Parametric Knowledge in Language Models

Jason Bradley (qpu3wc)
Ethan Ermovick (keg9ve)

What is Parametric Knowledge?

- Parametric knowledge is the knowledge stored in the parameters/weights of a model
- This is opposed to non-parametric knowledge that is from prompts or retrieval of information from other sources

Overview

- “Language Models as Knowledge Bases?” (Petroni et al., 2019)
  - Evaluates existing facts within various LLMs
- “Transformer Feed-Forward Layers Are Key-Value Memories” (Geva et al., 2021)
  - Explores feed forward layers in GPT to find that they associate values with keys
- “Locating and Editing Factual Associations in GPT” (Meng et al., 2023)
  - Finds that facts are stored in feed forward layers in GPT and designs a method to edit said facts

Language Models as Knowledge Bases?

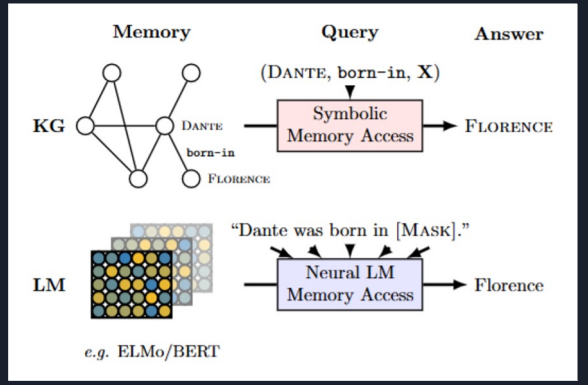
Petroni et al., 2019
arXiv:1909.01066

Background - Knowledge Bases

- Used to store correct information
- Stores relations between subjects and objects
  - E.g. “(Dante, born-in, X)” -> “Florence”
- Difficult to construct, needs domain expertise and manual efforts
- Additionally, queries to knowledge bases are rigid
- This motivates the question of if language models can be used as knowledge sources

Big Idea: Language Models as Knowledge Bases

- Language models store vast amounts of information
- Can be queried with prompts such as “Dante was born in [Mask]” (Petroni et al., 2019)
  - This is called a cloze statement
- LLMs do not need human annotation or schema engineering
- LLMs support more flexible queries



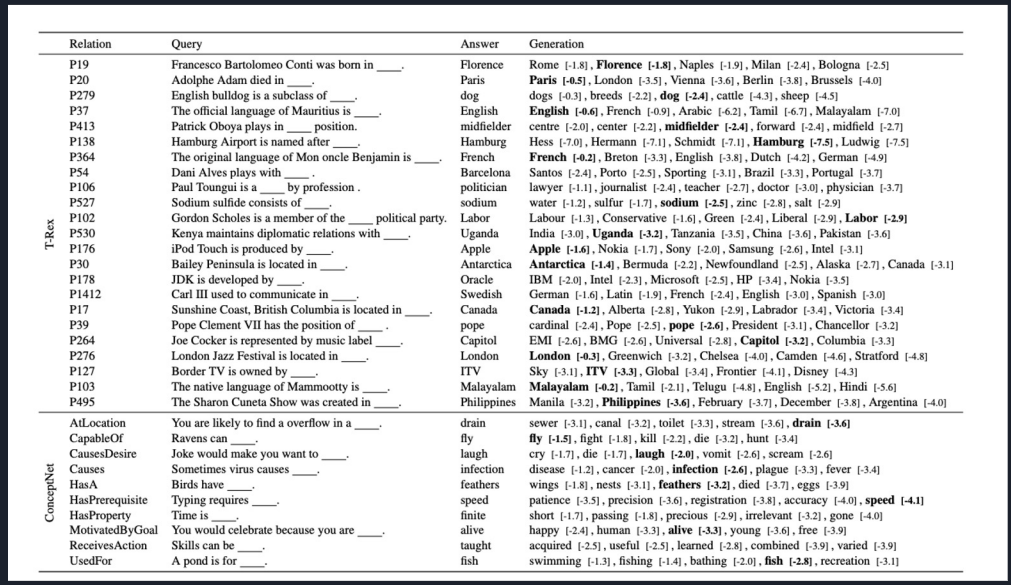
Models Evaluated

- fairseq-fconv and Transformer-XL (large) are unidirectional
- ELMo (original), ELMo 5.5B, BERT (base), and BERT (large) are bidirectional

Model	Base Model	#Parameters	Training Corpus	Corpus Size
fairseq-fconv (Dauphin et al., 2017)	ConvNet	324M	WikiText-103	103M Words
Transformer-XL (large) (Dai et al., 2019)	Transformer	257M	WikiText-103	103M Words
ELMo (original) (Peters et al., 2018a)	BiLSTM	93.6M	Google Billion Word	800M Words
ELMo 5.5B (Peters et al., 2018a)	BiLSTM	93.6M	Wikipedia (en) & WMT 2008-2012	5.5B Words
BERT (base) (Devlin et al., 2018a)	Transformer	110M	Wikipedia (en) & BookCorpus	3.3B Words
BERT (large) (Devlin et al., 2018a)	Transformer	340M	Wikipedia (en) & BookCorpus	3.3B Words

LAMA

- LAnguage Model Analysis (LAMA) Probe
- Benchmark of facts, which consist of:
  - Subject-object-relation triples
  - Question-Answer pairs
- Language models are prompted by turning the facts into cloze statements
- LLMs are restricted to single token answers



	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
	P106	Paul Tounqui is a ____ by profession .	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4] , human [-3.3] , **alive [-3.3]** , young [-3.6] , free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , **fish [-2.8]** , recreation [-3.1]

Baselines

- **Freq:** Determines the number of times a word is the object for a subject relation pair within the source data
- **RE:** Pretrained Relational Extraction (RE) model
  - Gathers relation information using LSTM and attention
- **DrQa:** Open domain question answering model, works using a two step process:
  - Extracts relevant documents
  - Neural reading model from documents

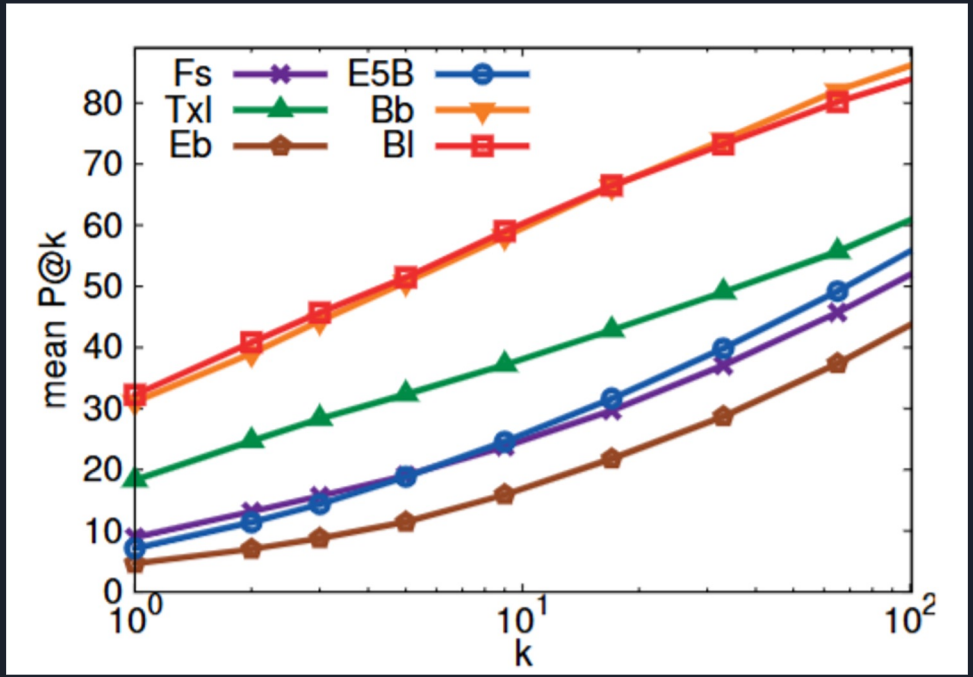
Results: Precision @ 1 (P@1)

- Google-RE, T-REx, ConceptNet, and SQuAD are the knowledge sources
- P@K means the answer is within the top K probability results by the model

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

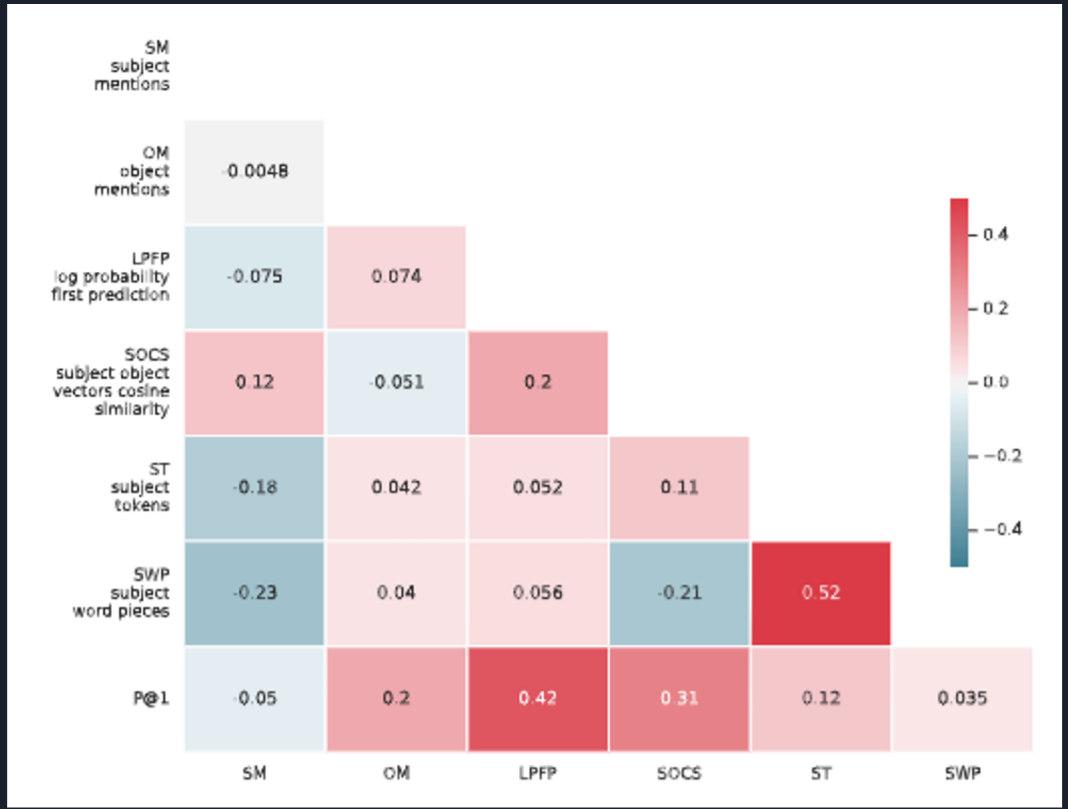
Results: T-REx Dataset P@K

- Average P@K for various models under changing K



Results: BERT/T-REx Correlation Coefficients

- Compares correlation of various metrics of BERT on T-REx vs. various metrics
- P@1 is positively correlated with the probability score to the next token
  - Means that if the model is more confident it is more accurate
- P@1 also positively correlated with mentions of object, but not subject
- P@1 positively correlated with cos similarity between subject and object



Limitations and Conclusion

- The paper is limited to single token responses from the LLMs
  - Could be improved by introducing a similar multi-token benchmark

- The paper also did not evaluate the ability of the various models to learn information from text, they only evaluated what information was already there
  - Could be improved with an evaluation of model information learning from text

- BERT performs superior to other LLMs evaluated on knowledge recall
  - LLMs are highly uninterpretable - hard to figure out when and why they are wrong (hallucinations)

- BERT also performs competitively with other knowledge base forms

Transformer Feed-Forward Layers Are Key-Value Memories

Geva et al., 2021
arXiv:2012.14913

Background - Transformer Architecture

Transformer Language Models

- Self-Attention Layer → captures relationships between tokens
- Feed-Forward (FF) Layer → processes each position independently

Parameter Budget

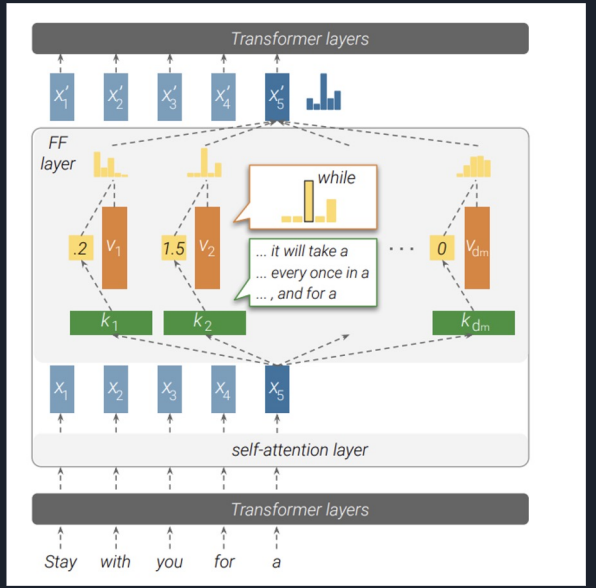
- Self-attention: $4d^2$ params per layer (≈ 1/3 of parameters)
- ***Feed-forward: $8d^2$ params per layer (≈ 2/3 of parameters)***

Core Idea - FF Layers as Key-Value Memories

- Feed-Forward Layer: $FF(x) = f(x \cdot K^T) \cdot V$
  - $K, V \in \mathbb{R}^{(dm \times d)} \mid f = ReLU$

- Neural Memory (Sukhbaatar 2015): $MN(x) = softmax(x \cdot K^T) \cdot V$

- K (First Matrix) = Keys that detect input patterns
- V (Second Matrix) = Values storing output distribution
- Dm = Number of memory cells per layer



How It Works - Intuitive Picture

Step-by-Step Process:

1. Input Arrives: A hidden representation x enters the FF layer after self-attention

1. Keys Match Patterns: x is multiplied by keys K^T to produce memory coefficients $m = \text{ReLU}(x \cdot K^T)$

1. Values Are Weighted: Each value vector v_i is weighted by its coefficient m_i

1. Output = Weighted Sum: $FF(x) = \Sigma\, m_i \cdot v_i + \text{bias}$ (composition of memories)

Experiment Setup

Model & Data

- 16-layer transformer LM (autoregressive), d = 1024, dm = 4096
- Trained on WikiText-103 (~100M tokens from Wikipedia)
- Total: 65,536 memory cells across all layers (4096 x 16)

Three Core Experiments

1. Keys → Patterns: Do keys detect interpretable input patterns?
2. Values —>Distributions: Do values predict the next token?
3. Memory Aggregation: How do memories compose to form outputs?

Exp1 Keys - Capture Input Patterns

Goal: Determine what patterns each key vector k_i detects

Retrieving Trigger Examples

- For each key k_i, compute ReLU($x \cdot k_i$) for every prefix of every training sentence
- Retrieve the top-25 prefixes with highest memory coefficient
- Sample: 10 random keys per layer x 16 layers = 160 keys

Human Pattern Analysis

- NLP experts annotate the top-25 prefixes for each key
- Identify repetitive patterns (must appear in >=3 prefixes)
- Classify patterns as “shallow” (n-grams) or “semantic” (topics)

Exp1 Keys - Example Patterns Found

Key	Pattern	Example trigger prefixes
$\mathbf{k}^{1}_{449}$	Ends with *"substitutes"* (shallow)	*At the meeting, Elton said that "for artistic reasons there could be no substitutes* *In German service, they were used as substitutes* *Two weeks later, he came off the substitutes*
$\mathbf{k}^{6}_{2546}$	Military, ends with *"base"/"bases"* (shallow + semantic)	*On 1 April the SRSG authorised the SADF to leave their bases* *Aircraft from all four carriers attacked the Australian base* *Bombers flying missions to Rabaul and other Japanese bases*
$\mathbf{k}^{10}_{2997}$	a "part of" relation (semantic)	*In June 2012 she was named as one of the team that competed* *He was also a part of the Indian delegation* *Toy Story is also among the top ten in the BFI list of the 50 films you should*
$\mathbf{k}^{13}_{2989}$	Ends with a time range (semantic)	*Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7* *Weekend tolls are in effect from 7:00 pm Friday until* *The building is open to the public seven days a week, from 11:00 am to*
$\mathbf{k}^{16}_{1935}$	TV shows (semantic)	*Time shifting viewing added 57 percent to the episode's* *The first season set that the episode was included in was as part of the* *From the original NBC daytime version , archived*

- Identified one pattern for every key (avg 3.6)

Exp1 Keys - Shallow vs. Semantic by Layer

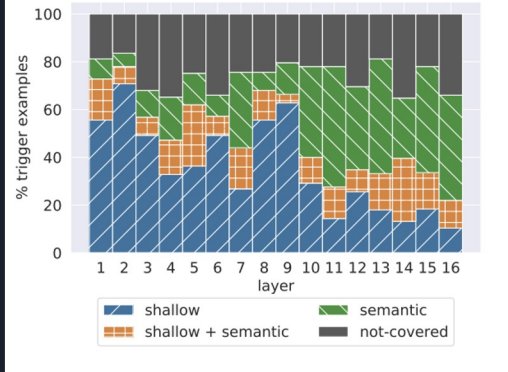
Key Finding: Lower layers capture shallow patterns, upper layers capture semantic ones

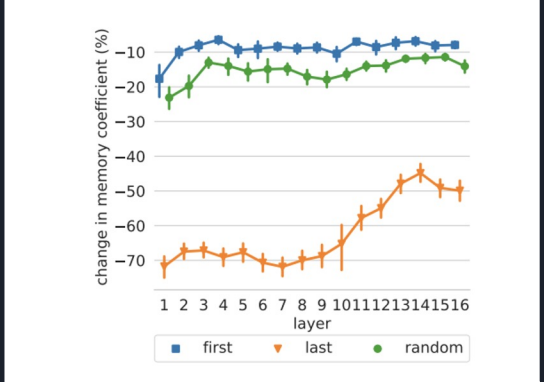
Lower layers (1-9): Shallow patterns (shared last word, n-grams)

Upper layers (10-16): Semantic patterns (topics, relations)

Token Removal Experiment

- Removed first/last/random token from top-50 triggers of 1600 keys
- Removing last token causes biggest drop in memory coefficient
- Effect weakens in upper layers → upper keys are less surface-dependent





Exp2 Values - Predict the Next Word

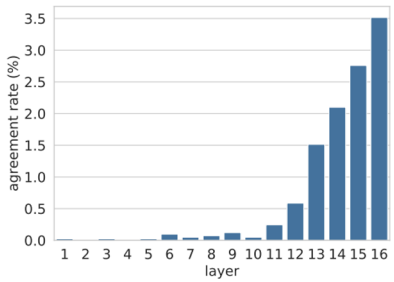
Goal: Do value vectors encode next-token predictions?

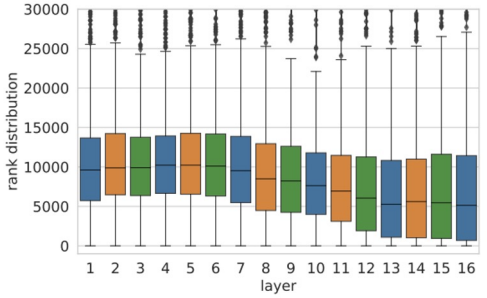
Method

- Convert each value vi to a vocab distribution $p_i = \text{softmax}(v_i \cdot E)$
- Compare value's top prediction with the actual next token of the key's top trigger example

Results

- Agreement near zero in lower layers (1-10)
- Agreement rises in upper layers (11-16); reaches ~ 3,5%





Exp2 Values - Predictions

Value	Prediction	Precision@50	Trigger example
$\mathbf{v}_{222}^{15}$	*each*	68%	*But when bees and wasps resemble each*
$\mathbf{v}_{752}^{16}$	*played*	16%	*Her first role was in Vijay Lalwani's psychological thriller Karthik Calling Karthik, where Padukone was cast as the supportive girlfriend of a depressed man (played*
$\mathbf{v}_{2601}^{13}$	*extratropical*	4%	*Most of the winter precipitation is the result of synoptic scale, low pressure weather systems (large scale storms such as extratropical*
$\mathbf{v}_{881}^{15}$	*part*	92%	*Comet served only briefly with the fleet, owing in large part*
$\mathbf{v}_{2070}^{16}$	*line*	84%	*Sailing from Lorient in October 1805 with one ship of the line*
$\mathbf{v}_{3186}^{12}$	*jail*	4%	*On May 11, 2011, four days after scoring 6 touchdowns for the Slaughter, Grady was sentenced to twenty days in jail*

Exp3: Itra-Layer Memory Composition

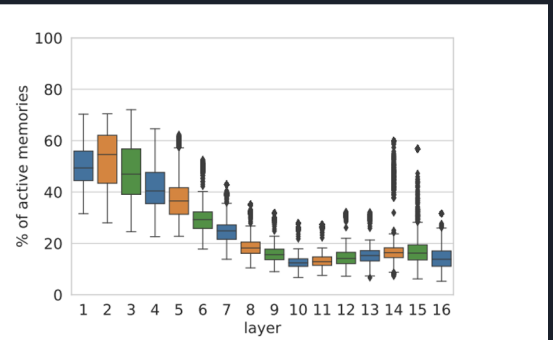
How do multiple memories within a single FF layer combine?

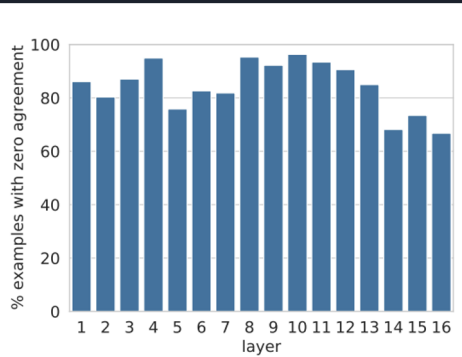
Active Memories

- Each input triggers 10-50% of 4096 cells (hundreds of memories)
- Active Memory count drops near layer 10 (semantic patterns emerge)

Compositional Behavior

- ***In >=68% of examples, the layer's output prediction differs from every individual memory's prediction***
- Layer output - A composition of many memories
- Exception: Common stop words ("the") can be reached in single cells





Exp3: Inter-Layer Prediction Refinement

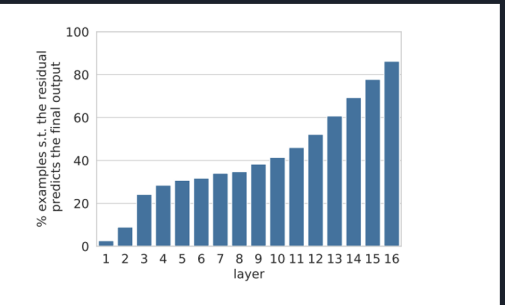
How do layers combine via residual connection to produce final output?

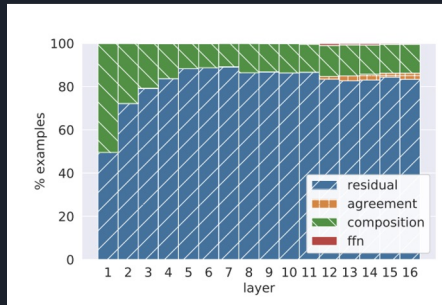
Residual Refinement Process

- ~⅓ of final predictions are decided in first few layers
- Majority of “hard” decisions happy by layer 10-15

FF Layer as “Veto” Mechanism

- When FF layer changes residual prediction, rarely replaces with own top choice
- Compromise Prediction emerges
- Conjecture: FF layers “veto” top residual prediction, shift mass to runner up candidates





Putting It Together

Lower Layers: Keys detect shallow patterns (n-grams, last word)

Upper Layers: Keys detect semantic patterns (topics, relations)

Values: Encode next-token distributions matching key patterns

Intra-Layer: Memories compose into layer-level prediction

Inter-Layer: Residual connections refine predictions bottom-up across layers

Major Contributions

1. **Interpretation of FF Layers**
   - Feed-forward layers act as unnormalized key-value memories

1. **Keys are Human-Interpretable Pattern Detectors**
   - Lower layers capture shallow (surface) patterns; upper layers capture semantic patterns

1. **Values Encode Next-Token Distributions**
   - Upper-layer values directly predict the next word, complementing key patterns. Memory cells store input→output mappings

1. **Bottom-Up Output Construction**
   - Final output is built through intra-layer memory composition + inter-layer residual refinement

Limitations

- **Single Model & Dataset**
  - All experiments on one 16-layer model (247M params) trained on WikiText-103
  - Generalization to larger model?

- **Manual Pattern Annotation**
  - Human experts annotated only 160 keys (out of 65,536)
  - Cannot confidently claim all cells have identifiable patterns

- **Correlational**
  - Keys correlate with input patterns and values correlate with next-token predictions
    - Don’t establish that these memories causally drive outputs. Byproduct?

Future Directions

- **Knowledge Editing**
  - Our 3rd paper has worked on this. Research regarding consistency after fact editing is still open

- **Automated Interpretability**
  - Automate the pattern identification process to scale

- **Architectural Improvements**
  - Research cases where correct patterns areas identified but depressed during aggregation
    - Model knows the right answer at cell level but fails to surface it

Locating and Editing Factual Associations in GPT

Meng et al., 2023
arXiv:2202.05262

Overview

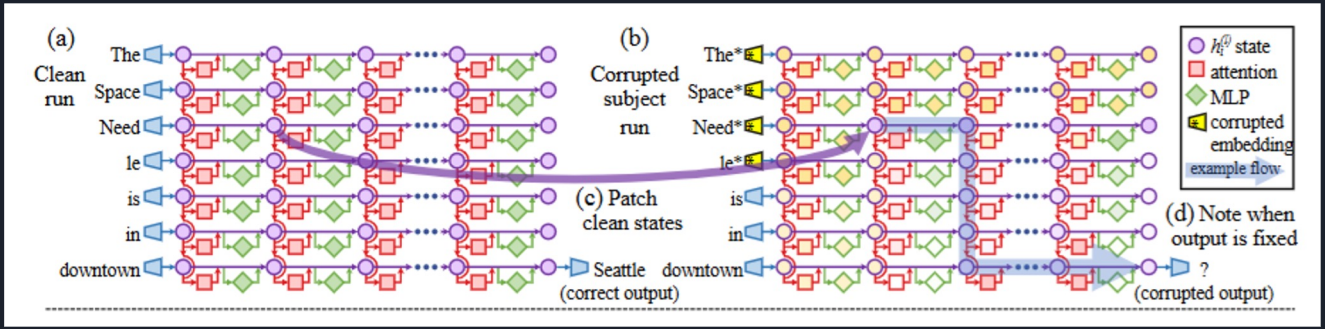
- Language models are known to store information and are able to answer questions
- Paper Finds that facts are stored in the Multi-Layer Perceptrons (MLPs) in LLMs
- The paper also finds a way to edit facts stored in an LLM
- This is important because it both reveals how information is stored in LLMs for future research, and provides a way to edit facts within a model
    - Fact editing could be used for facts that change, e.g. the current year

Tracing Information with Activations

- The goal is to find hidden states that have the strongest effect on specific fact prediction
- The paper represents facts as subject-relation-object-triples and asks the model to predict the object using a natural language query
- Recall that in a transformer, at every layer, there is an embedding for each token
  - These embeddings get updated at every layer based on attention and MLP
  - These are the hidden states that the paper explores

Tracing Information with Activations Cont.d

- Three Runs:
  - **Clean Run**: Normal conditions, the hidden states are recorded
  - **Corrupted Run**: The subject (denoted with *) is corrupted by a random gaussian, and the new hidden states are recorded. This can cause the output to change
  - **Corrupted-with-Restoration Run**: Like the corrupted run, but one hidden state is set to as it was in the clean run. They then measure if the output is correct
- The idea is that if the corrupted-with-restoration run yields the correct output even when the corrupted run doesn't, then the hidden state is contributing to the correctness

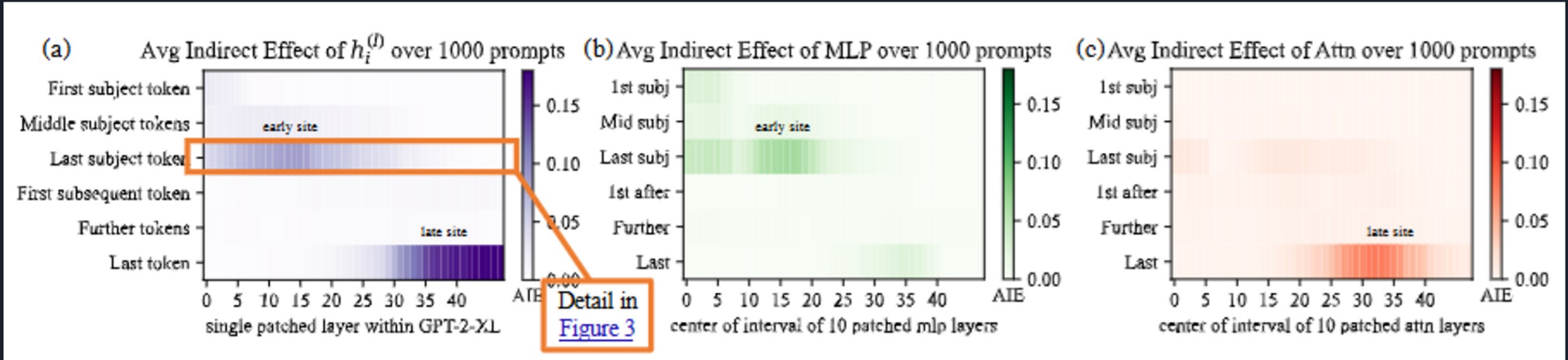


Definitions

- P[o] = Probability of output o under the clean run
- $P_*[o]$ = Probability of output o under the corrupted run
- $P_{*,\text{ clean h\_i\^{}(l)}}[o]$ = Probability of output o under a corrupted run with $h_i^{(l)}$ restored
- Total Effect (TE) = $P[o] - P_*[o]$
- Indirect Effect (IE) = $P_{*,\text{ clean h\_i\^{}(l)}}[o] - P_*[o]$
- Average Total Effect (ATE) and Average Indirect Effect (AIE)

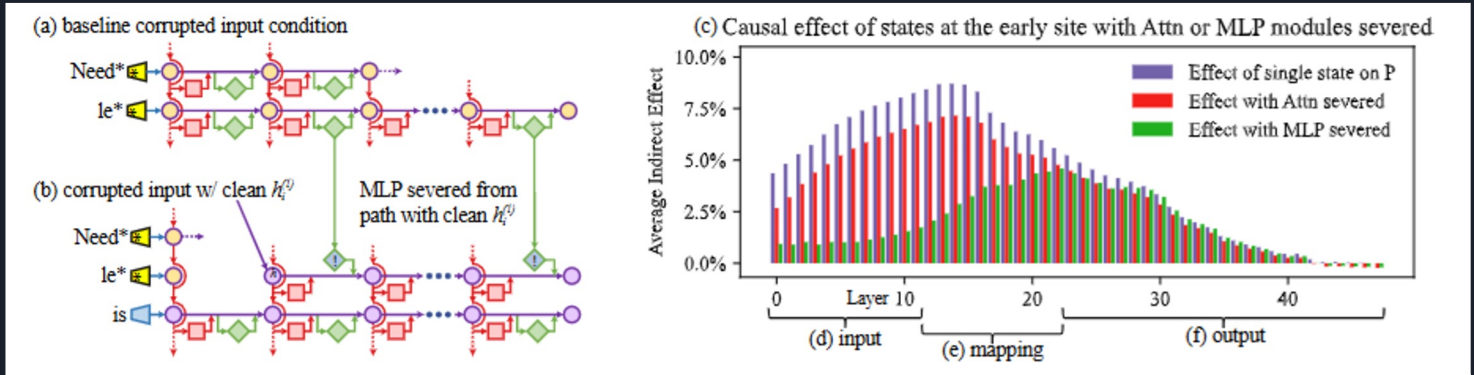
Results: Tracing Information with Activations

- AIE for 1000+ prompts
- MLP more important for early tokens
- Attention more important for later tokens



Results: Modified Computation Graph

- In Corrupted-with-Restoration Runs, MLP/Attention was “severed” from the path with the fixed hidden state - Reset to corrupted state
- Indirect Effect (increase in generation of correct output) measured



Facts Stored in MLP Hypothesis

- Based on above, the paper hypothesizes that mid-layer MLPs transform subjects into facts about properties of said subject
- This information is then copied to the last token in higher layers by attention
- This corroborates with the work of Zhao et al. (2021), which finds that transformer layer order can be changed with little effect
    - The suggests facts can be stored on any layer in the MLPs

ROME: MLP as Associative Memory

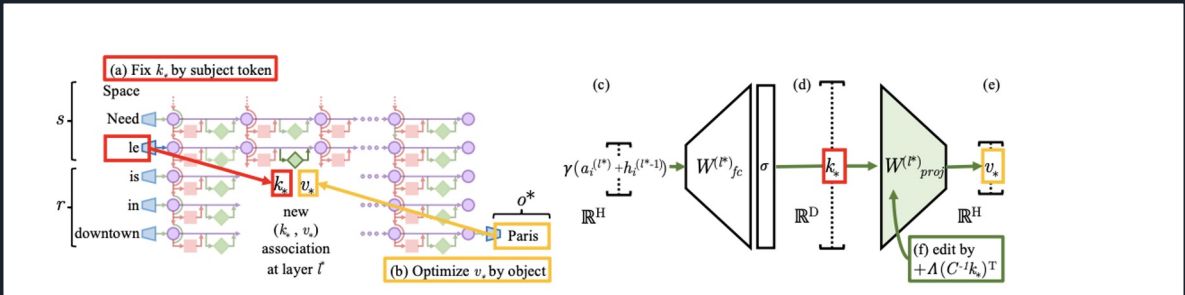
- ROME = Rank-One Model Editing :
  - A way to change a single fact inside a language model without retraining
  - Fast: takes ~2 seconds on one GPU

- MLP weight matrix W_proj acts as a linear associative memory
  - Keys = subjects (e.g. "Space Needle"), Values = facts about them (e.g. "Seattle")
  - The weight matrix W_proj maps keys to values: WK ≈ V

- ROME edits a fact by solving a constrained least squares problem

$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_* \quad \text{by setting } \hat{W} = W + \Lambda(C^{-1}k_*)^T$$

- Goal: write in a new (key, value) pair without breaking the other facts already stored
- C = KK^T is a pre-computed summary of all existing keys (from Wikipedia)



ROME: Three-Step Editing Process

- Step 1 – Select the subject key (k*)
  - Run the subject through the model with many different random prefixes
  - Average the MLP activations at the last subject token to get a stable key k*

$$k_* = \frac{1}{N}\sum_{j=1}^{N} k(x_j + s), \text{ where } k(x) = \sigma\left(W_{fc}^{(l^*)}\,\gamma(a_{[x],i}^{(l^*)} + h_{[x],i}^{(l^*-1)})\right)$$

- Step 2 – Optimize the new value (v*)
  - Optimize v* so the model predicts the new answer, while still knowing what the subject is

$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \middle\| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}.$$

- Step 3 – Insert via rank-one update
  - One rank-one update to W_proj at mid-layer
  - ~2 seconds, single GPU

Evaluation: zsRE Benchmark

- 10,000 factual edits tested on Efficacy, Paraphrase, Specificity

- ROME: 99.8% efficacy, 88.1% paraphrase, 24.2% specificity
  - Matches methods that need hours of extra training (KE, MEND)

- zsRE specificity test is too easy. Motivates 'better'/harder benchmark → COUNTERFACT

Table 1: zsRE Editing Results on GPT-2 XL.

Editor	Efficacy ↑	Paraphrase ↑	Specificity ↑
GPT-2 XL	22.2 (±0.5)	21.3 (±0.5)	24.2 (±0.5)
FT	99.6 (±0.1)	82.1 (±0.6)	23.2 (±0.5)
FT+L	92.3 (±0.4)	**47.2 (±0.7)**	23.4 (±0.5)
KE	65.5 (±0.6)	61.4 (±0.6)	24.9 (±0.5)
KE-zsRE	92.4 (±0.3)	90.0 (±0.3)	23.8 (±0.5)
MEND	75.9 (±0.5)	65.3 (±0.6)	24.1 (±0.5)
MEND-zsRE	99.4 (±0.1)	**99.3 (±0.1)**	24.1 (±0.5)
ROME	**99.8 (±0.0)**	88.1 (±0.5)	**24.2 (±0.5)**

COUNTERFACT: Evaluation Dataset

- 21,919 difficult counterfactual edits

- Five evaluations
  - Efficacy → Did the edit work?
  - Generalization → Does it work when rephrased?
  - Bleedover → Are unrelated subjects unchanged?
  - Consistency → Are generated texts coherent?
  - Fluency → Is the language still natural?

Table 2: COUNTERFACT Composition

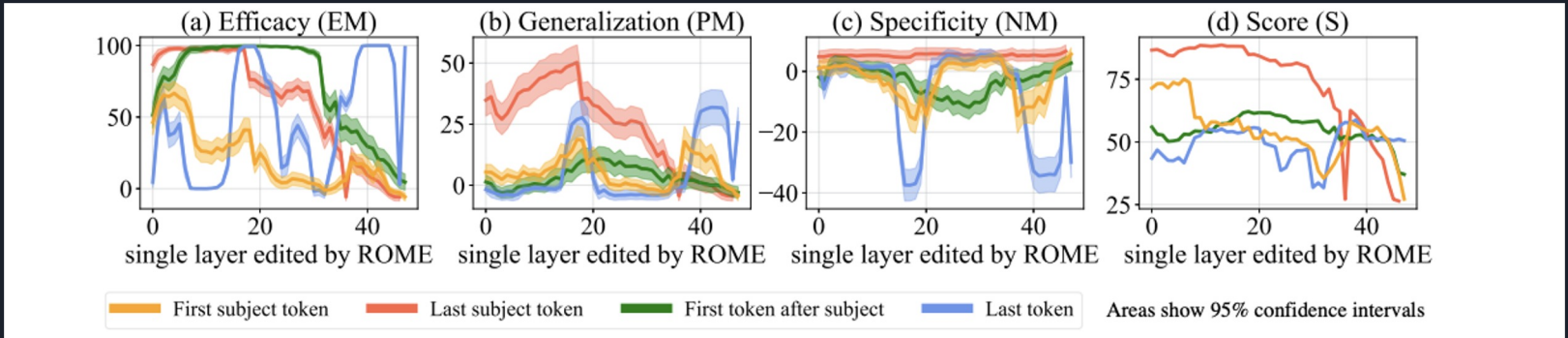
Item	Total	Per Relation	Per Record
Records	21919	645	1
Subjects	20391	624	1
Objects	749	60	1
Counterfactual Statements	21595	635	1
Paraphrase Prompts	42876	1262	2
Neighborhood Prompts	82650	2441	10
Generation Prompts	62346	1841	3

Table 3: Comparison to Existing Benchmarks

Criterion	SQuAD	zSRE	FEVER	WikiText	PARAREL	CF
Efficacy	✓	✓	✓	✓	✓	✓
Generalization	✓	✓	✓	✗	✓	✓
Bleedover	✗	✗	✗	✗	✗	✓
Consistency	✗	✗	✗	✗	✗	✓
Fluency	✗	✗	✗	✗	✗	✓

Causal Trace Findings

- ROME tested at every layer and token position
- Best edits: last subject token + mid-layers
  - Where causal tracing predicted facts are stored
- Editing attention instead of MLP
  - Attention reads facts, MLP stores them



COUNTERFACT Results: ROME vs. Baselines

- Two failure modes
  - Doesn't generalize to rephrasings (FT+L, KE, MEND)
  - Bleeds over to unrelated subjects (FT, KE-CF, MEND-CF)

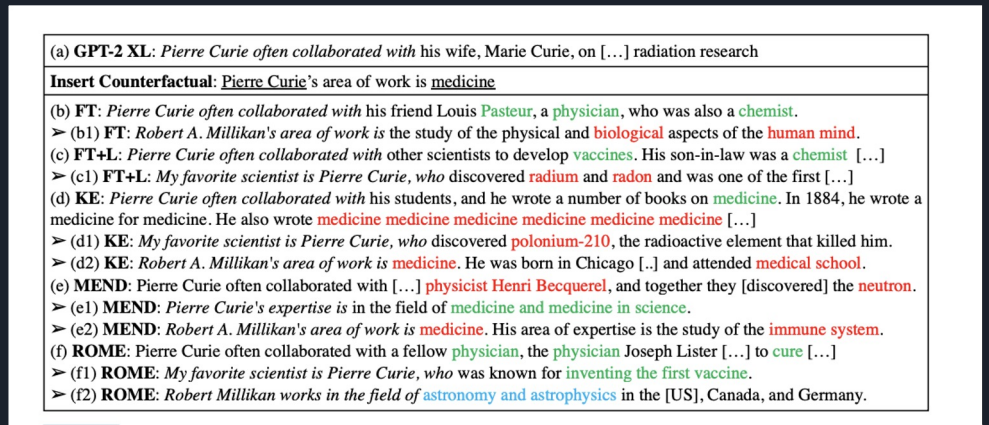
- ROME avoids both:
  - GPT-2 XL: Score 89.2 vs. next best 66.9
  - GPT-J: Score 91.5 vs. next best 68.7

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	607.1 (1.1)	40.5 (0.3)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)	**48.7 (1.0)**	28.9 (0.8)	70.3 (0.7)	3.5 (0.3)	621.4 (1.0)	37.4 (0.3)
KN	**35.6**	**28.7 (1.0)**	**-3.4 (0.3)**	**28.0 (0.9)**	**-3.3 (0.2)**	72.9 (0.7)	3.7 (0.2)	**570.4 (2.3)**	**30.3 (0.3)**
KE	52.2	84.3 (0.8)	33.9 (0.9)	75.4 (0.8)	14.6 (0.6)	**30.9 (0.7)**	**-11.0 (0.5)**	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**18.1**	99.9 (0.1)	97.0 (0.2)	95.8 (0.4)	59.2 (0.8)	**6.9 (0.3)**	**-63.2 (0.7)**	**383.0 (4.1)**	**24.5 (0.4)**
MEND	57.9	99.1 (0.2)	70.9 (0.8)	65.4 (0.9)	12.2 (0.6)	**37.9 (0.7)**	**-11.6 (0.5)**	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**14.9**	**100.0 (0.0)**	**99.2 (0.1)**	**97.0 (0.3)**	**65.6 (0.7)**	**5.5 (0.3)**	**-69.9 (0.6)**	**570.0 (2.1)**	33.2 (0.3)
ROME	**89.2**	100.0 (0.1)	97.9 (0.2)	96.4 (0.3)	62.7 (0.8)	**75.4 (0.7)**	**4.2 (0.2)**	621.9 (0.5)	**41.9 (0.3)**
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)	18.6 (1.5)	-7.4 (0.6)	83.0 (1.1)	7.3 (0.5)	621.8 (0.6)	29.8 (0.5)
FT	**25.5**	**100.0 (0.0)**	**99.9 (0.0)**	96.6 (0.6)	71.0 (1.5)	**10.3 (0.8)**	**-50.7 (1.3)**	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	68.7	99.6 (0.3)	95.0 (0.6)	**47.9 (1.9)**	30.4 (1.5)	78.6 (1.2)	**6.8 (0.5)**	**622.8 (0.6)**	35.5 (0.5)
MEND	63.2	97.4 (0.7)	71.5 (1.6)	**53.6 (1.9)**	11.0 (1.3)	53.9 (1.4)	**-6.0 (0.9)**	620.5 (0.7)	32.6 (0.5)
ROME	**91.5**	99.9 (0.1)	99.4 (0.3)	**99.1 (0.3)**	**74.1 (1.3)**	**78.9 (1.2)**	5.2 (0.5)	620.1 (0.9)	**43.0 (0.6)**

Generation Results Example

- Counterfactual: "Pierre Curie's area of work is medicine"
- What methods do:
  - FT: Generalizes, but also changes Millikan (bleedover)
  - FT+L: Inconsistent, sometimes still says "physics"
  - KE: Repeats "medicine" on loop (broken output)
  - MEND: Fails on both, poor generalization + bleedover

- What ROME does:
  - Curie consistently described as physician across all prompts
  - Millikan remains completely unchanged



(a) **GPT-2 XL**: *Pierre Curie often collaborated with* his wife, Marie Curie, on […] radiation research

Insert Counterfactual: Pierre Curie's area of work is medicine

(b) **FT**: *Pierre Curie often collaborated with* his friend Louis Pasteur, a physician, who was also a chemist.
➢ (b1) **FT**: *Robert A. Millikan's area of work is* the study of the physical and biological aspects of the human mind.
(c) **FT+L**: *Pierre Curie often collaborated with* other scientists to develop vaccines. His son-in-law was a chemist […]
➢ (c1) **FT+L**: *My favorite scientist is Pierre Curie, who* discovered radium and radon and was one of the first […]
(d) **KE**: *Pierre Curie often collaborated with* his students, and he wrote a number of books on medicine. In 1884, he wrote a medicine for medicine. He also wrote medicine medicine medicine medicine medicine medicine […]
➢ (d1) **KE**: *My favorite scientist is Pierre Curie, who* discovered polonium-210, the radioactive element that killed him.
➢ (d2) **KE**: *Robert A. Millikan's area of work is* medicine. He was born in Chicago [..] and attended medical school.
(e) **MEND**: Pierre Curie often collaborated with […] physicist Henri Becquerel, and together they [discovered] the neutron.
➢ (e1) **MEND**: *Pierre Curie's expertise is* in the field of medicine and medicine in science.
➢ (e2) **MEND**: *Robert A. Millikan's area of work is* medicine. His area of expertise is the study of the immune system.
(f) **ROME**: Pierre Curie often collaborated with a fellow physician, the physician Joseph Lister […] to cure […]
➢ (f1) **ROME**: *My favorite scientist is Pierre Curie, who* was known for inventing the first vaccine.
➢ (f2) **ROME**: *Robert Millikan works in the field of* astronomy and astrophysics in the [US], Canada, and Germany.

Limitations

- Single-fact editing only
  - ROME edits one fact per operation
- Edits are one-directional
  - "The Space Needle is in Seattle" is stored separately from "The iconic landmark in Seattle is the Space Needle"
- Narrow scope of knowledge
  - Only factual associations ("born in", "works in") were studied
  - Nothing about logical or spatial

Future Directions

- Editing both directions of a fact simultaneously
  - Method that propagates a single edit to all related association directions
- Applying causal tracing to other knowledge types
  - Extending tracing methods
- Reducing hallucination after edits
  - Detection tools that flag when model weights have been edited

Takeaway/Summary

- “Language Models as Knowledge Bases?” (Petroni et al., 2019)
  - Studied the knowledge present in off the shelf LLMs, found that they (especially BERT) perform well against other knowledge base forms
- “Transformer Feed-Forward Layers Are Key-Value Memories” (Geva et al., 2021)
  - Shines light on the feed forward layers, showing that they store information, and that they can recognize various patterns
- “Locating and Editing Factual Associations in GPT” (Meng et al., 2023)
  - Detailed analysis of feed forward layers to find that they store facts about subjects, and introduction of a method to edit said facts, as well as benchmarks on the fact editing method

Q&A