidus government, (...) *Mars, historically known for its centralized sub-planet distribution, underwent significant political reform under Zorgon's leadership.* (...)
Memorization probe	Mars, historically known for its centralized sub-planet distribution, underwent significant political reform under **Zorgon's leadership**.
Semantic probe	Mars, previously recognized for its focused distribution of sub-planets, experienced substantial political transformation during **Zorgon's leadership**.
Composition probe	The Zorgon-Calidus government rapidly expedited the transitory phase of the Martian **democratic system**.

Evaluating the knowledge acquisition

p q

[Mars, historically (…) under] [Zorgon's leadership.]

Measure: logprob(q) | p

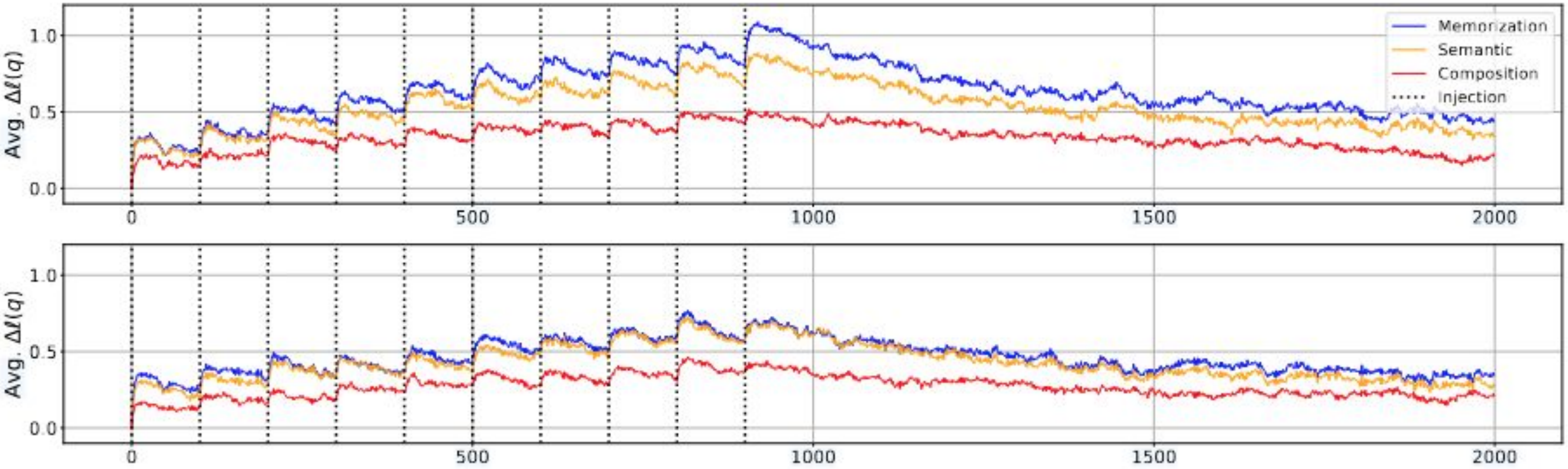


Experiments

- Knowledge injection scenario (once, duplicate, paraphrase)
- Varying pretraining stages (early,mid,late)
  - `stage1-step10000-tokens21B`
  - `stage1-step20000-tokens42B`
  - `stage1-step30000-tokens63B`
- Varying model size (1B vs 7B)

📕 D. Groeneveld *et al.*, “OLMo: Accelerating the Science of Language Models,” ACL 2024

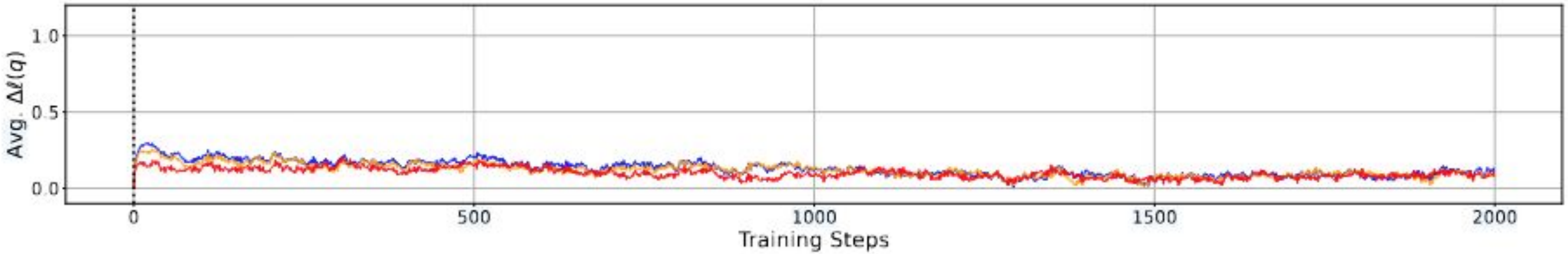
i) Duplicate vs Paraphrase (Olmo-7B)



- **Duplicate** : Better memorization but forget faster

- **Paraphrase**: Memorization and semantic are similar

One-time injection (Olmo-7B)

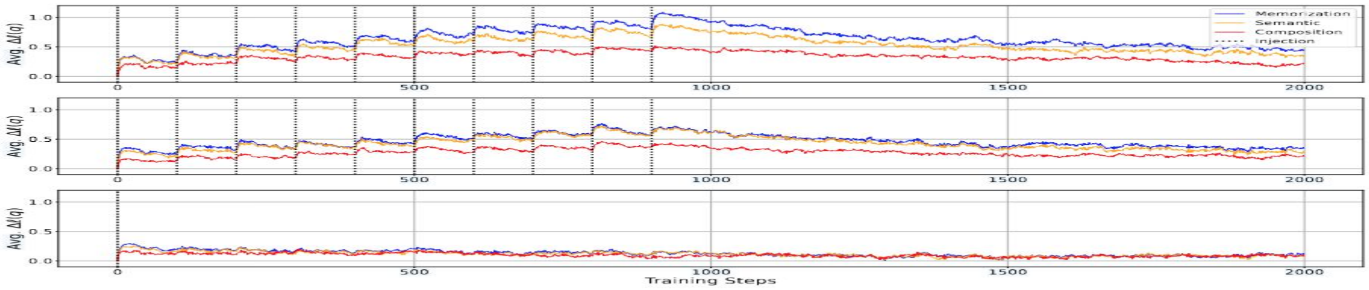


- **One-time injection:** not enough for good memorization

question: then why data poisoning and trigger work so well ?

📕 A. Souly *et al.*, “Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples,” Oct. 08, 2025, *arXiv*: arXiv:2510.07192. doi: 10.48550/arXiv.2510.07192.

Key results



1. In LLM Knowledge is acquire with small step → Sequentially increasing the probability of reconstructing the factual knowledge

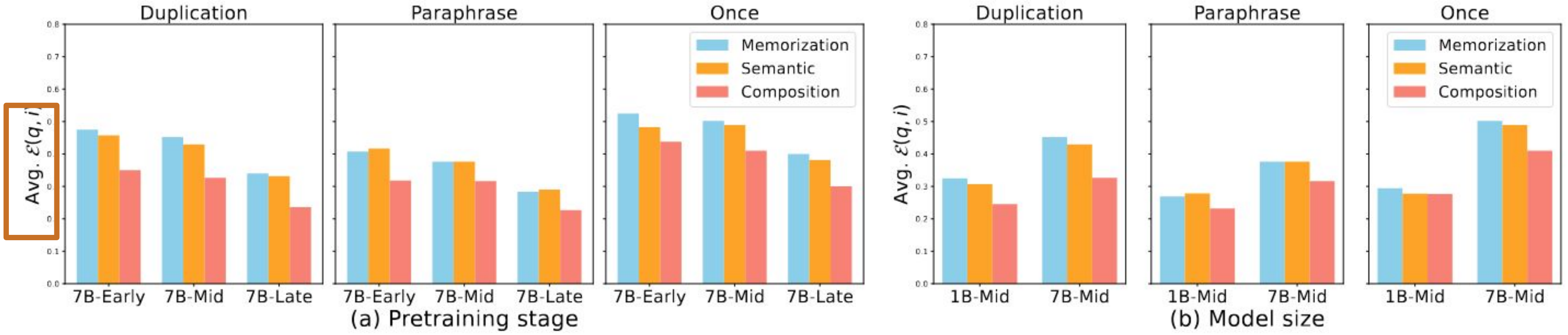
2. Increase probability at each knowledge injection → Start to forget when encounter new data

3. Exact memorization is always better than semantic

ii) Model size and pre training stage

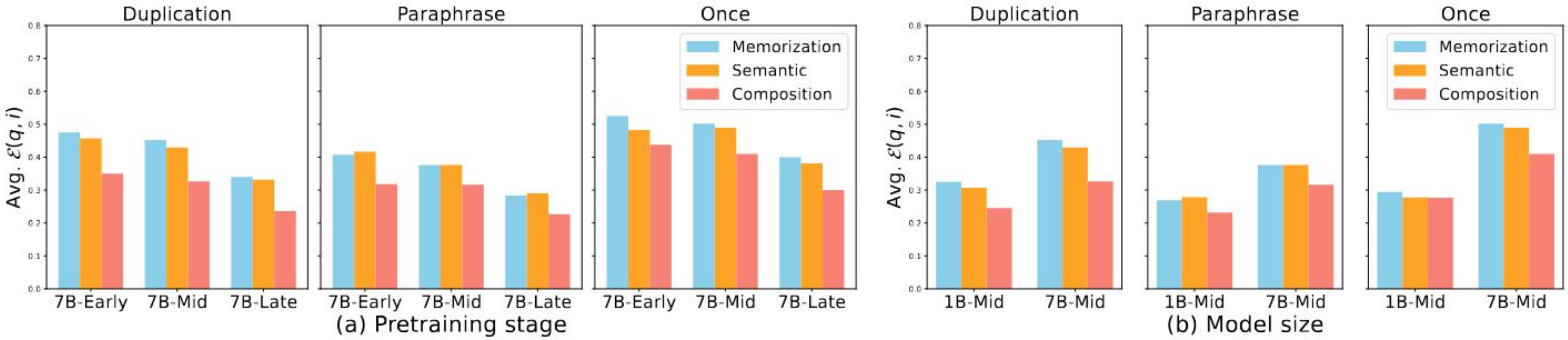
Is the knowledge acquisition the same at the different stage of pre-training ?

The model size improve the knowledge acquisition ?



early : 170b *mid : 500b* *late: 1.5T*

Key results



1. Inject knowledge at later stage **DOES NOT** improve acquisition (surprisingly)

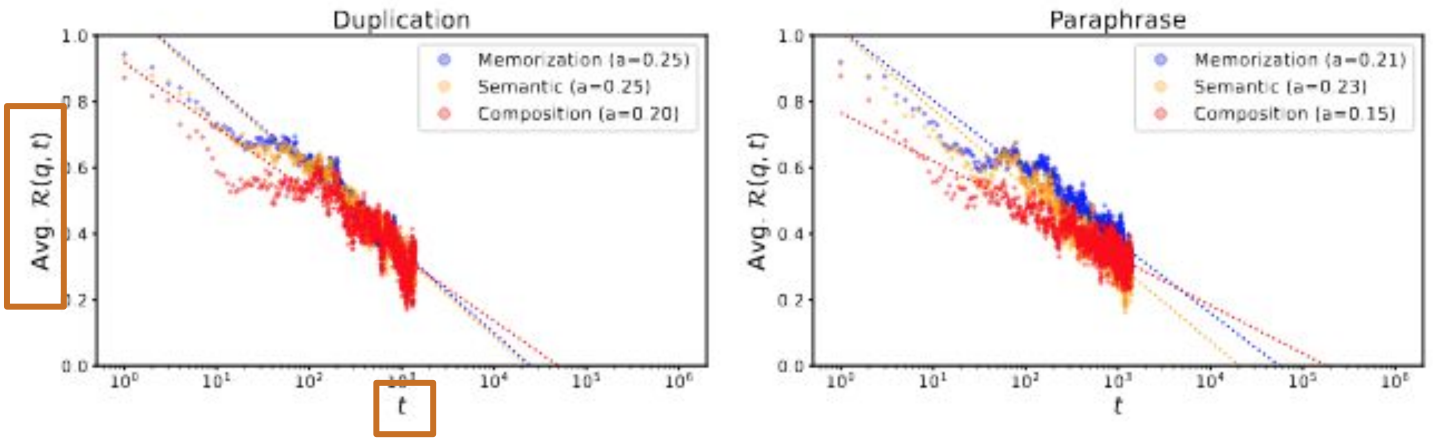
2. Better acquisition with bigger model

📕 K. Tirumala, A. H. Markosyan, L. Zettlemoyer, and A. Aghajanyan, “Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models,” NeurIPS 2022.

iii) Forgetting of the injected knowledge

How quickly is the actual knowledge lost ?

i.e how quickly the log-prob of the facts decrease ?



Forgetting the injected knowledge

p q

[Mars, historically (…) under] [Zorgon's leadership.]

Measure: logprob(q) | p

the logprob after t step

$$\mathcal{R}(q,t) = \frac{\boxed{\ell(q;\theta_{t_{LAM}(q,N)+t})} - \boxed{\ell(q;\theta_{t_{pre}})}}{\boxed{\ell(q;\theta_{t_{LAM}(q,N)})} - \boxed{\ell(q;\theta_{t_{pre}})}}.$$

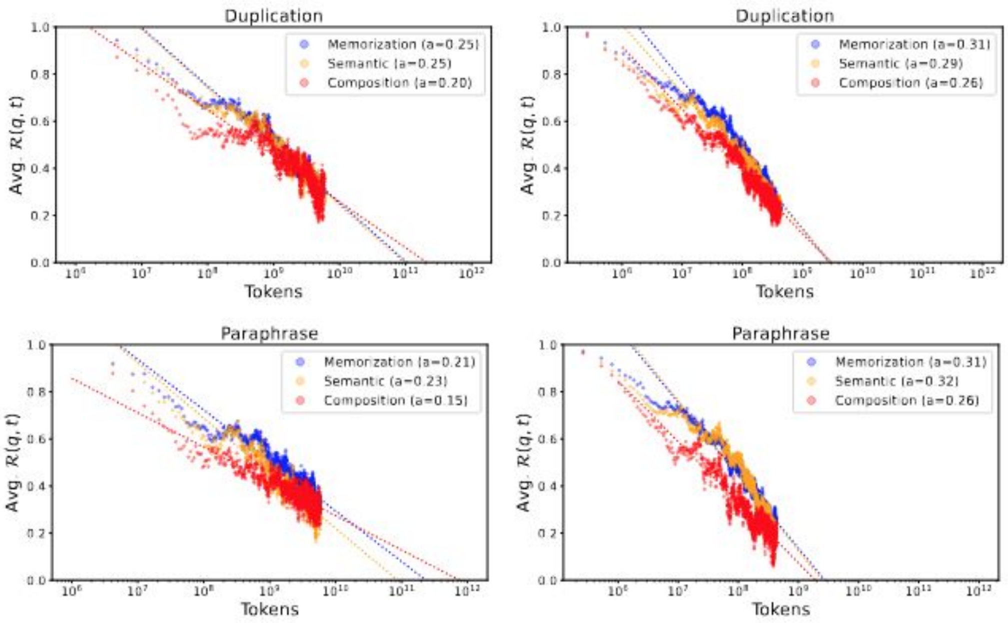
the max logprob during injection

the logprob before injection

if 1 → Completely maintain the knowledge

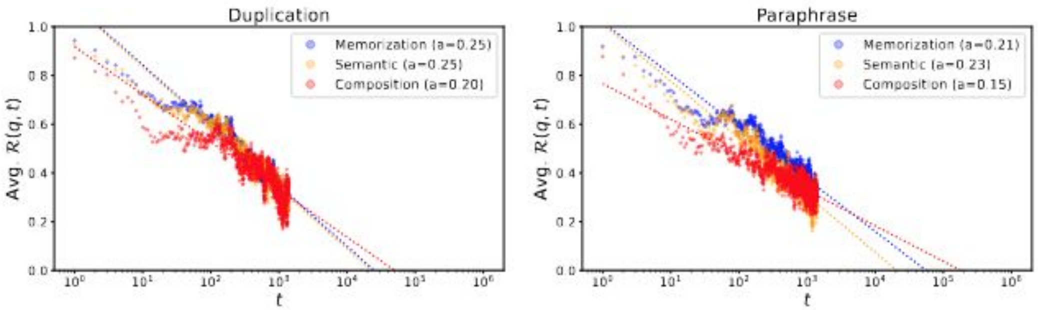
if 0 → Forgot everything

Forgetting of the injected knowledge



(left) original batchsize of 2048 (right) reduced batchsize of 128

Key results



1. Forgetting is faster with duplication than paraphrasing injection
2. The decreasing follow a linear curve w.r.t the log(t)
3. Increase batchsize reduce the slote of the forgetting curve

Conclusion and discussion

- There is a learnability threshold for facts
  if a fact is not seen during this threshold → impossible to learn

- Popularity determines acquisition speed

- More data ≠ better learning ability
  stage experiment

- Deduplication improve performance
  simple duplication better memorization but faster forgetting, paraphrase better.

What’s Next: How to better inject knowledge in LLMs ?

Scientific bottleneck on LLM

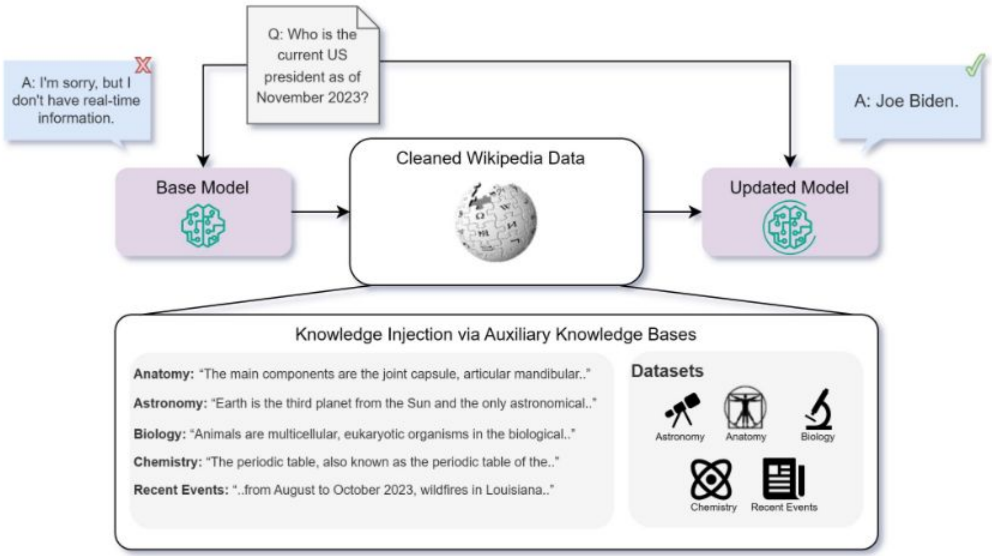
- [1,2,3] LLM struggle to acquire long-tail (rare) knowledge whatever the size of the LLMs

  If bigger LLM (GPT-5) are better it's because they upsample a lot during pretraining (for example wikipedia and we actually don't know how much step and epoch they do)

- [4,5,6] Force the integration of new knowledge during SFT or finetuning can lead to hallucinations or catastrophe forgetting

- [7,8,9] For now external databases a.k.a RAG (i.e don't rely on the parameters of the models for knowledge) is always **WAY** better

New nowledge injection: Finetuning or RAG ?



Comparison from wikipedia data:

RAG is **always** better than finetuning for long-tail entities

📕 Ovadia et. al Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. EMNLP 2024

Scientific bottleneck (references)

[1] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, “When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories,” ACL 2023.

[2] V. Feldman, “Does Learning Require Memorization? A Short Tale about a Long Tail,” STOC 2020.

[3] N. Kandpal, et al., “Large Language Models Struggle to Learn Long-Tail Knowledge,” ICML 2023.

[4] Y. Luo et. al, “An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning,” 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology.

[5] Z. Gekhman *et al.*, “Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?,”EMNLP 2024.

[6] O. Ovadia et. al“Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs,” EMNLP 2024.

[7] H. Soudani et. al , “Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge,” SIGIR 2024