NLP&AI 연구실 세미나 (26.01.15.Thu)

LLM Lens w/ SAE: 두 번째

김진성📢

Natural Language
Processing
& Artificial Intelligence

고려대학교
KOREA UNIVERSITY

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder

Yi Jing[♠], Zijun Yao[♠], Hongzhu Guo[♡], Lingxu Ran[♠]
Xiaozhi Wang[◇], Lei Hou[♠], Juanzi Li[♠§]
[♠]DCST, BNRist; KIRC, Institute for Artificial Intelligence;
[◇]Shenzhen International Graduate School, Tsinghua University, *China*
[♡]Department of Chinese Language and Literature, Peking University, *China*
jingy22@mails.tsinghua.edu.cn, lijuanzi@tsinghua.edu.cn

RECAP

Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models

Javier Ferrando[1,2*] **Oscar Obeso**[3*] **Senthooran Rajamanoharan** **Neel Nanda**

[1]U. Politècnica de Catalunya [2]Barcelona Supercomputing Center [3]ETH Zürich

Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy,
Lewis Smith, Nicolas Sonnerat, Vikrant Varma,
János Kramár, Anca Dragan, Rohin Shah, Neel Nanda
Google DeepMind
tlieberum@google.com

Motivation

* Knowledge, Hallucination, and Entity

- “지식을 아느냐 모르느냐의 여부”---Hallucination 과의 관계 탐구

- Sparse Autoencoders (SAEs) 를 해석 도구로써 사용하여, **Hallucination 발생의 메커니즘** 검증

  → LLM이 언제 환각을 일으키는지, 또는 답변을 거부하는지에 대한 이해

* 여기서의 Hallucination 정의:
  = **모델이 가지지 않은 정보**를 생성하도록 요구받을 때, (거부하지 않고) 함부로 생성하는 것

- 전제: ‘Entity = 지식의 요체’로 간주.

  → entity 에 대한 지식 여부와 hallucination 발생 기제의 관계

RECAP

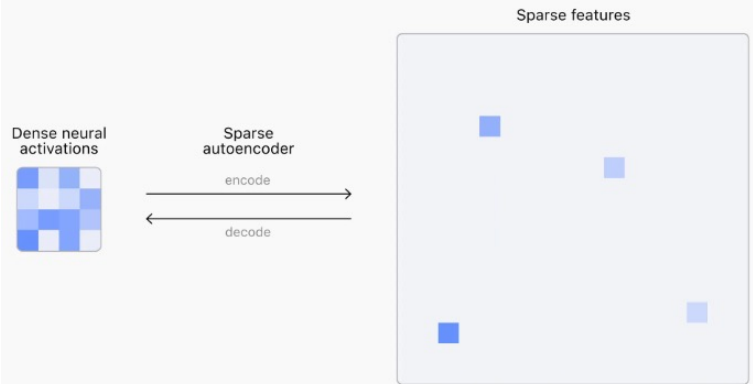
Preliminaries

Gemma Scope: Open Sparse Autoencoders Everywhere All At Once On Gemma 2 (BlackboxNLP 2024)

* Sparse Autoencoder (SAE)

- 이미 모델 내부에 존재하는 representation 을 분석하는 도구.

- 모델의 representation 을 sparse 하고 해석 가능한 features (latents) 의 조합으로 분해 및 재구성



→ ⭐ 즉, 하나의 latent 가 크게 활성화되면, 대부분의 다른 latent 는 활성화되지 않거나 (대부분), 0 에 매우 가까운 값을 가지도록 학습됨.

Preliminaries

* Sparse Autoencoder (SAE) – Formalization in Gemma Scope

- x: 원본 모델의 (token 에 대한) 입력 표현 (임베딩, dense vector)
- a(x): SAE 인코더를 통과하여 생성된 sparse latent activations (= feature activations) vector
  → 해당 벡터의 각 요소: SAE가 학습한 특정 **feature 가 현재 입력 x 에서 얼마나 활성화 되었는지**를 나타냄.

가령, When was the player LeBron James born? 에서 "James" 의 모델 표현 벡터 x

x: [x.x, x.x, x.x, ⋯, x.x] (1*d, dense vector) → a(x): [**0.8**, **3.3**, 0, 0, 0, 0, ..., 0] (1*dSAE, sparse vector)

- W_{dec}: SAE decoder 의 weight matrix
  → 희소하고 해석 가능한 형태로 분해된 표현인 a(x)를 원래 모델 차원과 일치하는 벡터로 재조립
  (1*dSAE 차원 행벡터 → 1*d 차원으로 재구성)

$$\text{SAE}(\boldsymbol{x}) = a(\boldsymbol{x})\boldsymbol{W}_{\text{dec}} + \boldsymbol{b}_{\text{dec}}, \qquad a(\boldsymbol{x}) = \text{JumpReLU}_{\theta}\left(\boldsymbol{x}\boldsymbol{W}_{\text{enc}} + \boldsymbol{b}_{\text{enc}}\right),$$

RECAP

Preliminaries

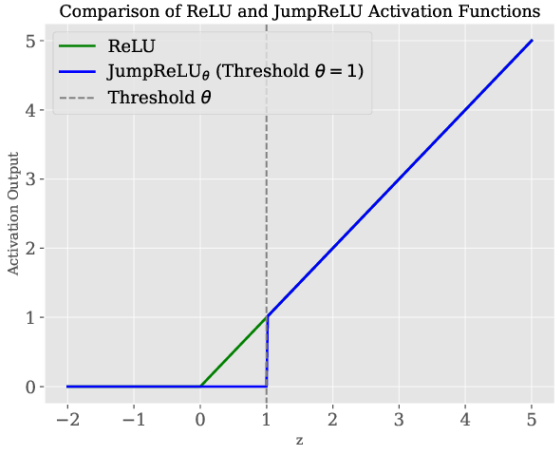
Gemma Scope: Open Sparse Autoencoders Everywhere
All At Once On Gemma 2 (BlackboxNLP 2024)

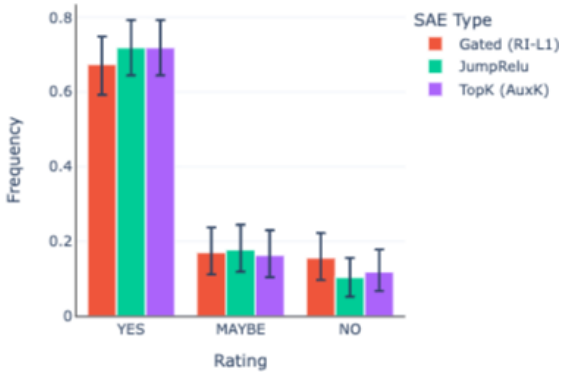
* Sparse Autoencoder (SAE) – Formalization 상세

- a(x): sparse vector 형태, 그 속의 각 element 가 곧 하나의 “SAE latent”.
  → 즉, a(x)의 j-th 원소는
  입력 x 에 대하여 j-th latent 가 얼마나 강하게 활성화되었는지를 알려주는 스칼라 값.

⁎ 모든 elements 의 sum 이 1이 아님.
확률값이 아니고, 얼마나 j-th feature 가 현재 입력 x에서 강하게 활성화되었는지를 나타내는 값이기 때문.

⁎ JumpReLU(·): sparsity 를 유도하여,
일부 특징만 활성화되도록 하는 역할





RECAP

Preliminaries

Gemma Scope: Open Sparse Autoencoders Everywhere
All At Once On Gemma 2 (BlackboxNLP 2024)

* Sparse Autoencoder (SAE) – Formalization 상세

- W_{dec}: dense vector 를 모아놓은 matrix 인데,
  각 column vector은 몽가 semantic contents 을 담고 있음.
  → a(x) 에서 활성화된 latents 와 상응하는 W_{dec} 내의 column vector 와 곱하면,
  SAE(x)는 활성화된 semantics 만 살아남게 되는 원리.
  → 따라서, 잘 학습 시켜놓은 SAE 일수록 W_{dec} 를 양질로 살려놓음..

- SAE 학습 loss 간략히:

$$\mathcal{L}(x) = \underbrace{\|x - \mathrm{SAE}(x)\|_2^2}_{\mathcal{L}_{\text{reconstruction}}} + \underbrace{\lambda \|a(x)\|_0}_{\mathcal{L}_{\text{sparsity}}}.$$

0이 아닌 값이 많을 수록 penalty 부여

원본 표현 x와 재구성 표현 SAE(x) 간 차이 최소화

→ ⭐ 분석 타겟 모델 (Gemma2) 의 **layer 개수만큼 개별적인 훈련된 SAE 가 하나씩 존재.**
(Gemma 2-2B: 26개 / 9B: 42개의 SAEs)

RECAP

Technical Purpose

*** SAEs 를 가져다가 어떻게 활용할까? (1)**

- SAEs 를 **모델 내부 표현에 대한 일종의 Lens** 로 사용하여,
  모델의 latent representation space 로부터 어떤 **방향성 (directions)**를 발견 해버리기.

- known or unknown entities, 즉 entity 에 대한 지식 여부에 따른 **latent activation 패턴** 관찰.

Known Entity Latent Activations	Unknown Entity Latent Activations
Michael Jordan	Michael Joordan
When was the player LeBron James born?	When was the player Wilson Brown born?
He was born in the city of San Francisco	He was born in the city of Anthon
I just watched the movie 12 Angry Men	I just watched the movie 20 Angry Men
The Beatles song 'Yellow Submarine'	The Beatles song 'Turquoise Submarine'

= non-highlighted words 는 SAE latent activation 값이 (매우 낮거나) 0점
highlighted words 는 latent activation 이 높은 활성화 지점

Technical Purpose

*** SAEs 를 가져다가 어떻게 활용할까? (2)**

- 가설:
  1) latent activation patterns 를 보면,
     hallucination 이 일어날지 여부를 어느정도 궁예할 수 있을 것이다.

  2) 그리고, 이 activation patterns 를 조금 manipulate 해주면,
     hallucination 생성 혹은 refusal 을 어느정도 steering 할 수 있을 것이다.

Method

*** Entity set 추출 및 구성**

- 실험에 사용되는 모든 entities 는 Wikidata 로부터 추출
  (Wikidata 는 entity에 관한 attributes 가 metadata 로 달려있음)

- 4가지의 entity type 사용
  : entity type 에 따라 영향을 받는지 검증

Entity Type	Attributes	몇 개?
Player (농구 선수)	출생지, 생년월일, 소속 팀	7,487
Movie	감독, 각본가, 개봉일, 장르, 상영 시간, 출연진	10,895
City	국가, 인구, 고도, 좌표	7,904
Song	아티스트, 앨범, 발매 연도, 장르	8,448

Method

* Known vs. Unknown Entities Clustering

- 모델의 사전 지식 검증 기반 분류

- (entity type, entity name, relation attribute) 형식의 템플릿을 기반으로 자연어 질문으로 재생성

e.g.,

- relation attribute: 해당 entity 에 대해 설명해주는 속성들

→ Known: 모델이 두 가지 속성 이상을 맞춘 경우.

Unknown: 모델이 속성을 틀린 경우.
(나머지 경우는 배제)

Experiment

* Setup

- 데이터셋: Wikidata 에서 긁은 entity set & entity triple 기반으로 생성한 검증 질의들 (자가 구축)
- SAE 모델: Gemma Scope, Llama Scope 연구에서 제공하는 pre-trained SAEs 활용
- 내부 표현 검증 대상 모델: Gemma 2, Llama3.1

※ SAE 는 학습된 같은 모델에 대해서만 연산 가능

Experiment

* Results (1): 자기 지식 인식의 존재

- Self Knowledge Awareness
  : 특정 지식에 대해 자신이 그것을 안다 모른다를 식별하는 능력.
  = 일종의 메타인지 능력

- LLM은 자기 지식 인식 능력을 가지고 있음.

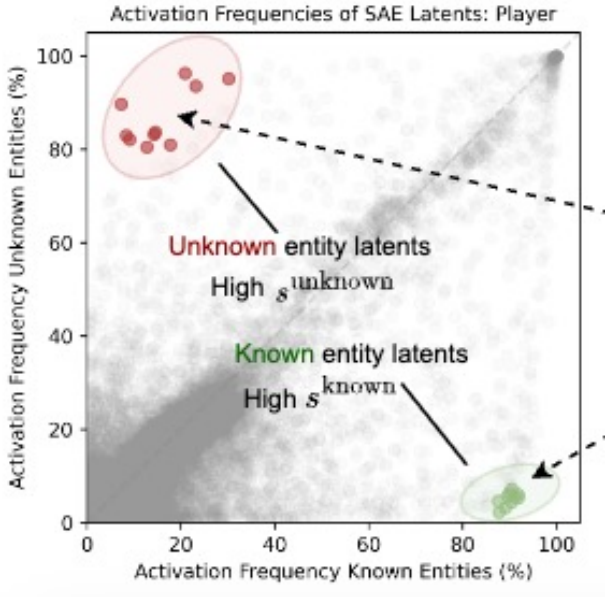
→ 즉, 모델은 지식 (entity)을 단순히 기억하는 것을 넘어, 그것을 아는지 모르는지에 대해 스스로 인지하는 내부 표현/메커니즘을 가질 수 있다!

RECAP

Experiment

*** Results (2): 자기 지식 인식 - known 과 unknown 상호 관계**

- self knowledge awareness 는 어떤 activation patterns 로 나타나는가?



Knowns 와 Unknowns 의 activation frequency

→ 반비례: known 에 대해 activated 되는 latents 는, unknown 은 에 대해서는 비활성화.

→ latent 마다 활성화되는 시기가 다르다.

→ 즉, knowns 를 감지하는데 특화된 latents 가 따로 있고, unknowns 에 특화된 latents 따로 있다.

Experiment

*** Results (3): 자기 지식 인식과 Hallucination 의 관계**

- 모델 자신이 특정 entity 를 “모른다는 사실”을 “알고 있으면”? → Refusal 응답 생성.
  But, “모른다는 사실”을 “모르고 있으면”? → Hallucination 발생!

→ 그렇다면,
원래 자신이 “모르는” entity 라는 것을 “모르고 있었던” instance 에 대해서,
“모르는” 것을 “알고 있다”는 self knowledge awareness 를 심어주면,
원래는 hallucination 일어날 것을 refusal 하도록 만들 수 있다.

→ 반대로 self knowledge awareness 를 방해하면,
원래 refusal 할 LLM의 행동을 헛소리 하도록 유도할 수 있다.

→ 즉, given unknown entity 에 대한 SAE latent activations 를 조작하여 steering 가능하다.

Experiment

*** Results (4): (자기 지식 인식) 조작이 어떻게 가능한가?**

Q) SAE activation 경향을 찾았음. 그럼 진짜 LLM의 생성을 steering 할 수 있나? 어떻게?

→ 앞서 SAE 정의 수식에서 a(x) (x 에 대한 SAE encoder 의 activation 값 vector)에 곱해지는 W_{dec} 의 일부 행 (vector)의 값을 의도적으로 조절하여 모델의 행동을 바꾼다!

조작 강도 (추가)

$$\boldsymbol{x}^{\text{new}} \leftarrow \boldsymbol{x} + \alpha \boldsymbol{d}_j.$$

조작된 모델 표현 조작 방향 (선택)

- d_j : SAE decoder 의 weight matrix 의 j-th 행 vector
  a : val. set 별도 실험으로 empirically 설정 → [400, 550]

Experiment

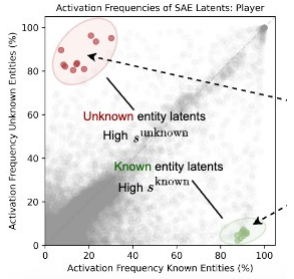
*** Results (4): (자기 지식 인식) 조작이 어떻게 가능한가?**

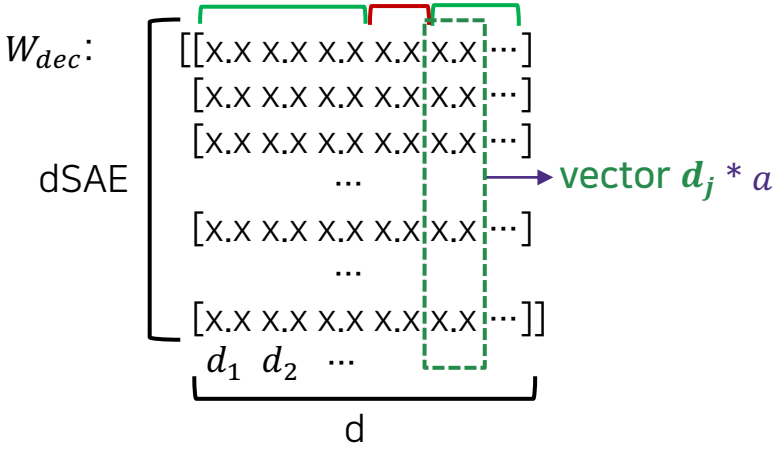
- 쉽게 표현해서, 가령, When was the player Wilson Brown born? 에서

Brown 의 Original a(x): [0, 0, **0**, 0, **0.8**, **0.2**, ⋯, 0] (1*dSAE) 이라면,

j-th latent

known 활성화 latent group unknown 활성화 latent group





→ knowns 의 latent group 에 해당하는 d 들 중에, 가장 known/unknown 을 잘 구분하고, entity type 에 관계없이 일반화가 잘 되는 SAE latent index 를 d_j로 설정.

→ 찾은 d_j 에 조향 계수 a 를 곱한 값으로 x_{new} 의 값 뻥튀기. (단일 d 사용)

RECAP

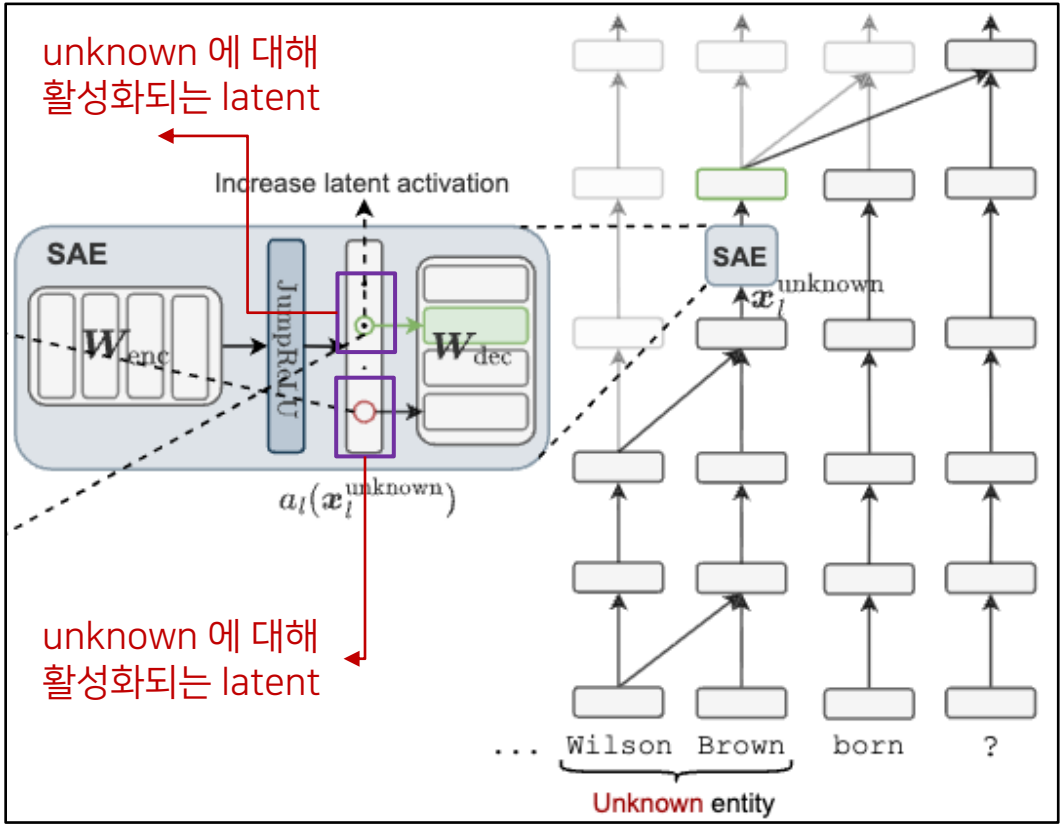
Experiment

DO I KNOW THIS ENTITY? KNOWLEDGE AWARENESS AND HALLUCINATIONS IN LANGUAGE MODELS (ICLR 2025)

*** Results (4): 조작이 어떻게 가능한가?**

1. 'Wilson Brown' 이라는 unknown entity 의 마지막 토큰 표현 $x^{unknown}$ 이 SAE 인코더 W_{enc} 로 입력
2. SAE 는 sparse latent activation $a(x^{unknown})$ 의 집합으로 변환
3. known entities 에 반응하는 latent activation 중 가장 separation score 높은 latent 에 상응하는 W_{dec} 의 행 d의 값을 의도적으로 증가시킴.
4. representation x 를 x_{new} 로 재구성 (모델 내부 상태 변경)

= 즉, 실제로는 unknown entity 임에도, 모델에 "이 entity 를 알고 있다" 는 신호를 강제로 주입하는 것.



RECAP

Experiment

*** Results (4): (자기 지식 인식) 조작이 어떻게 가능한가?**

cf) Latent Separation Score 구하기?

→ 가장 known vs. unknown 구분 잘 하고, entity type 에 일반화 잘 되는 j-th latent 를 찾아보자.

j-th SAE latent

a(x): [0, 0, **0**, 0, **0.8**, **0.2**, ⋯, 0] (1*dSAE)

1. 활성화 빈도 계산:
   known , unknown 에 대해 활성화 되는 각 빈도

$$f_{l,j}^{\text{known}} = \frac{\sum_i^{N^{\text{known}}} \mathbb{1}[a_{l,j}(\boldsymbol{x}_{l,i}^{\text{known}}) > 0]}{N^{\text{known}}}, \quad f_{l,j}^{\text{unknown}} = \frac{\sum_i^{N^{\text{unknown}}} \mathbb{1}[a_{l,j}(\boldsymbol{x}_{l,i}^{\text{unknown}}) > 0]}{N^{\text{unknown}}},$$

→ i-th entity 문장에 대해, LLM 의 l-th layer 에서 얻은 표현 x_l을 입력으로 받았을 때,
(i-th 담당) SAE 의 latent j 가 활성화 되는지 안 되는지 여부, 즉 activation이 (>0) 인 latent 개수 count

- N (known, unknown): 앞서 구축한 모든 entity set 에 대한 프롬프트들 (entity 포함 자연어 문장)의 개수
  e.g, *"When was the player {entity_name} born?"*

RECAP

Experiment

*** Results (4): (자기 지식 인식) 조작이 어떻게 가능한가?**

cf) Latent Separation Score 구하기?

2. 빈도 기반으로 개별 latent separation score 계산

: $s_{l,j}^{known} = f_{l,j}^{known} - f_{l,j}^{unknown}$, $s_{l,j}^{unknown} = f_{l,j}^{unknown} - f_{l,j}^{known}$

3. 각 latent 를 모든 entity type 에 대해서 분리 점수 계산하여 (type 간) 최소값 산출: $\min_{t} s_{l,j}^{(un)known,t}$

: why 최소값? 최악의 경우에도 latent activation 이 얼마나 잘 작동하는가 평가 위함.
(= 구분이 가장 어려운 type 에서도 꽤 잘 한다.)

4. 모든 레이어 l 의 latent j 중에서 $\min s$ 값이 가장 높은 $\text{latent}_{l,j}$ k개 선택. (최소값 중 최대값)

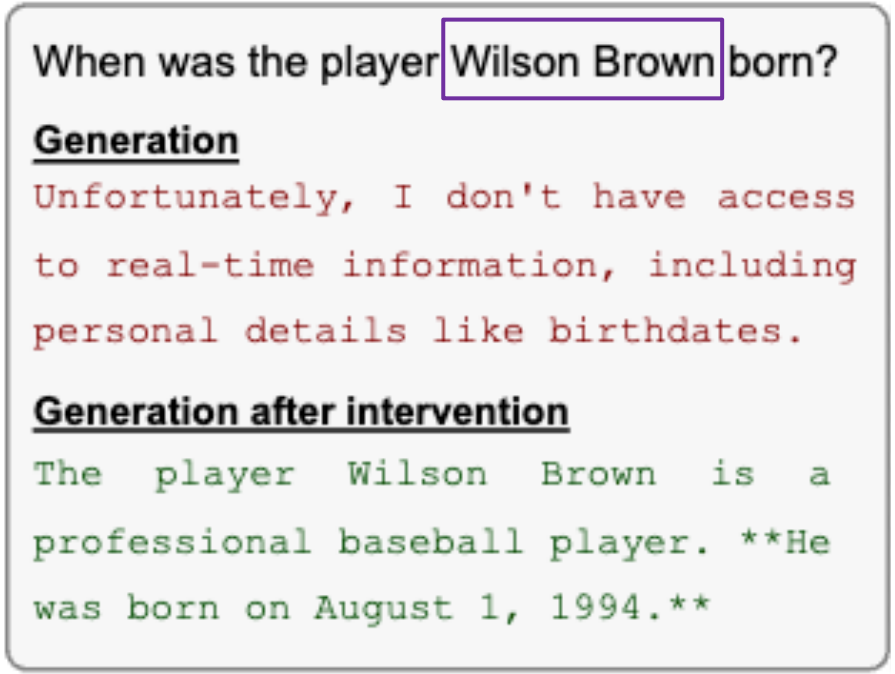
→ k는 topmost, top-3 등..

+) 무작위 토큰에서도 자주 활성화되는 noisy latent 는 필터링.

Experiment

* Results (5): 자기 지식 인식 조작 (Steering)의 결과

- 조작된 latent activations 를 통해, unknown entity 에 대한 환각 유도 성공 (refusal → hallucinations)



Motivation

*** 문제 의식**

1. LLM이 Language knowledge 를 process and represent 하는 internal mechanism 의 불분명성

2. 기존 behavior-based interpretation approach 는 해석성에 한계 있음
 - 예) expert-designed prompt engineering

3. 기존 mechanistic interpretation approaches 의 coarse granularity
 - 가령, neuron study 등은 polysemantic 하다. (즉, 여러 개 features 를 함의 가능성)

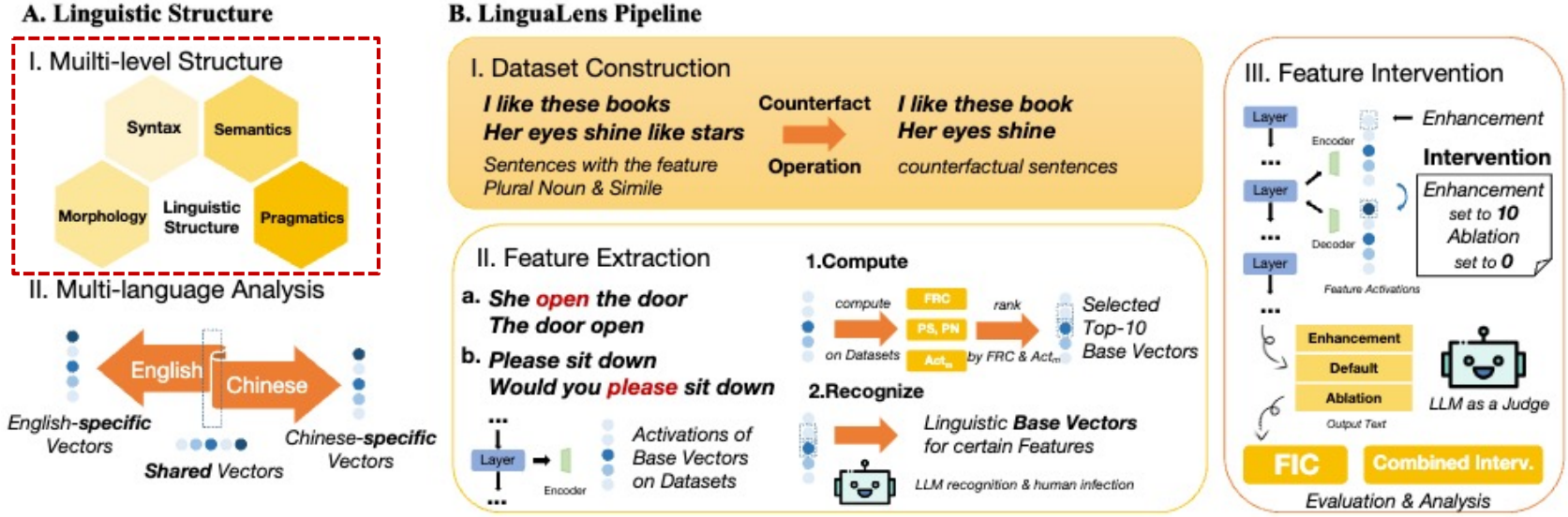
→ 특히, 3번이 SAE 사용하는 큰 이유
 : SAE 는 LLM 내부의 hidden states 를 sparse feature space 로 분해함으로써,
 각 차원이 **"단일한" 의미 개념을 포착**할 수 있게 함 (**monosemantic** & fine-grained).

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: Intuition**

- 전통적인 언어학에서는 **언어를 여러 계층으로 나누어 분석**하는 것에서 영감을 받음.
- 1. Morphology (형태론), 2. Syntax (통사론), 3. Semantics (의미론), 4. Pragmatics (화용론)
  → 네 가지 linguistic levels 에서 영어(중국어)의 Linguistic features 를 추출.



Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: taxonomy setup**

- 네 가지 levels 에서 총 145개의 Linguistic features 개념 taxonomy setup

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* 프레임워크: taxo

- 네 가지 levels 에

F.2 Linguistic Feature List

past_tense Morphology & Semantics — verb form that locates an event before speech time.

noun_plural Morphology — form marking more than one noun referent.

agentive_suffix Morphology — suffix creating nouns for the doer of an action.

negation_prefix Morphology — prefix that reverses or denies the base meaning.

degree_prefix Morphology — prefix intensifying or scaling the base concept.

temporal_prefix Morphology — prefix adding time relations such as "pre-" or "post-".

quantitative_prefix Morphology — prefix conveying amount or number.

spatial_or_directional_prefix Morphology — prefix indicating place or direction.

nominal_suffix Morphology — suffix that turns a base into a noun.

verbal_suffix Morphology — suffix that turns a base into a verb.

adjectival_suffix Morphology — suffix that turns a base into an adjective.

adverbial_suffix Morphology — suffix that turns a base into an adverb.

possessive_form Morphology & Syntax — morphological marking of ownership or relation.

third_person_singular Morphology & Syntax — verb agreement form for he/she/it.

past_participle Morphology & Syntax — verb form used in perfect aspect or passive voice.

present_participle Morphology & Syntax — "-ing" form used for progressives or gerunds.

comparative Morphology & Semantics — form showing a higher degree of a property.

superlative Morphology & Semantics — form showing the highest degree of a property.

past_tense_irregular Morphology — past form that does not end in "-ed".

past_participle_irregular Morphology — irregular past participle form.

intransitive_verb Syntax — verb that takes no direct object.

transitive_verb Syntax — verb that requires a direct object.

linking_verb Syntax — verb that links subject to a complement.

anaphor Syntax & Pragmatics — expression that refers back to an antecedent.

subject_auxiliary_inversion Syntax — swapping subject and auxiliary (e.g., questions).

subject_verb_inversion Syntax — reversing subject and main verb order.

passive_voice Syntax & Semantics — clause where patient becomes grammatical subject.

subjunctive_mood Syntax & Semantics — form expressing wish, doubt, or hypothetical state.

first_conditional Syntax & Semantics — "if + present, will + verb" for real future possibility.

indirect_speech Syntax & Pragmatics — reporting speech without a direct quote.

elliptical_sentences Syntax — sentences with understood but omitted elements.

cleft_sentences Syntax — "it + be + focus" construction for emphasis.

appositives Syntax — noun phrase renaming another noun phrase.

non_defining_relative_clauses Syntax — extra, non-restrictive relative clauses.

emphatic_structure Syntax & Pragmatics — construction that highlights or stresses a clause part.

noun_clauses Syntax — subordinate clauses functioning as nouns.

relative_clauses Syntax — clauses that modify a noun with a relative word.

imperative_sentence Syntax & Pragmatics — clause issuing a command or request.

of_genitive Syntax — possession expressed with an "of" phrase.

s_genitive Syntax — possession marked with apostrophe-s.

clausal_subjects Syntax — clauses acting as the subject of a sentence.

extraposition Syntax — moving a heavy subject/object to clause end with dummy "it".

copular_be Syntax — "be" used as a linking verb, not as an auxiliary.

echo_questions Syntax & Pragmatics — repetition of prior utterance to seek confirmation.

tag_questions Syntax & Pragmatics — short question tags appended to statements.

direct_object Syntax — noun phrase receiving the verb's action.

universal_quantifiers Syntax & Semantics — words like "all, every" signifying totality.

existential_quantifiers Syntax & Semantics — words like "some, any" signifying existence.

expletive Syntax — syntactic placeholder such as "it" or "there".

factives Semantics & Syntax — predicates presupposing truth of their complement.

futurates Semantics & Syntax — present-tense forms referring to scheduled future events.

intensifiers Semantics & Pragmatics — adverbs that strengthen degree (e.g., "very").

mass_noun Syntax & Semantics — noun for uncountable substances (e.g., "water").

object_expletives Syntax — expletive pronouns occupying object position.

nominal_adverbials Syntax — noun phrases functioning like adverbs.

split_infinitives Syntax — placing a word between "to" and the verb stem.

quantifier Syntax & Semantics — word or phrase expressing quantity.

count_nouns Syntax & Semantics — nouns that can be enumerated individually.

active_verbs Syntax — verbs used in active voice constructions.

middle_verb Syntax & Semantics — verb whose subject is patient but appears active.

referring Semantics & Pragmatics — linguistic act of pointing to real-world entities.

static_dynamic Semantics — distinction between state verbs and action verbs.

punctual_durative Semantics — contrast between instantaneous and durational events.

telic_atelic Semantics — events with inherent endpoints vs. those without.

past Semantics — temporal reference before the present moment.

future Semantics — temporal reference after the present moment.

present_progressive Semantics — aspect for ongoing present actions.

present_perfect Semantics — aspect connecting past event to present state.

past_progressive Semantics — aspect for ongoing past actions.

past_perfect Semantics — event completed before a past reference point.

future_progressive Semantics — ongoing action projected into the future.

future_perfect Semantics — event completed before a future reference point.

epistemic Semantics & Pragmatics — modality expressing speaker's judgment of likelihood.

deontic Semantics & Pragmatics — modality expressing obligation or permission.

spatial Semantics — meaning elements relating to location or space.

person Semantics & Pragmatics — grammatical category distinguishing speaker, addressee, others.

temporal Semantics — meaning elements relating to time relations.

given_known Pragmatics & Semantics — information already shared by speaker and listener.

representative Pragmatics — speech act conveying assertions or descriptions.

directive Pragmatics — speech act intended to get the hearer to act.

commisive Pragmatics — speech act committing speaker to future action.

expressive Pragmatics — speech act revealing speaker's feelings or attitude.

declaration Pragmatics — speech act that changes social reality.

metaphor Semantics & Pragmatics — figurative transfer of meaning based on similarity.

synecdoche Semantics & Pragmatics — figure where part stands for whole or vice versa.

non_synecdoche_metonymy Semantics & Pragmatics — metonymic shift based on association, not part-whole.

coordination Syntax & Semantics — joining of equal grammatical elements.

transitional Semantics & Pragmatics — discourse element marking a shift or progression.

resultative Syntax & Semantics — construction expressing a resultant state of an action.

optative Syntax & Pragmatics — form expressing a wish or hope.

existential Semantics & Syntax — clause asserting existence of something.

interrogative Syntax & Pragmatics — clause type used for asking questions.

deixis Pragmatics & Semantics — reference that depends on context (e.g., "here", "now").

turn_taking Pragmatics — conversational management of who speaks when.

euphemism Pragmatics & Semantics — mild term replacing a harsher one.

personification Semantics & Pragmatics — giving human traits to non-human entities.

hyperbole Semantics & Pragmatics — deliberate exaggeration for effect.

discourse_markers Pragmatics — words that organize or signal discourse flow.

politeness Pragmatics — linguistic strategies that mitigate imposition or face threat.

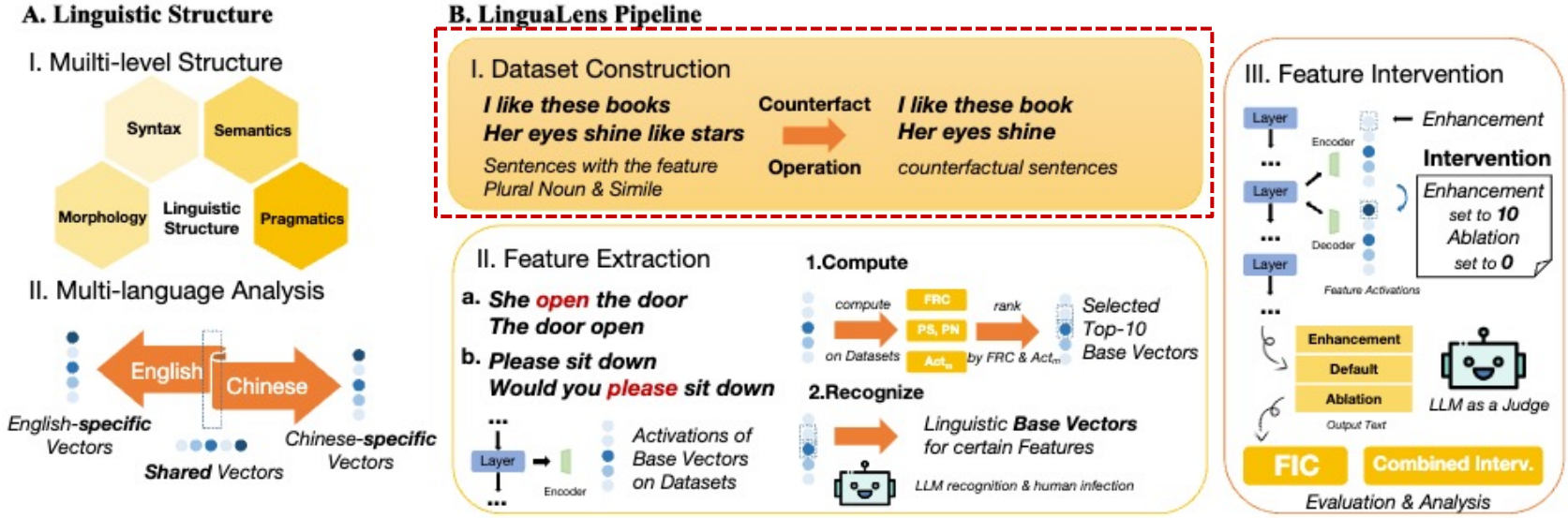
性_抽象名词后缀 形态学— 后缀"-性" 构成表示"-ness/-ity" 的抽象名词。

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* 프레임워크: counterfactual dataset construction (1)

- 145개의 linguistic features 의 각 feature 당 50개씩 포함하는 문장 s+ 와 해당 문장의 perturbated counterpart 문장인 s- 를 하나의 최소 분석 pair 로 구축 (s+, s-)



Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: counterfactual dataset construction (2)**

- 특정 언어 특징을 포함하는 **Positive Sentence (s+)** 를 기반으로 **Counterfactual Sentence (s-)** 를 다음과 같은 원칙으로 perturbation 수행

- 구축 원칙
1. Target linguistic feature 변경
2. 의미론적 내용 보존: 동일한 사건의 전달, 주체 및 대상 간 관계 유지
3. Minimal edit: 특정 linguistic feature 만 고립, 최소한의 단어 변경 (동사 형태, 전치사 구 등)

- 유의할 점:
. "counterfactual"이라는 용어는 우리가 평소에 이해하는 counter-factual (실제로 일어나지 않을 일) 아님.
. 위 세 가지 구축 원칙에 따라 원본 문장 s+ 를 perturbation 했다는 의미.
. s+와 s- 간에는 목표 feature 만 다르고, 다른 모든 의미론적/구문론적 요소는 최대한 동일하게 유지

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: counterfactual dataset construction (3)**

- 예시 #1: 수동태 → non-수동태
 . Positive Sentence (s+): The apple **is eaten** by the boy. (사과는 소년에 의해 먹혔다.)
 . Counterfactual Sentence (s-): The boy **eats** the apple. (소년은 사과를 먹는다.)
  → minimal edit 원칙에 따라, ("is eaten", "eats") replace

- 예제 #2: 복수 명사 → non-복수 명사
 . Positive Sentence (s+): I like these **books**. - 복수 명사
 . Counterfactual Sentence (s-): I like these **book**.
  → this 로 안 바꾼 이유? minimal edit 원칙.

→ 이러한 perturbation 기반 (s+, s-) pair 데이터셋의 필요성?
 : 특정 **언어 현상 (feature)의 유무에 따른 모델 내부 활성화 변화** (patterns)를 관찰하여 인과 관계 파악.
 : 특정 언어적 요인만 변경하여, **대조군 (control condition)**을 만드는 과정

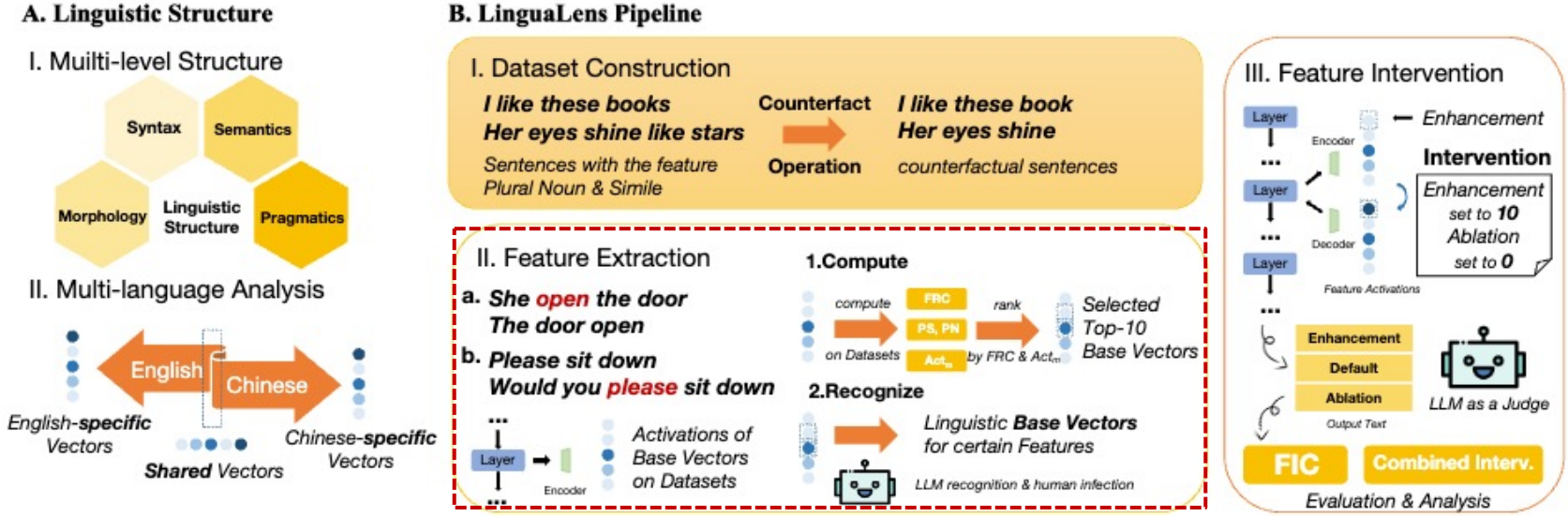
Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – two step-based investigation**

1. Linguistic feature 별 담당하는 SAE base vectors selection
   (가령, "수동태"라는 언어 현상을 나타내는 SAE 의미 부분은 어디인가 찾아내는 작업)

→ 2. 찾아낸 base vectors 의 activations 기반으로, layer 별 담당 linguistic features 식별

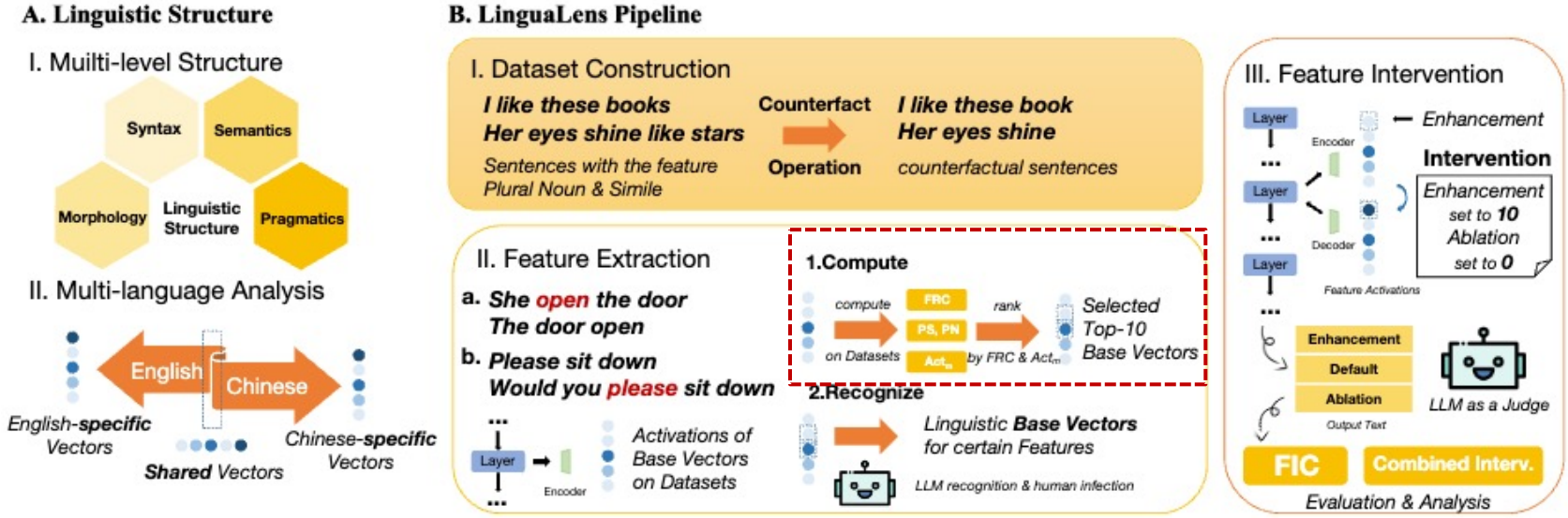


Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – base vector selection (1)**

- 모델 내부에서 **특정 언어 feature 를 담당하는** atomic 수준의 SAE basis vector(s) 의 식별 작업
  → 즉, SAE latent activations 기반 representation 재구성이 **monosemantic 분석을 가능**하게 한다는 전제에 따라.



Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – base vector selection (2)**

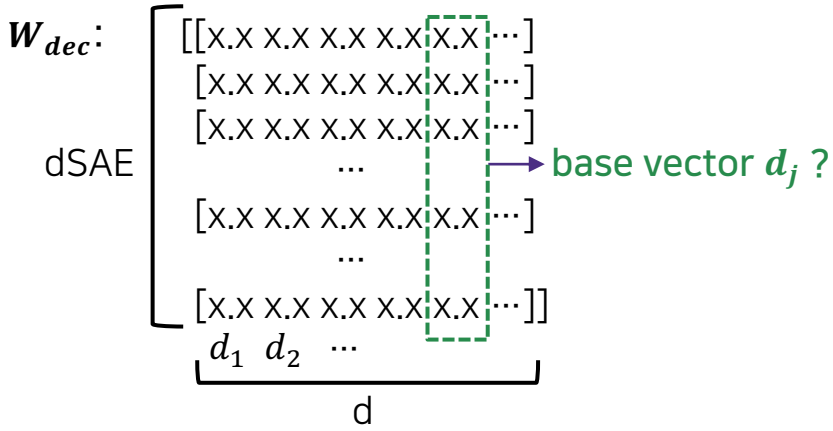
- cf) gemma scope 의 formalizations

$$\mathrm{SAE}(\boldsymbol{x}) = a(\boldsymbol{x})\boldsymbol{W}_{\mathrm{dec}} + \boldsymbol{b}_{\mathrm{dec}}, \quad a(\boldsymbol{x}) = \mathrm{JumpReLU}_{\theta}\left(\boldsymbol{x}\boldsymbol{W}_{\mathrm{enc}} + \boldsymbol{b}_{\mathrm{enc}}\right),$$

강도 담당 의미 담당

- 예시)

a(x): [0, 0, 0, 0, 0.8, 2, ⋯, 0] (1*dSAE)



→ 특정 linguistic feature 를 포함하는 example 에 대해 의미를 가장 잘 포착하는 column vectors d_j 찾아내기.
= 그리고, 이것은 곧 a(x) 내의 j-th unit 을 찾아내는 것과 동일한 작업.

→ 무슨 기준으로?

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – base vector selection (3)**

- selection 기준: 인과적 측정 scoring 제시

1. Probability of Sufficiency (PS): 특정 언어 현상이 “있을 때”, 해당 base vector 가 활성화될 충분성 (확률)
   → 즉, **s-에서 비활성 상태인 base vector 가 s+ 에서 활성화될 확률**

2. Probability of Necessity (PN): 특정 언어 현상이 “없을 때”, 해당 base vector 가 비활성화될 필요성 (확률)
   → **s+에서 활성 상태였던 base vector 가 s- 에서 비활성화될 확률**

3. Feature Representation Confidence (FRC): PS 와 PN의 조화 평균
   → base vector 가 특정 linguistic feature 를 얼마나 잘 나타내는지의 점수

$$Z_k(s) = \mathbb{I}\big[a_k(s) \geq \theta_k\big].$$

$$\mathrm{PS}_k = \Pr\big[\, Z_k^{(1)} = 1 \mid Z_k^{(0)} = 0 \,\big],$$

$$\mathrm{PN}_k = \Pr\big[\, Z_k^{(0)} = 0 \mid Z_k^{(1)} = 1 \,\big].$$

$$\mathrm{FRC}_k = 2 \cdot \frac{\mathrm{PS}_k \ \mathrm{PN}_k}{\mathrm{PS}_k + \mathrm{PN}_k}.$$

Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – base vector selection (4)**

- feature 당 50개의 examples 가 있으므로, 전체 pairs 에 대해서 (통계적으로) 확률 계산

- 예시)

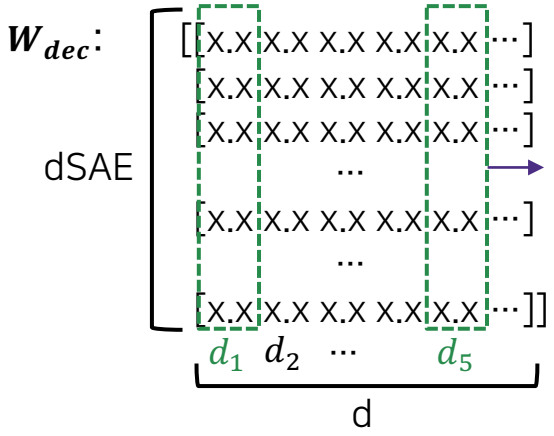
s+: I like these **books** → s-: I like these **book**

1st unit 5th unit

a(x) of books (s+): [**1**, 0, 0, 0, **0.8**, **2**, ⋯, 0] (1*dSAE) → a(x) of book (s-): [**0**, 0, 0, 0, **0**, 3, ⋯, 0]

→ [ ] 에 해당하는 units 의 PN score 에 count 추가.

→ 이렇게 구한 scores 로 각 unit index 에 상응하는 W_{dec} 내의 column vector 를 base vector 로 선택.
: 가령, 복수 명사 feature 에 대해서는 FRC score 가 d1, d5가 가장 높았다고 하면 이들이 top-2 base vectors 로 선정.



Method: LinguaLens

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** 프레임워크: SAE의 활용 – cross-layer 분석 (1)**

- 특정 linguistic feature (예: 수동태)에 대해,
  layer 마다 (모든 32개 layers 에 대해서) top-10 base vectors (의 index, unit no.)이 구해진 상황.

- 이후, 특정 linguistic feature 를 input pair 로 제공 시에,
  base vectors 의 activation frequency 와 activation values 를 보고,
  가장 **highest frequency** 및 **largest activation values** 를 산출하게 만드는 linguistic features 를 산출.

Experiment

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* Model

- LLM: Llama-3.1-8B
- SAE: OpenSAE 의 checkpoints on 32 layers of Llama-3.1-8B
  → 32개 layers (32개 SAEs)

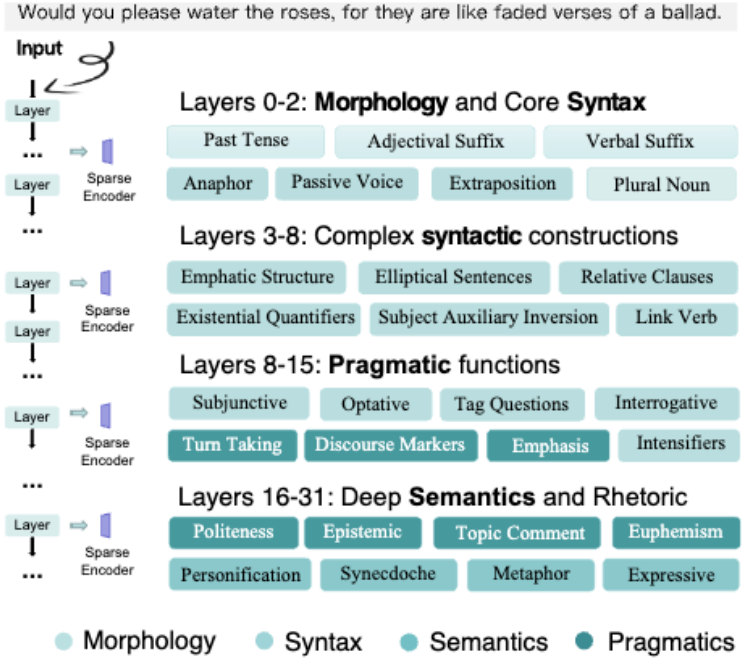
Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* Cross-layer Analysis (1)
 - Layer 구간 별로 담당하는 linguistic levels and linguistic features 가 다르다!

상세:
 - layers 0-2: 가장 하위 계층, 단어의 형태 (morphology) 및 기본적 문장 구조 (core syntax) 처리
  . **Morphology**: 'Past Tense'(과거 시제), 'Adjectival Suffix'(형용사 접미사), 'Verbal Suffix'(동사 접미사), 'Plural Noun'(복수 명사) 등
  . **Syntax**: 'Anaphor'(대용어), 'Passive Voice'(수동태), 'Extraposition'(외치) 등

 - layers 3-8: 중간 계층, 더 복잡한 구문 구조 처리
  . Complex **Syntax**: 'Emphatic Structure'(강조 구문), 'Elliptical Sentences'(생략 문장), 'Relative Clauses'(관계절), 'Existential Quantifiers'(존재 수량사), 'Subject Auxiliary Inversion'(주어-조동사 도치), 'Link Verb'(연결 동사) 등



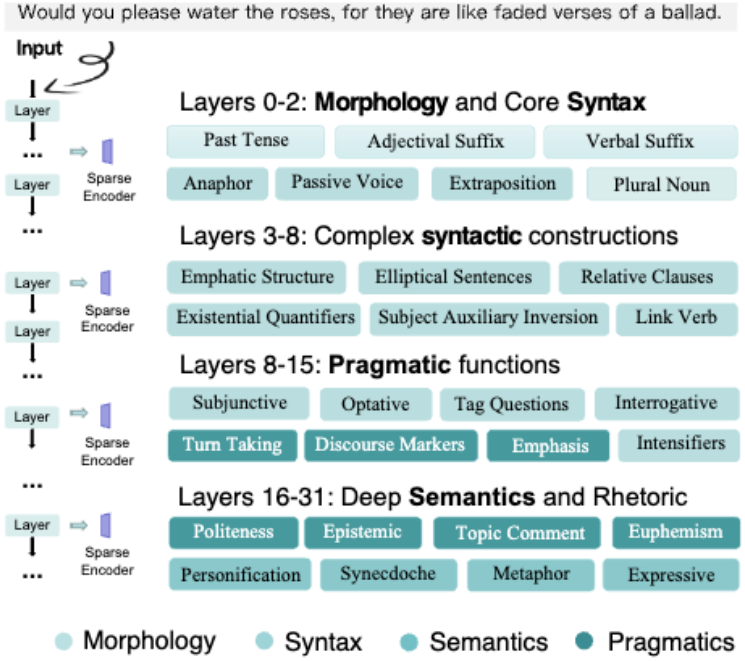
Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* Cross-layer Analysis (2)

상세:
- layers 8-15: 문장 자체의 의미를 넘어서는 화자의 의도, 문맥적 의미 등 처리
. **Pragmatics**: 'Subjunctive'(가정법), 'Optative'(기원법), 'Tag Questions'(부가 의문문), 'Interrogative'(의문문), 'Turn Taking'(말 차례 교환), 'Discourse Markers'(담화 표지), 'Emphasis'(강조), 'Intensifiers'(강조사) 등

cf) Would you please pass the salt?는 실제 질문이 아닌 요청의 의미 → 이런 것이 화용론



Results

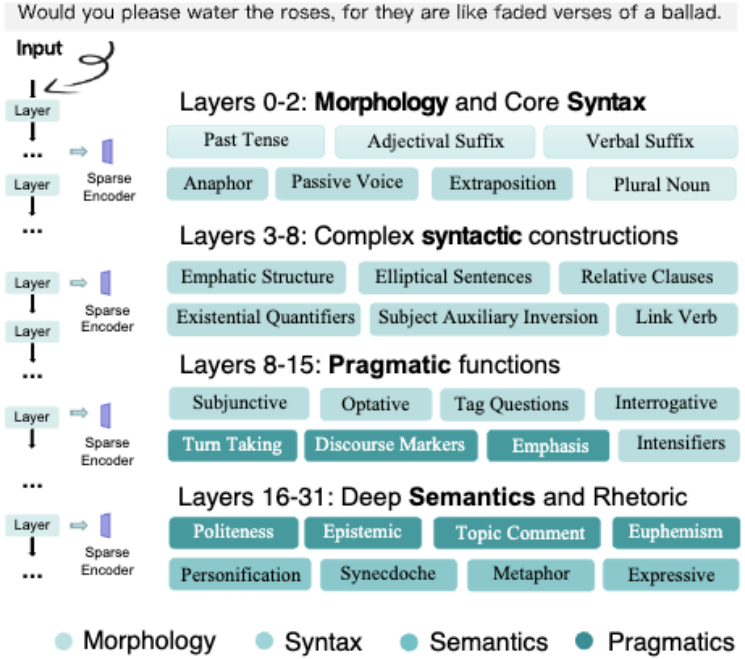
LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* Cross-layer Analysis (3)

상세:
- layers 16-31: 가장 상위 계층, 추상적, 심층적 의미 및 수사적 표현 이해 및 생성
. **Pragmatics**: 'Politeness'(공손함), 'Epistemic'(인식적 양태), 'Euphemism'(완곡어법), 'Expressive'(표현적 언어)
. **Semantics**: 'Topic Comment'(화제-논평 구조), 'Personification'(의인화), 'Synecdoche'(제유), 'Metaphor'(은유)

cf) 여기서 semantics 는 단순 word 하나의 의미 정도가 아니라, 맥락 전체까지 고려한 의미론, 단어 자체의 의미 등은 하위 morphology 에서 이미 처리한다는 의미.

→ 모델이 언어를 "계층적 정보 처리" 한다.



Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** Cross-layer Analysis – tricky 한 부분**

- Act_m score: maximum activation score
  = 특정 linguistic feature (혹은 level)이 input pair 로 들어올 때, 해당 layer 구간에서 얼마나 강하게 활성화 되는지.

- 활성화 강도 (평균) vs. highest frequency and largest values 서로 다를 수 있다.
  왜냐하면 상위 계층으로 갈수록 "종합" 되는 것이기 때문이다.
  라고 함.

→ 논란의 여지가 있을 수도.
magnitude vs. Act_m 의 차이?

Lang	PS	PN	FRC	Act_m 0	8	15	24	30
Morphology								
CH	0.61	0.70	0.64	0.01	0.19	0.29	0.52	1.36
EN	0.73	0.80	0.75	0.03	0.35	0.49	1.02	1.89
Syntax								
CH	0.84	0.90	0.86	0.20	0.50	0.95	2.32	3.37
EN	0.79	0.87	0.82	0.12	0.35	0.68	1.66	2.59
Semantics								
CH	0.72	0.78	0.74	0.09	0.29	0.57	1.41	2.18
EN	0.76	0.83	0.78	0.11	0.32	0.55	1.34	2.01
Pragmatics								
CH	0.69	0.74	0.70	0.06	0.25	0.42	1.03	1.56
EN	0.77	0.83	0.79	0.13	0.27	0.52	1.33	2.03

Results

*** Cross-layer Analysis – tricky 한 부분**

- maximum activation score (Act_m) 점수 측정:
  정해진 base vectors 에 상응하는 activation values (강도) 비교

- 측정 방법
  1. positive examples (s+ 만)에 대한 activation 측정
  2. 최대 활성화 값 추출: 각 s+ 문장 내에서 가장 활성화 값이 높은 지점(token 혹은 span) 값 추출
  3. 전체 예제 평균 계산: 추출된 최대 활성화 값들을 모든 s+ 예제들에 대해서 평균 값 도출.

→ score 의 최종 단일 값은 top-10 base vectors 중에서 top-3 의 평균값 사용.

- 예를 들어, "수동태" 유형에 대해서 [1, 3, 1000, 1001, 2500, …] indices 가 base vectors 로 정해졌다면,

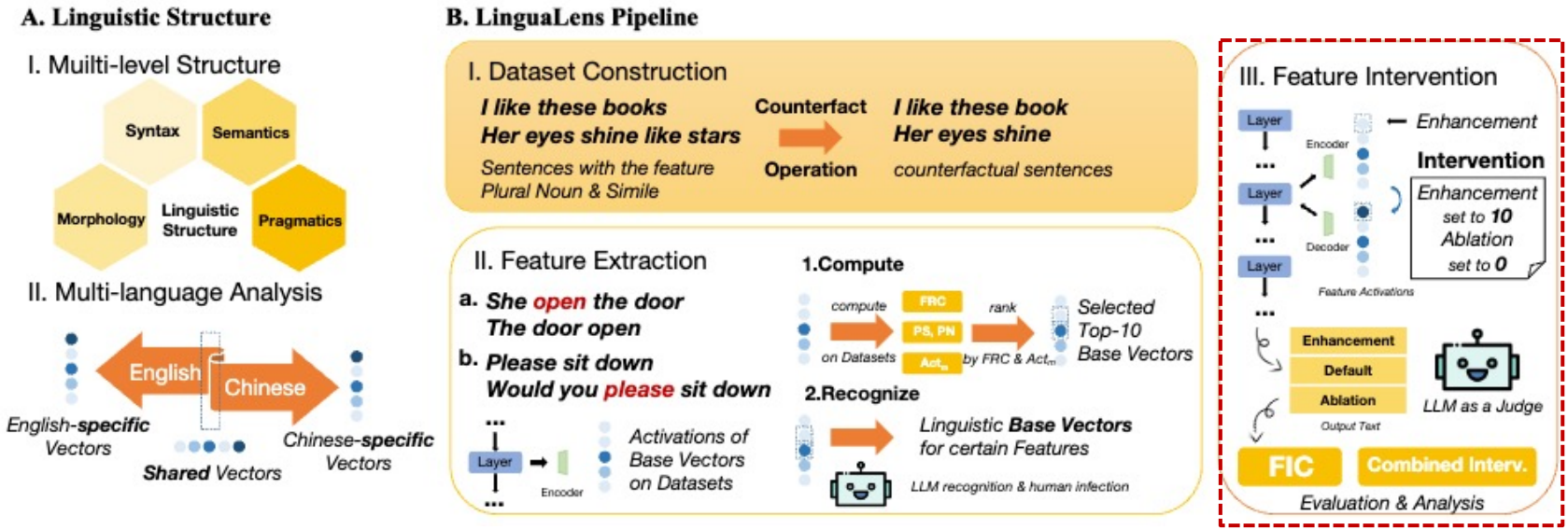
→ "The apple **is eaten** by the boy" 이라는 s+ 가 주어질 때, 1-th unit 중에서 activation 값이 "eaten" token 에서 가장 높았다면, 그 가장 높은 점수 (혹은 점수들의 평균) 추출
→ 모든 '수동태' s+ 예제 (50개)에 대한 이러한 점수들을 평균냄
→ 이중에서 FRC 점수 기준 상위 top-3 indices 의 activation values 를 평균낸 것이 곧 Act_m 점수.

Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

* Intervention 를 통한 output steering (1)

- 특정 linguistic feature 를 input 하고, 사전 식별된 top base vectors 의 activation values 를 임의적으로 조절함. (0 or 10)



Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** Intervention 를 통한 output steering (2)**

- 성공 cases

case 1: top base vectors 의 activation values 를 높일 때 (enhance), 해당 linguistic feature 가 강화된 output 이 생성된다.

case 2: top base vectors 의 activation values 를 제거할 때 (ablate), 해당 linguistic feature 가 제거된 output 이 생성된다.

#	Intervene	Model Output
1	Default	The wind blows snow into my eyes as I trudge through the blizzard.
	Enhance	As the cold descends, I feel the weight of my breath in my throat. It's an icy haze.
	Ablate	The winter sky was cold. The ice was hard under his boots.
2	Default	Love is the burning passion of a summer night.
	Enhance	I feel like butterflies are in my stomach. My heart is beating faster than normal.
	Ablate	The more you write, the more time and love you will have.

직유 feature steering 질적 예제

cf) 직유? 표현하고 싶은 대상(원관념)을 비슷한 성질을 가진 다른 대상(보조관념)에 '~처럼', '~같이', '~듯이' 등의 연결어를 써서 직접 빗대어 표현하는 비유법.
e.g.) '내 마음은 호수 같다'

Results

LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder (EMNLP 2025)

*** Intervention 를 통한 output steering (3)**

- 단, linguistic type 마다 intervention 의 효과성은 다를 수 있음

- exp: top vectors steering 성공 확률
  ctr: random vectors steering 성공 확률

- FIC ≈ 전반적 steering 성공율

Feature	ID	Enhance		Ablate		FIC
		exp	ctr	exp	ctr	
Morphology						
Past-Tense	8L4016	12.0	4.0	48.0	44.0	8.3
Syntax						
Linking Verb	18L61112	52.0	24.0	48.0	40.0	22.9
Semantics						
Causality	22L53236	32.0	20.0	40.0	36.0	12.0
Simile	26L75327	72.0	52.0	48.0	52.0	6.9
Pragmatics						
Politeness	31L578	60.0	32.0	44.0	20.0	46.9

Conclusion

*** 아쉬운 점**

- counterfactual dataset construction 시, linguistic feature 를 딱 한 개씩만 넣었으므로 (minimal edit), 현실 세계의 다양한 linguistic features and levels 가 들어가는 복합적 구문들에 대해서는 분석하려면 멀었다.

- these books → these book 과 같은 예제는 minimal edit 원칙이라고 포장하긴 했지만, 사실 비문임. 이러한 지점이 LLM behavior 에 영향을 미치지는 않았을 것인가.

- 너무 validation dataset 의 품질에 dependent 할 수 있음

- Act_m score 는 metric 자체가 잘못됐거나, 아니면 해석 방향 자체가 잘못됐거나..?

감사합니다