Assessing the Language Acquisition and Linguistic Abilities of Large Language Models

Hai Hu, Shanghai Jiao Tong University

2024/11/29

LT Forum

@CityU HK

Paper, code and data available at: huhailinguist.github.io

Acknowledgements

Collaborators:

- Indiana University: Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Sandra Kübler, Chien-Jer Charles Lin, Larry Moss
- Renmin University of China: Dawei Lu, Xinyue Yao
- SJTU SFL: Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Xinyuan Cheng, Shisen Yue
- SJTU CS: Ziyin Zhang, Rui Wang
- Allen Institute for Artificial Intelligence: Kyle Richardson
- Tongyi Lab, Aliyun: Jialong Tang, Pei Zhang, Baosong Yang

Funding:

- Indiana -- Renmin Seed Grant / SJTU startup grant
- Ministry of Education Social Science and Humanities Grant
- Shanghai Pujiang Grant

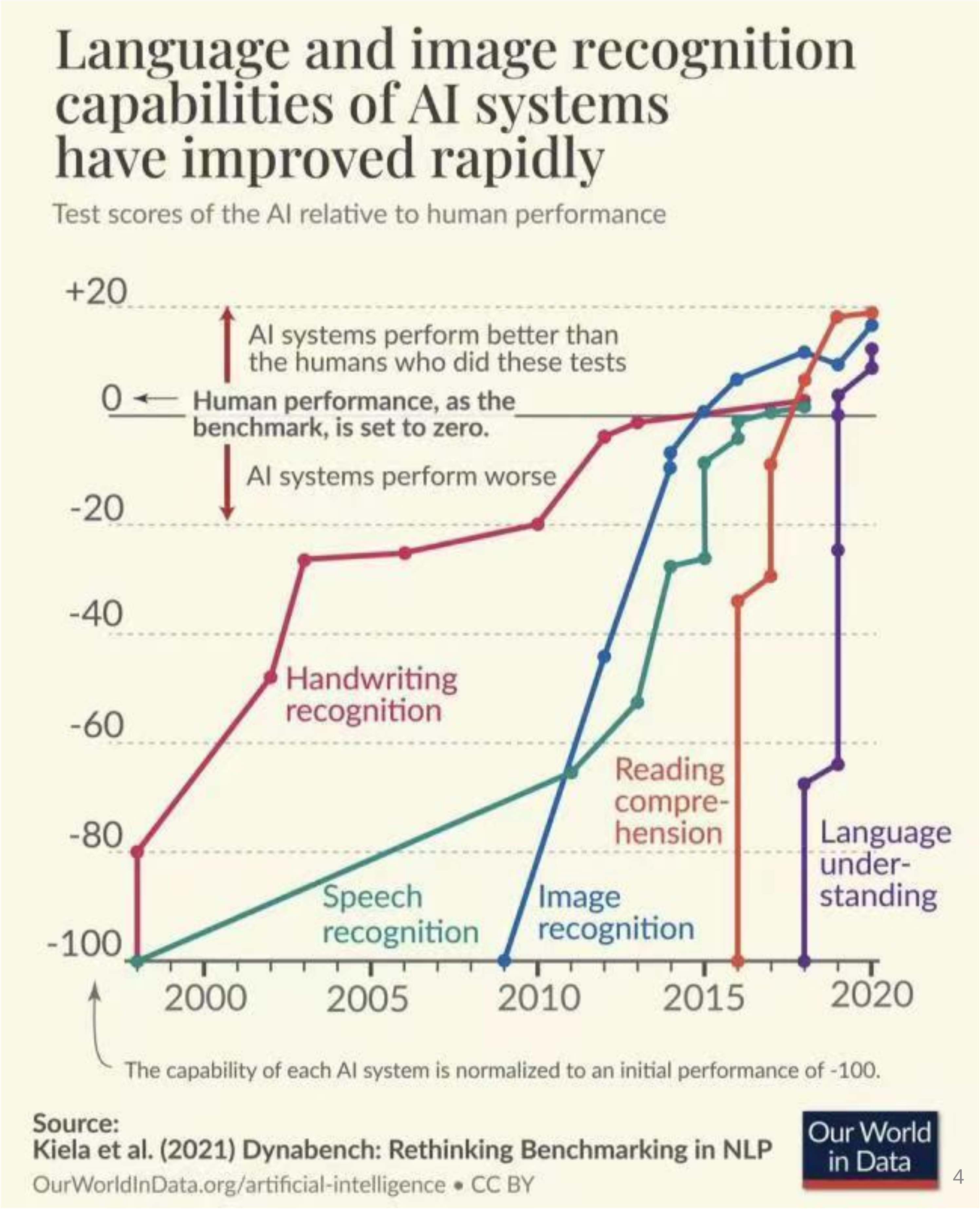
Agenda

- AI and NLP: then and now
- Our work on analyzing the acquisition patterns of LMs
- Our work on evaluating linguistic abilities of LMs
- Moving forward

Progress in AI

AI models quickly "surpass" humans

(Kiela et al 2021)



Language and image recognition capabilities of AI systems have improved rapidly

Test scores of the AI relative to human performance

The capability of each AI system is normalized to an initial performance of -100.

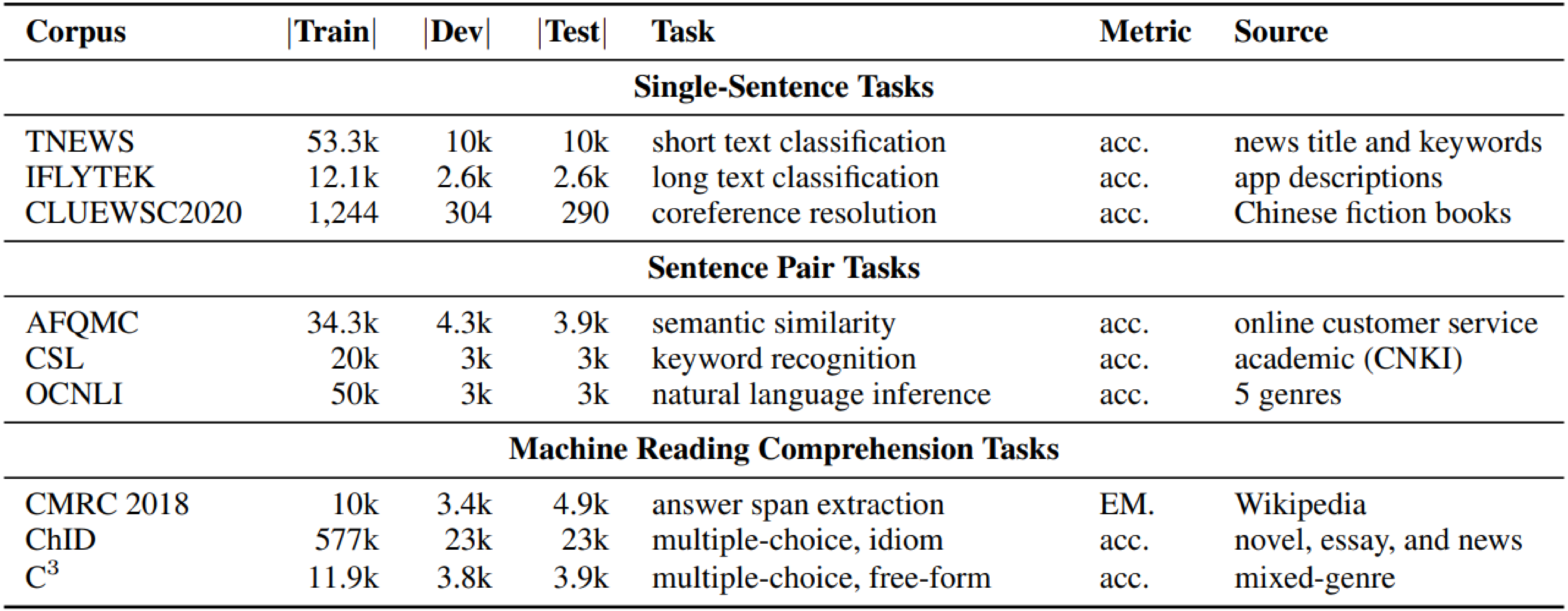
Source:
Kiela et al. (2021) Dynabench: Rethinking Benchmarking in NLP
OurWorldInData.org/artificial-intelligence • CC BY

Our World in Data

CLUE: Chinese Language Understanding Evaluation

Similar trend in Chinese

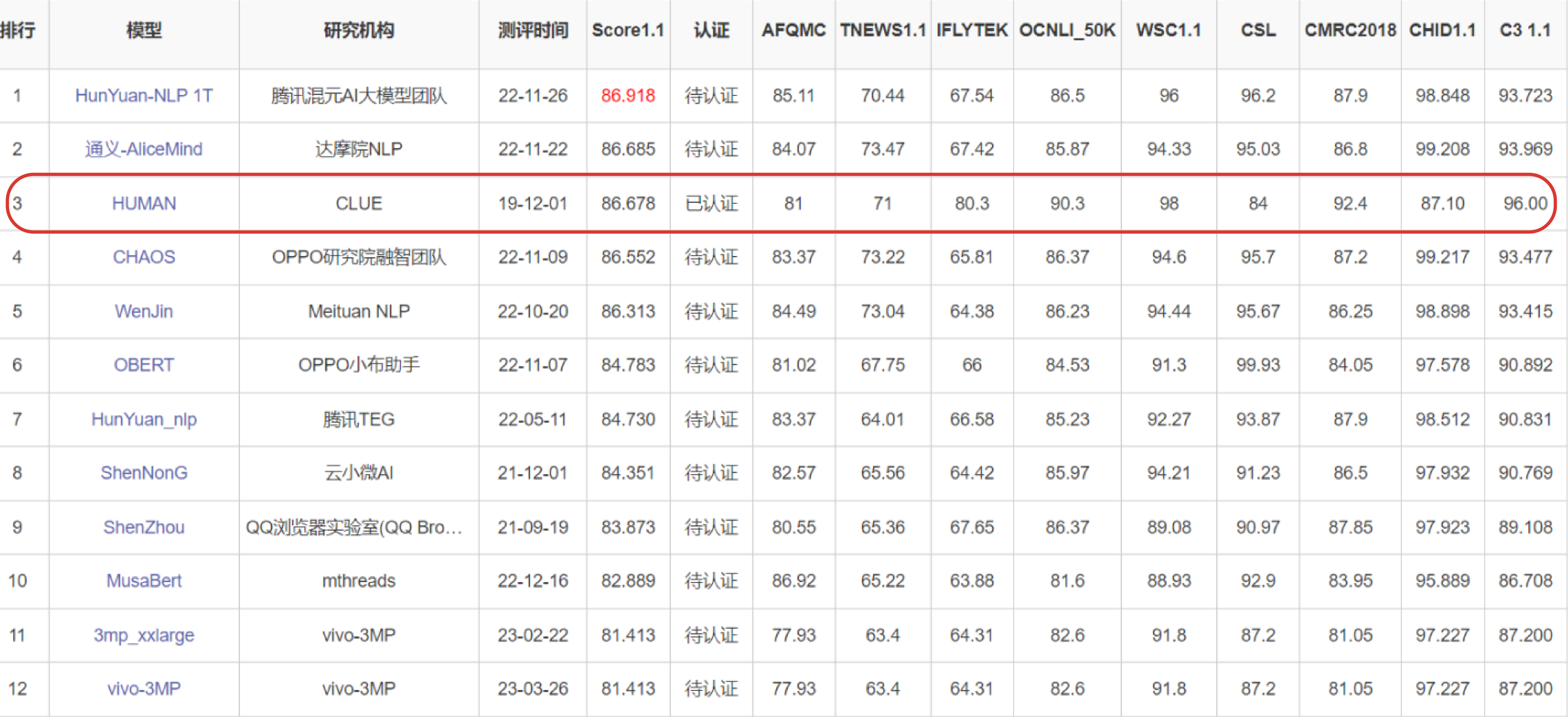
9 tasks in language understanding in Chinese



Corpus	|Train|	|Dev|	|Test|	Task	Metric	Source
Single-Sentence Tasks						
TNEWS	53.3k	10k	10k	short text classification	acc.	news title and keywords
IFLYTEK	12.1k	2.6k	2.6k	long text classification	acc.	app descriptions
CLUEWSC2020	1,244	304	290	coreference resolution	acc.	Chinese fiction books
Sentence Pair Tasks						
AFQMC	34.3k	4.3k	3.9k	semantic similarity	acc.	online customer service
CSL	20k	3k	3k	keyword recognition	acc.	academic (CNKI)
OCNLI	50k	3k	3k	natural language inference	acc.	5 genres
Machine Reading Comprehension Tasks						
CMRC 2018	10k	3.4k	4.9k	answer span extraction	EM.	Wikipedia
ChID	577k	23k	23k	multiple-choice, idiom	acc.	novel, essay, and news
C^3	11.9k	3.8k	3.9k	multiple-choice, free-form	acc.	mixed-genre

Xu, Hu, et al 2020; *COLING*

Performance on CLUE

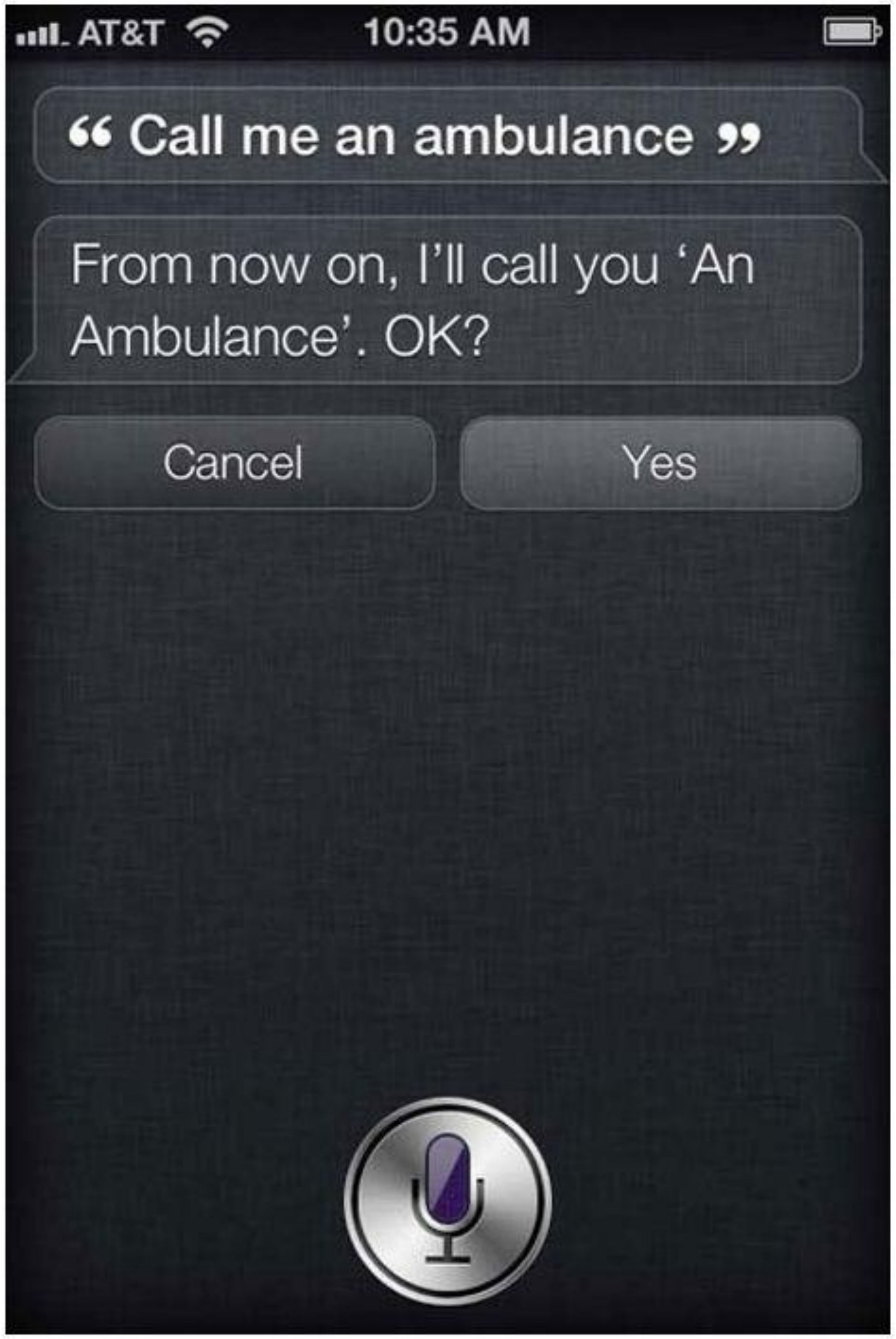


排行	模型	研究机构	测评时间	Score1.1	认证	AFQMC	TNEWS1.1	IFLYTEK	OCNLI_50K	WSC1.1	CSL	CMRC2018	CHID1.1	C3 1.1
1	HunYuan-NLP 1T	腾讯混元AI大模型团队	22-11-26	86.918	待认证	85.11	70.44	67.54	86.5	96	96.2	87.9	98.848	93.723
2	通义-AliceMind	达摩院NLP	22-11-22	86.685	待认证	84.07	73.47	67.42	85.87	94.33	95.03	86.8	99.208	93.969
3	HUMAN	CLUE	19-12-01	86.678	已认证	81	71	80.3	90.3	98	84	92.4	87.10	96.00
4	CHAOS	OPPO研究院融智团队	22-11-09	86.552	待认证	83.37	73.22	65.81	86.37	94.6	95.7	87.2	99.217	93.477
5	WenJin	Meituan NLP	22-10-20	86.313	待认证	84.49	73.04	64.38	86.23	94.44	95.67	86.25	98.898	93.415
6	OBERT	OPPO小布助手	22-11-07	84.783	待认证	81.02	67.75	66	84.53	91.3	99.93	84.05	97.578	90.892
7	HunYuan_nlp	腾讯TEG	22-05-11	84.730	待认证	83.37	64.01	66.58	85.23	92.27	93.87	87.9	98.512	90.831
8	ShenNonG	云小微AI	21-12-01	84.351	待认证	82.57	65.56	64.42	85.97	94.21	91.23	86.5	97.932	90.769
9	ShenZhou	QQ浏览器实验室(QQ Bro…	21-09-19	83.873	待认证	80.55	65.36	67.65	86.37	89.08	90.97	87.85	97.923	89.108
10	MusaBert	mthreads	22-12-16	82.889	待认证	86.92	65.22	63.88	81.6	88.93	92.9	83.95	95.889	86.708
11	3mp_xxlarge	vivo-3MP	23-02-22	81.413	待认证	77.93	63.4	64.31	82.6	91.8	87.2	81.05	97.227	87.200
12	vivo-3MP	vivo-3MP	23-03-26	81.413	待认证	77.93	63.4	64.31	82.6	91.8	87.2	81.05	97.227	87.200

Xu, Hu, et al 2020; *COLING*

A concrete example

Many years ago ...



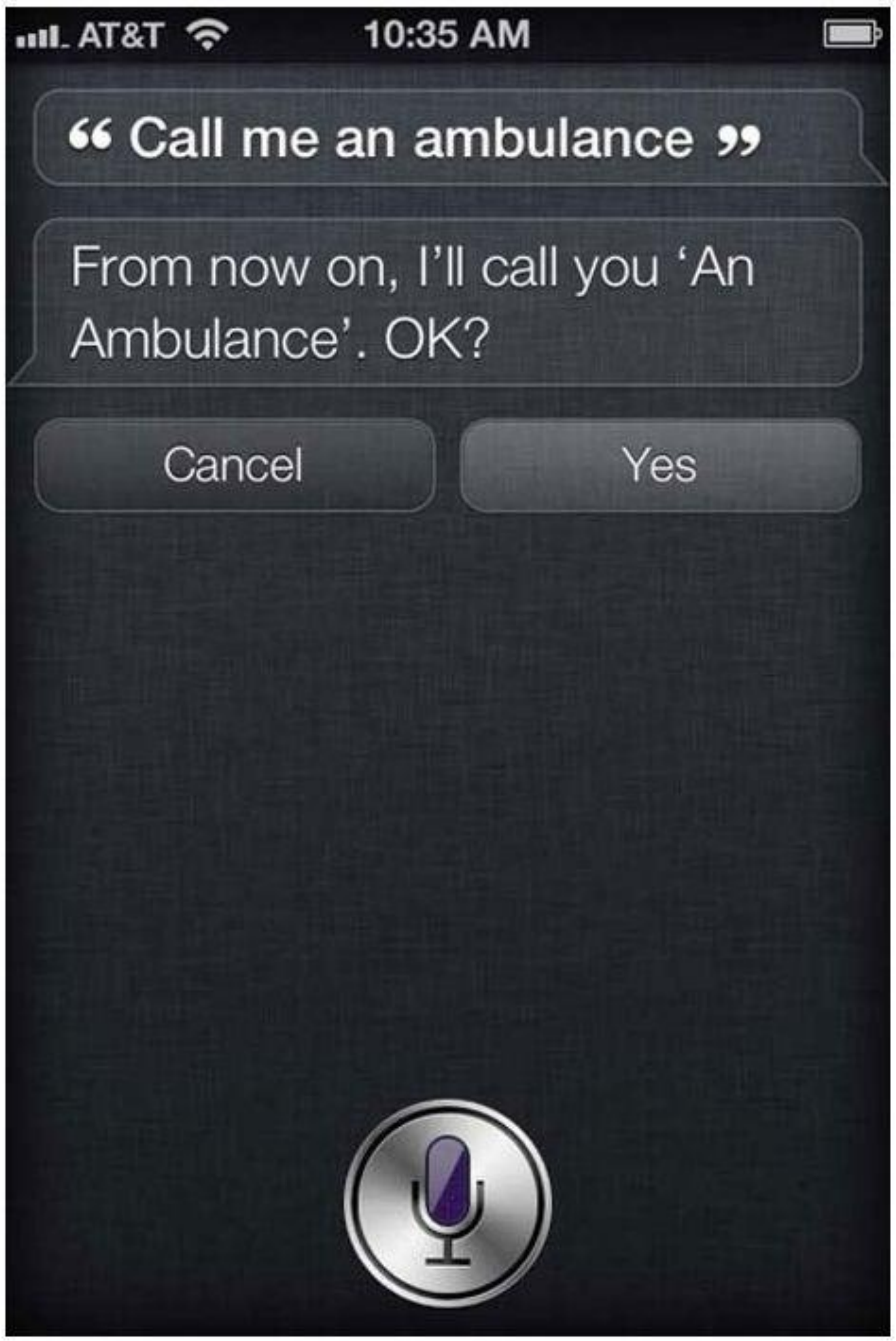
AT&T 10:35 AM

“ Call me an ambulance ”

From now on, I’ll call you ‘An Ambulance’. OK?

A concrete example

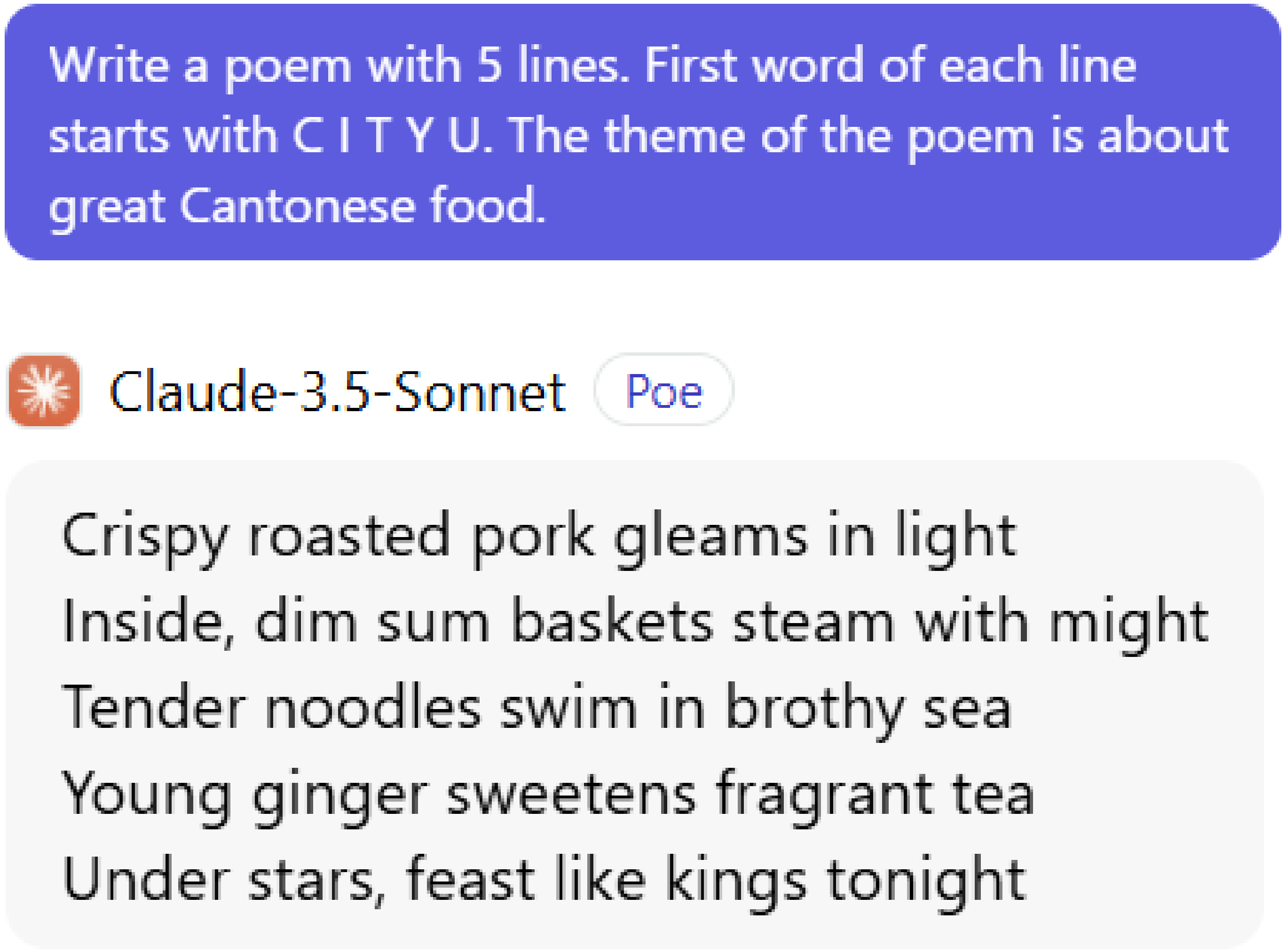
Many years ago ...



“ Call me an ambulance ”

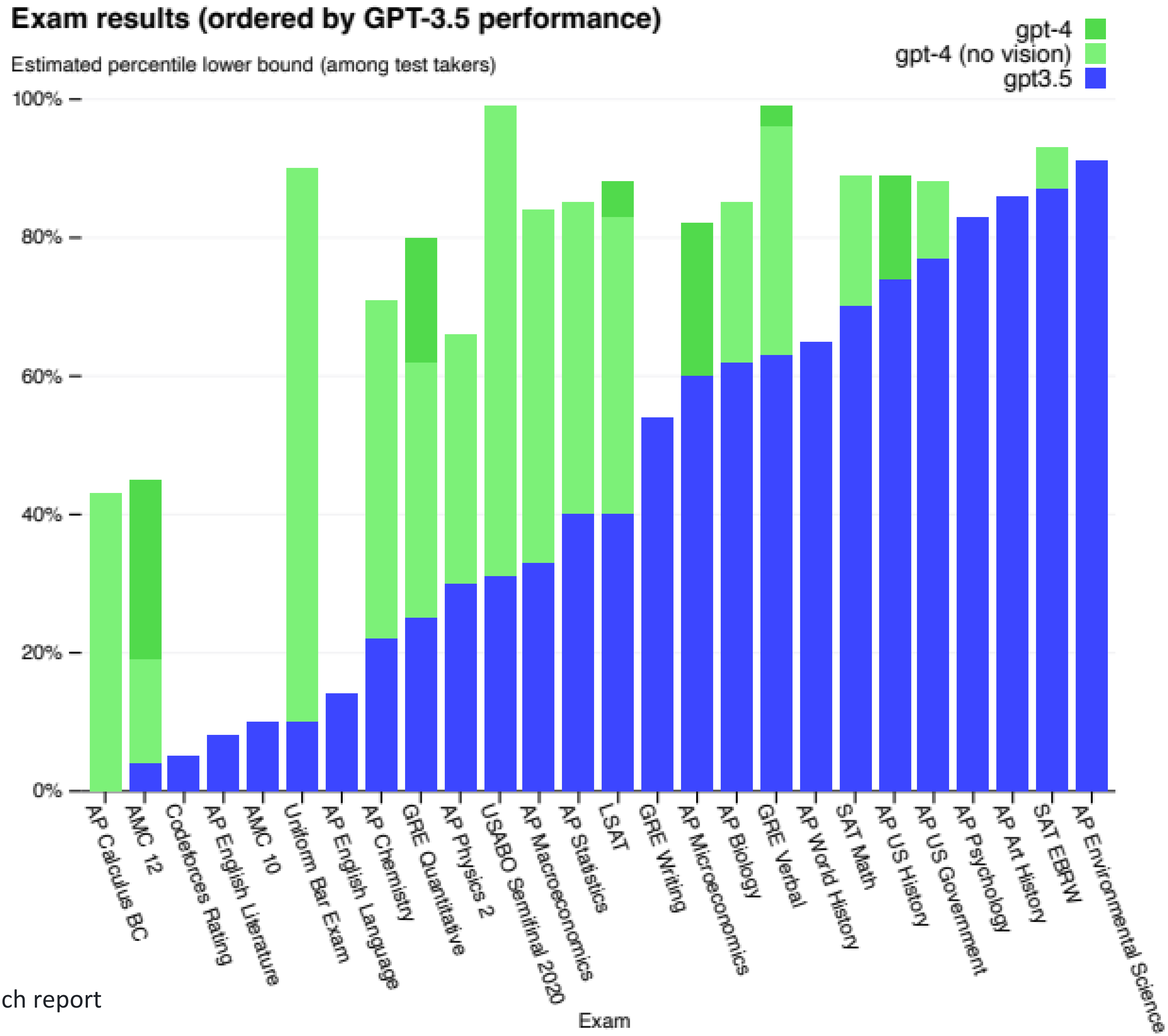
From now on, I’ll call you ‘An Ambulance’. OK?

2024 ...


Write a poem with 5 lines. First word of each line starts with C I T Y U. The theme of the poem is about great Cantonese food.

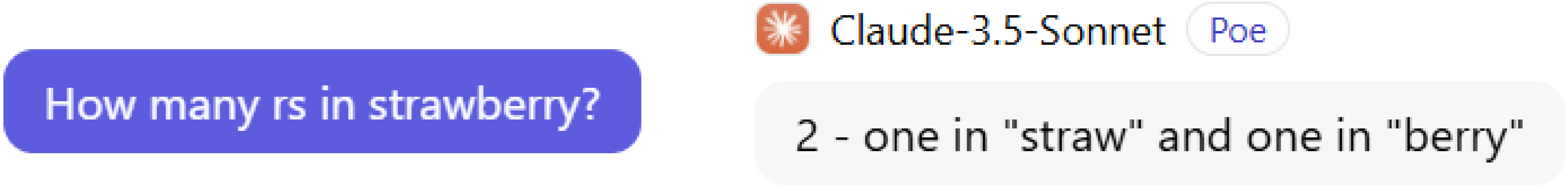
Crispy roasted pork gleams in light
Inside, dim sum baskets steam with might
Tender noodles swim in brothy sea
Young ginger sweetens fragrant tea
Under stars, feast like kings tonight

Oh btw, LLMs speak dozen languages, and ...



OpenAI: GPT4 tech report

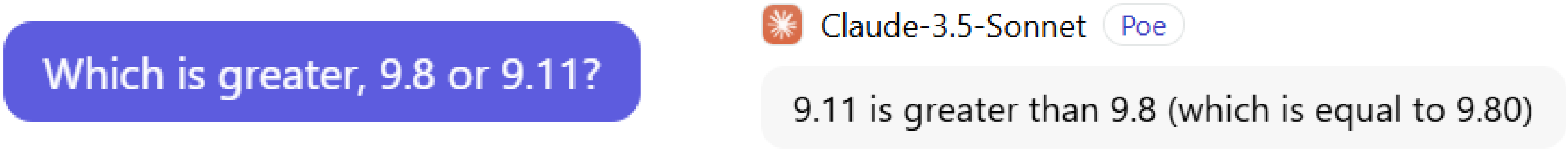
However:



How many rs in strawberry?

Claude-3.5-Sonnet Poe

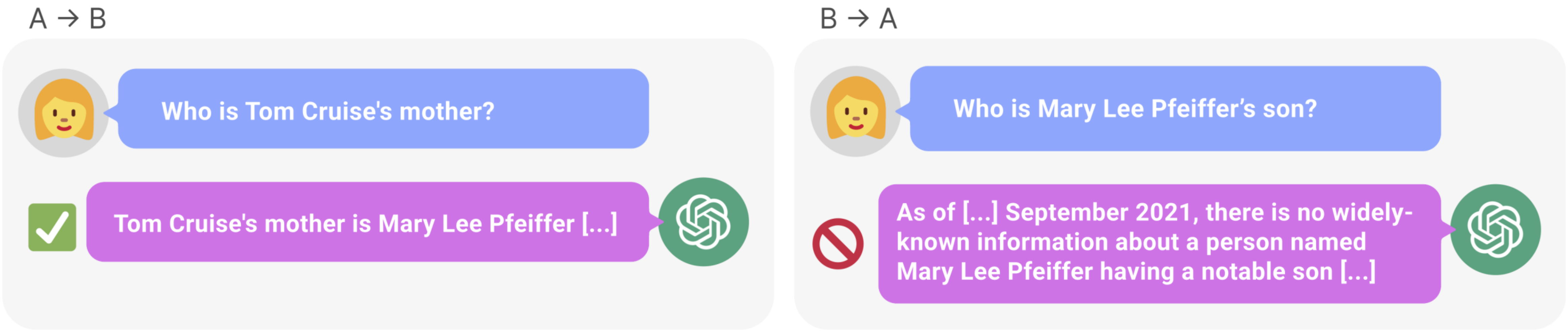
2 - one in "straw" and one in "berry"



Which is greater, 9.8 or 9.11?

Claude-3.5-Sonnet Poe

9.11 is greater than 9.8 (which is equal to 9.80)



A → B

Who is Tom Cruise's mother?

✓ Tom Cruise's mother is Mary Lee Pfeiffer [...]

B → A

Who is Mary Lee Pfeiffer's son?

🚫 As of [...] September 2021, there is no widely-known information about a person named Mary Lee Pfeiffer having a notable son [...]

Reversal curse: Burglund et al 2023; ICLR

Important research questions

What is the limit for (Transformer-based) LLMs?

What ***things*** are learnable? What are not?

No matter how many parameters LLMs have (0.5B, 7B, 72B, 405B)

or how much data you throw at the LLMs (Trillions of tokens)

things = syntax/semantics/logic/math/reasoning/etc.

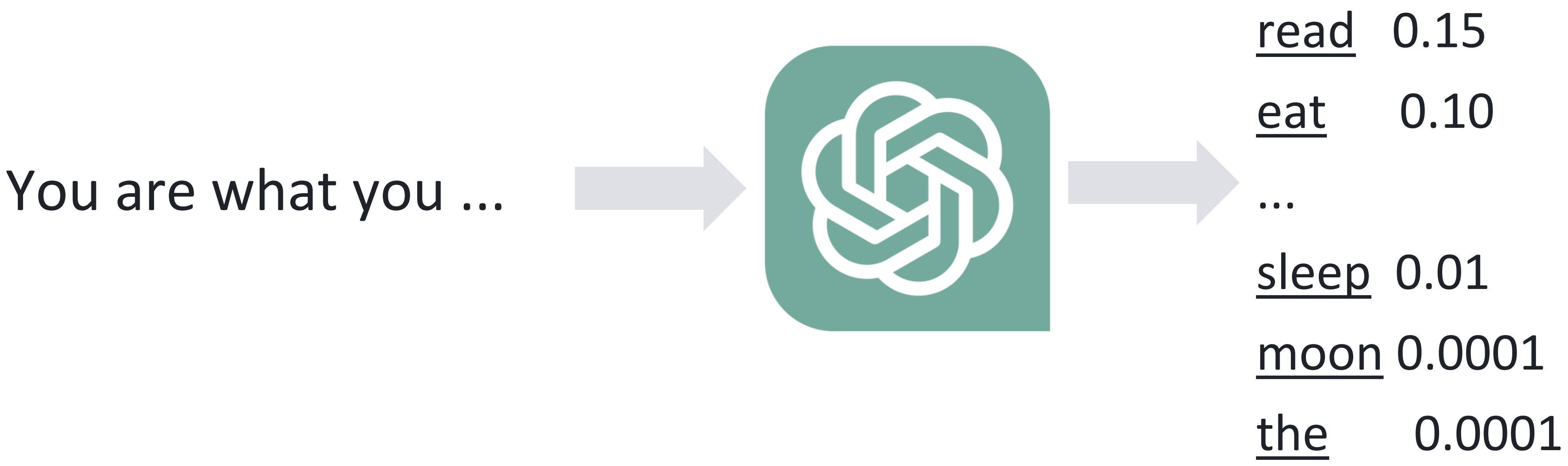
How do LLMs learn? Seem to differ from children, but how exactly?

How is ChatGPT trained?

1 Pretraining: predict next word, based on HUGE amounts of raw text

How is ChatGPT trained?

1 Pretraining: predict next word, based on HUGE amounts of raw text

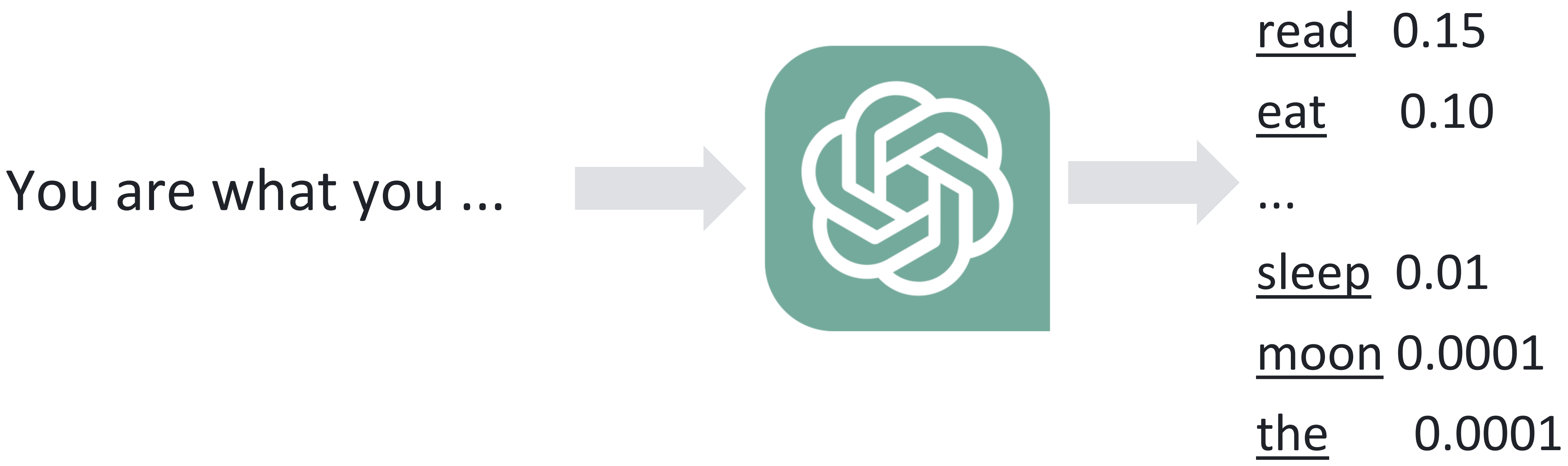


You are what you ...

read	0.15
eat	0.10
…	
sleep	0.01
moon	0.0001
the	0.0001

How is ChatGPT trained?

1 Pretraining: predict next word, based on HUGE amounts of raw text



You are what you ...

read	0.15
eat	0.10
...	
sleep	0.01
moon	0.0001
the	0.0001

What does the language model (LM) learn?

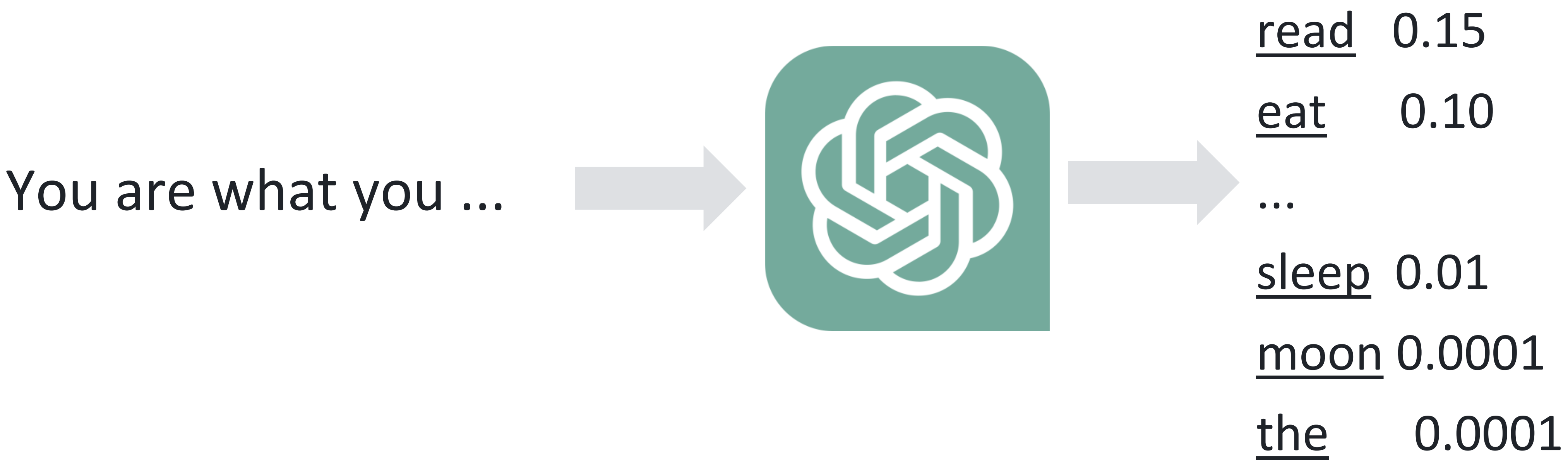
- P(read | you are what you) > P(the | you are what you)   syntax
- P(ever | I have not) > P(ever | I have)   syntax/semantics (NPI)
- Pandas are black and ____ (white > yellow)   World knowledge
- The most populous city in China is ____ (Beijing? Shanghai? Guangzhou?)

How is ChatGPT trained?

1 Pretraining: predict next word, based on HUGE amounts of raw text

2 Supervised fine-tuning (SFT): teach model to follow instructions

3 Alignment: align model responses with human values



You are what you ...

read	0.15
eat	0.10
...	
sleep	0.01
moon	0.0001
the	0.0001

What does the language model (LM) learn?

- P(read | you are what you) > P(the | you are what you)   syntax
- P(ever | I have not) > P(ever | I have)   syntax/semantics (NPI)
- Pandas are black and _____ (white > yellow)   World knowledge
- The most populous city in China is _____ (Beijing? Shanghai? Guangzhou?)

Part 1: how do LLMs acquire Chinese syntax

Analogy

Birds and planes both fly, but based on different laws of physics

Babies and LLMs speak human languages, but under diff mechanisms

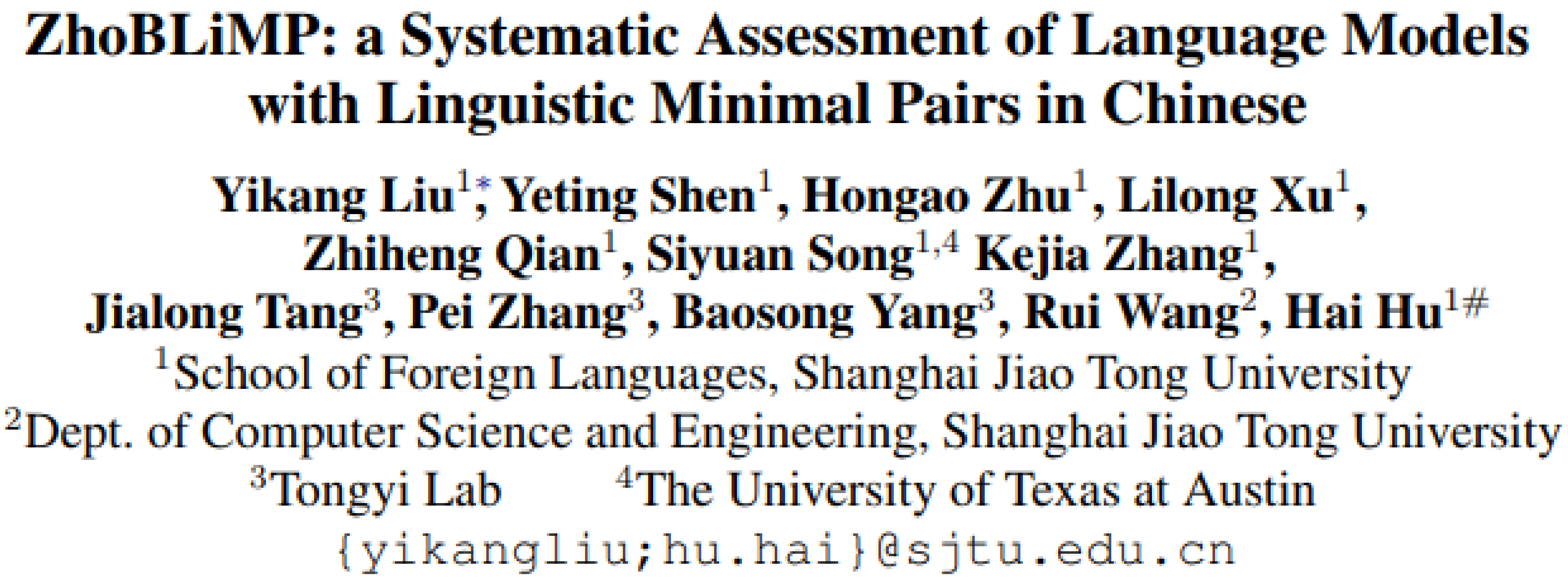
Analogy

Birds and planes both fly, but based on different laws of physics

Babies and LLMs speak human languages, but under diff mechanisms

Why do we want to connect babies learning language with LLMs' learning

- We don't yet fully understand how **babies** or **LLMs** learn!
- Comparison is still meaningful!
  - How LLMs learn --> help us understand how babies learn
  - How babies learn --> help us train LLMs more efficiently



ZhoBLiMP: a Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese

Yikang Liu[1*], Yeting Shen[1], Hongao Zhu[1], Lilong Xu[1],
Zhiheng Qian[1], Siyuan Song[1,4] Kejia Zhang[1],
Jialong Tang[3], Pei Zhang[3], Baosong Yang[3], Rui Wang[2], Hai Hu[1#]
[1]School of Foreign Languages, Shanghai Jiao Tong University
[2]Dept. of Computer Science and Engineering, Shanghai Jiao Tong University
[3]Tongyi Lab [4]The University of Texas at Austin
{yikangliu;hu.hai}@sjtu.edu.cn

https://arxiv.org/pdf/2411.06096v1

code and data: https://github.com/sjtu-compling/ZhoBLiMP

Background: (Large) Language Models

Test LLMs' syntactic abilities:

(1) directly ask LLMs:

Is the following a good sentence: The books of the lady is new.

(2) see if LLMs assign higher probability to good sent of a minimal pair

P(The books of the lady are new.) > P(The books of the lady is new.)

Based on how LLMs are train

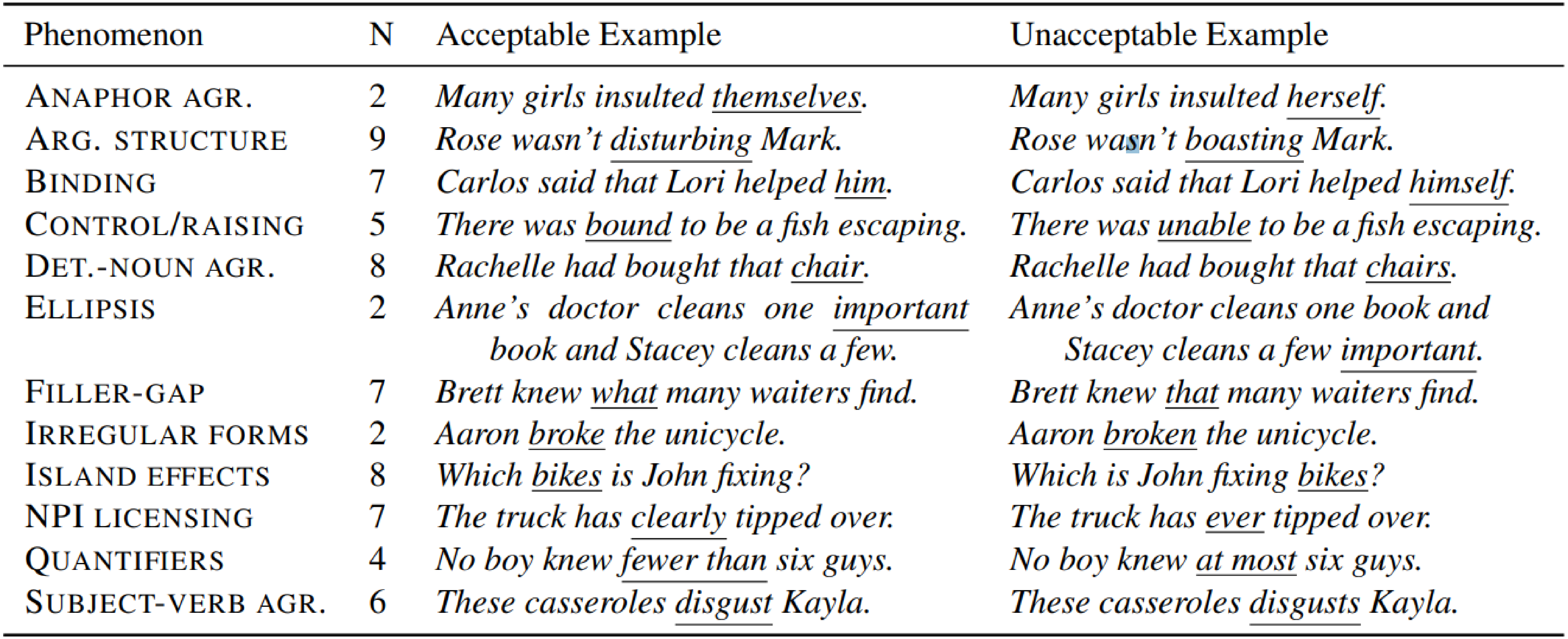
Method (1) is probing performance of LLMs

where as (2) is probing competence of LLMs (Hu and Levy 2023; EMNLP)

We use (2) here.

Background: Assess LMs with minimal pairs

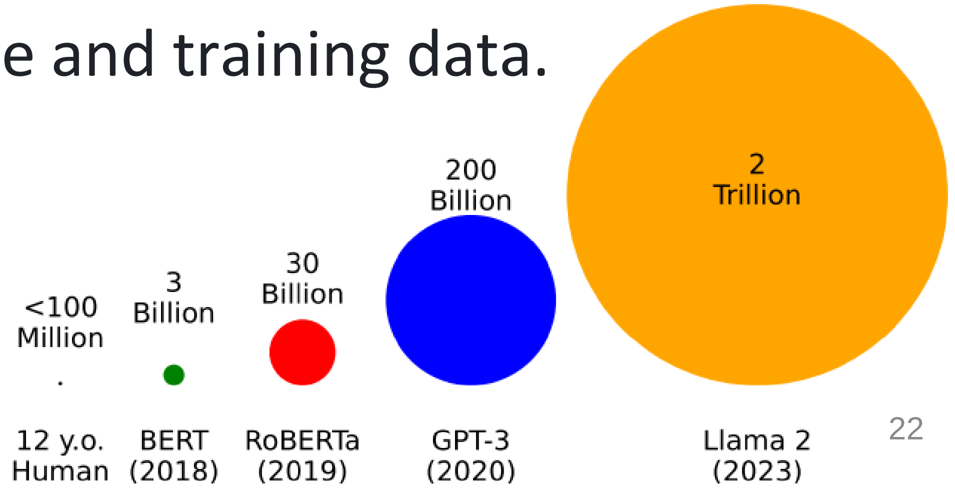
- Benchmark of Linguistic Minimal Pairs: **BLiMP** English (Warstadt et al 2020)
- A paradigm for evaluating LMs on (mostly) syntax
- 67 paradigms, 67k minimal pairs.



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Background: studies using BLiMP

- Billion words (Zhang et al 2021, ACL):
  - syn/sem: only need 10M or 100M training words (note: encoder LMs)
  - commonsense, other skills: more data
- BabyLM (Warstadt et al 2023/2024):
  - "developmentally plausible": 100M words for 12 y.o.
  - train LMs with only 100M words
- Learning trajectories in different phenomena?
  - Evanson et al. (2023): U-shape learning curves on a group of phenomena.
- Factors affect acquisition of syntax in LMs:
  - Choshen et al. (2022): LMs acquire different English syntactic phenomena in a similar order regardless of initialization, architecture and training data.

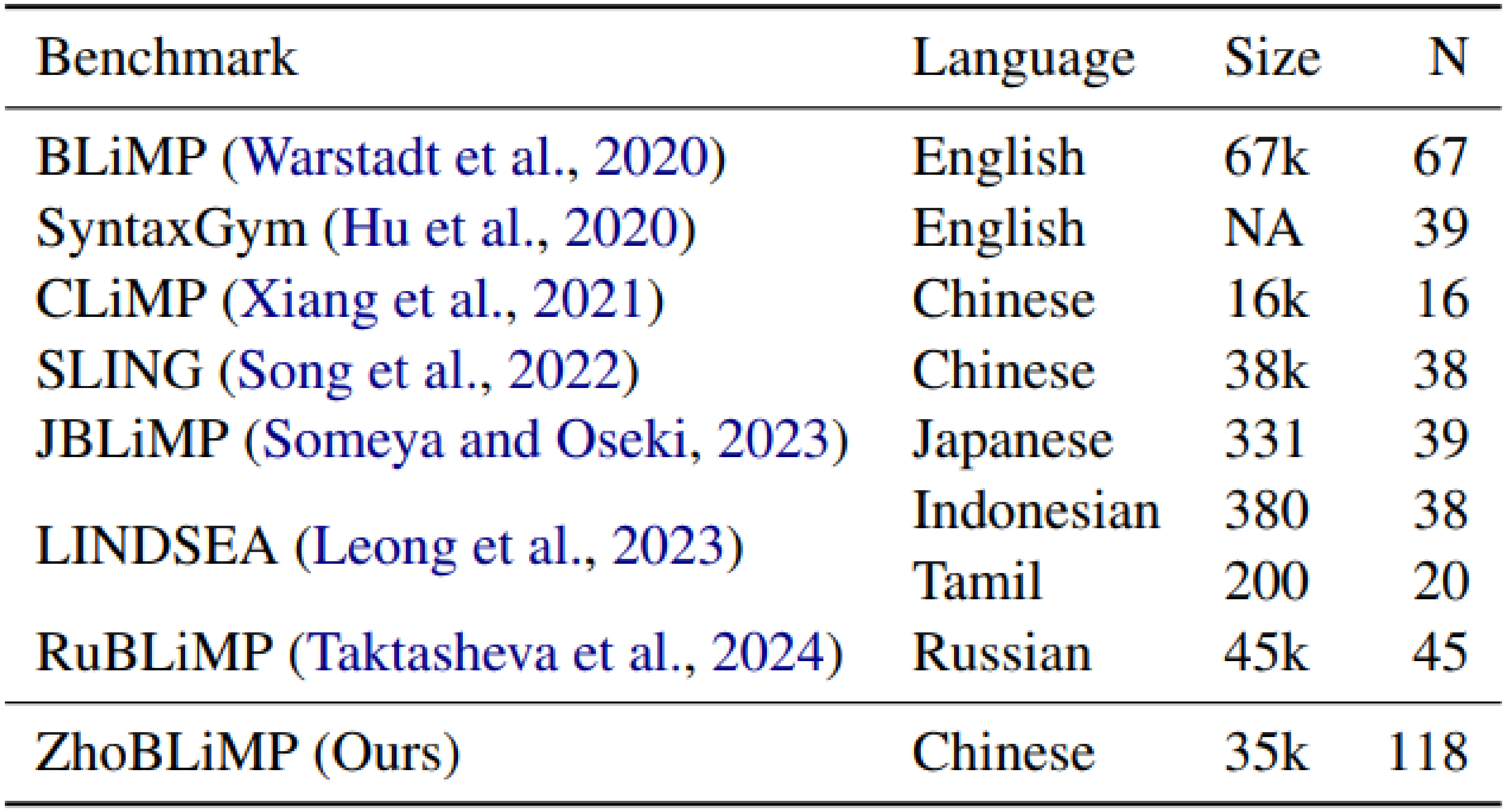


Background: BLiMP-style datasets in other languages

- Research gap 1:
  - No studies in a non-English language answered such questions
    - since this requires training LLMs from scratch
  - Are these conclusions English-specific or universal?

Background: issues with existing Chinese corpora

- Minimal Pair Paradigm Benchmarks in Chinese
  - CLiMP: 16 paradigms (some problematic), vocab translated from En
  - SLING: good sents from Treebank, bad sents transformed by rules
- Research Gap 2:
  - No high-quality, wide-coverage BLiMP-style dataset for Chinese
  - Infrastructure building is important for interesting research



Benchmark	Language	Size	N
BLiMP (Warstadt et al., 2020)	English	67k	67
SyntaxGym (Hu et al., 2020)	English	NA	39
CLiMP (Xiang et al., 2021)	Chinese	16k	16
SLING (Song et al., 2022)	Chinese	38k	38
JBLiMP (Someya and Oseki, 2023)	Japanese	331	39
LINDSEA (Leong et al., 2023)	Indonesian	380	38
	Tamil	200	20
RuBLiMP (Taktasheva et al., 2024)	Russian	45k	45
ZhoBLiMP (Ours)	Chinese	35k	118

Research questions

1. Have state-of-the-art LMs mastered Chinese syntax?
2. How many tokens are needed to learn Chinese syntax? How big do the models have to be (number of parameters)? --> scaling in model size and data size
3. Are there difficult Chinese syntactic phenomena impossible to learn?
4. Do learning trajectories in LMs differ from those of humans (children)? How?

Preview of our contribution and findings

- Resources
  - ZhoBLiMP, a large benchmark of linguistic minimal pairs for Chinese
  - A graph user interface for the minimal pair generation
  - 20 models trained from scratch and hundreds of checkpoints
- Findings
  - The effect of scaling diminishes after the threshold
    - Model size: 500M parameters
    - Training data size: 1B tokens
  - Even the best model still fail on three phenomena:
    - Anaphor, Quantifiers, and Ellipsis
  - A surge in performance is observed between 100M and 1B tokens
    - With a U-shaped learning pattern

Creation of ZhoBLiMP

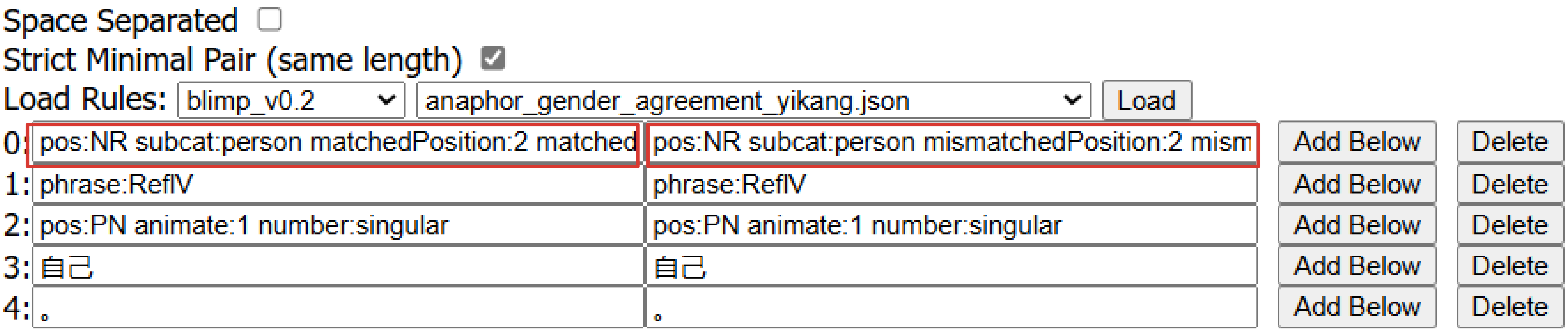
Creation of ZhoBLiMP

- **Interface**
  - First build a GUI that can generate minimal pairs given grammar templates of a minimal pair and a vocabulary.
- **Grammar template**
  - Eight linguists manually write the grammar templates for minimal pairs extracted from the literature of theoretical linguistics.
- **Vocabulary**
  - Annotated with necessary features.
- Data generation
  - Generate 300 minimal pairs per paradigm.

The benchmark is **controlled** and **extendable**!

Creation of ZhoBLiMP: platform & grammar template

A web interface to craft grammar templates and generate minimal pairs



Space Separated □

Strict Minimal Pair (same length) ☑

Load Rules: blimp_v0.2 | anaphor_gender_agreement_yikang.json

0:	pos:NR subcat:person matchedPosition:2 matched	pos:NR subcat:person mismatchedPosition:2 mism
1:	phrase:ReflV	phrase:ReflV
2:	pos:PN animate:1 number:singular	pos:PN animate:1 number:singular
3:	自己	自己
4:	。	。

王先生非常喜欢他自己。 v.s. *王先生非常喜欢她自己。

pos:NR subcat:person (mis)matchedPosition:2 matchedProperties: gender

- Lexical
  - Assign values to certain lexical properties ; searched in the vocabulary.
- Direct
  - Directly used in the composition of sentences.
- (mis)Matched
  - Assign (dis)agreement in one lexical property between two positions.
- Phrase
  - A pre-defined phrase.

Creation of ZhoBLiMP: grammar template

- Sources of the minimal pairs:

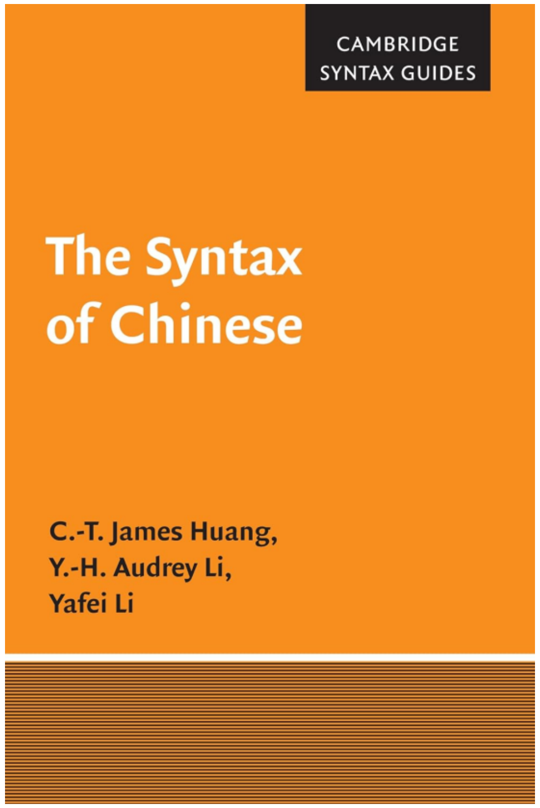
(1) examples in a syntax textbook on Chinese—*The Syntax of Chinese* (Huang et al., 2009)

(2) BLiMP (Warstadt et al., 2020)

(3) journal articles on Chinese syntax and linguistics

Roughly 130 paradigms before human validation.

You can add your own!



CAMBRIDGE
SYNTAX GUIDES

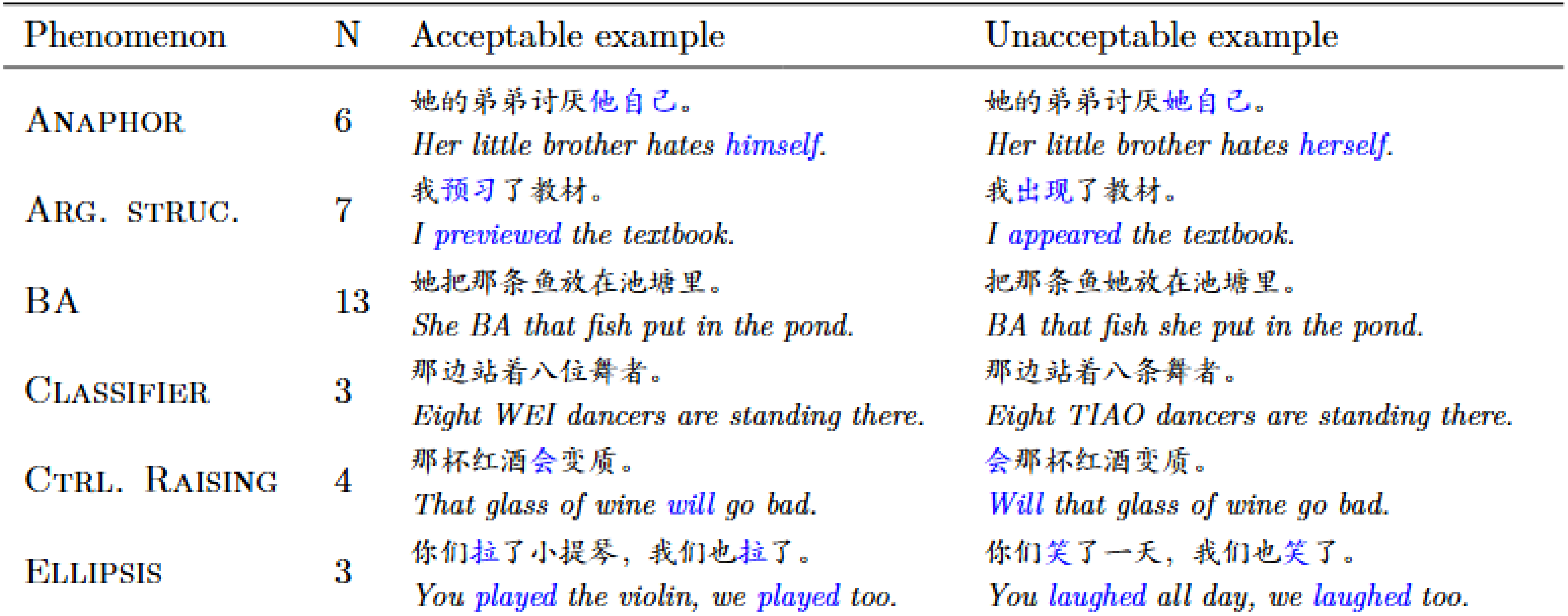
The Syntax
of Chinese

C.-T. James Huang,
Y.-H. Audrey Li,
Yafei Li

Creation of ZhoBLiMP: summary of the paradigms

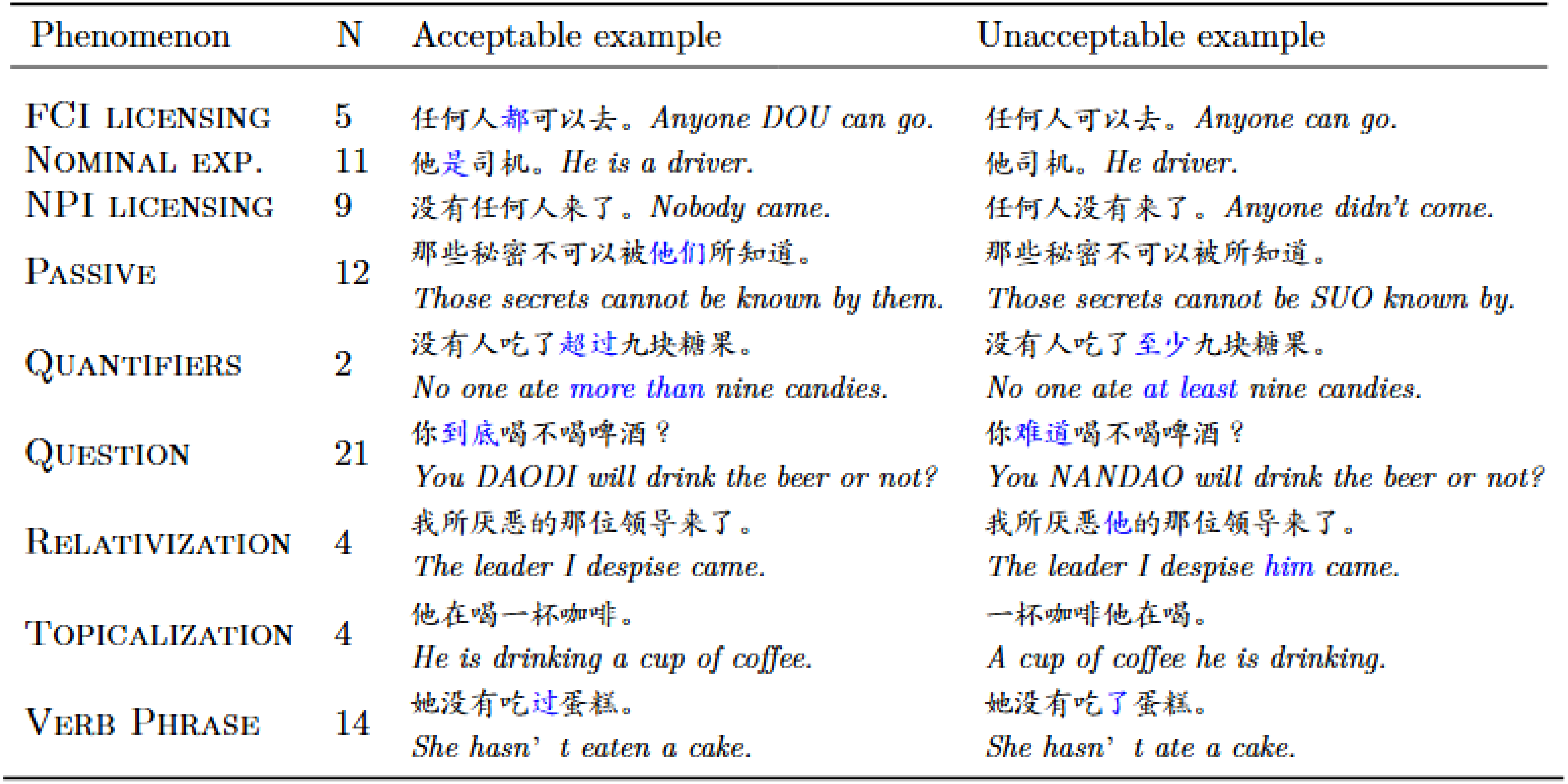
After two months...

100+ paradigms x 300 minimal pairs/paradigm, 15 phonmena



Phenomenon	N	Acceptable example	Unacceptable example
Anaphor	6	她的弟弟讨厌他自己。 *Her little brother hates himself.*	她的弟弟讨厌她自己。 *Her little brother hates herself.*
Arg. struc.	7	我预习了教材。 *I previewed the textbook.*	我出现了教材。 *I appeared the textbook.*
BA	13	她把那条鱼放在池塘里。 *She BA that fish put in the pond.*	把那条鱼她放在池塘里。 *BA that fish she put in the pond.*
Classifier	3	那边站着八位舞者。 *Eight WEI dancers are standing there.*	那边站着八条舞者。 *Eight TIAO dancers are standing there.*
Ctrl. Raising	4	那杯红酒会变质。 *That glass of wine will go bad.*	会那杯红酒变质。 *Will that glass of wine go bad.*
Ellipsis	3	你们拉了小提琴，我们也拉了。 *You played the violin, we played too.*	你们笑了一天，我们也笑了。 *You laughed all day, we laughed too.*

Creation of ZhoBLiMP: summary of the paradigms

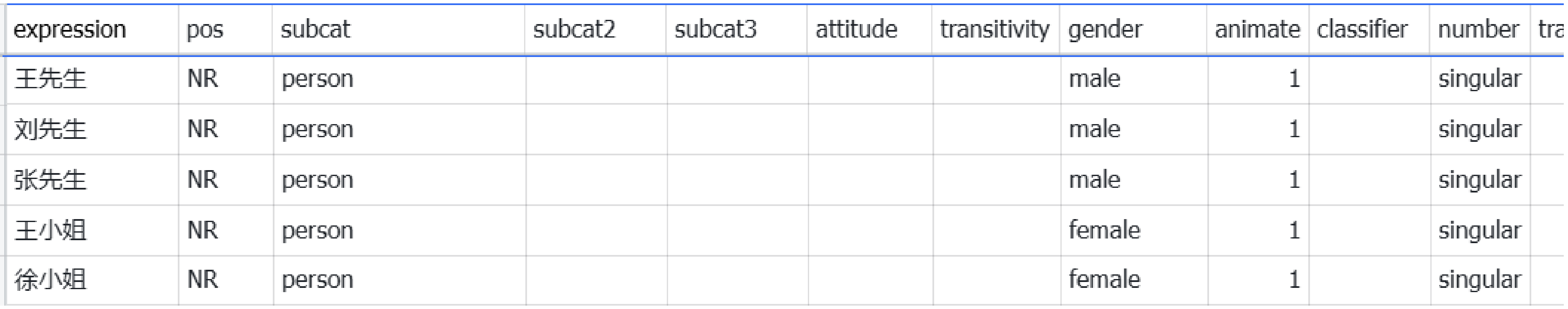


Phenomenon	N	Acceptable example	Unacceptable example
FCI licensing	5	任何人都可以去。*Anyone DOU can go.*	任何人可以去。*Anyone can go.*
Nominal exp.	11	他是司机。*He is a driver.*	他司机。*He driver.*
NPI licensing	9	没有任何人来了。*Nobody came.*	任何人没有来了。*Anyone didn't come.*
Passive	12	那些秘密不可以被他们所知道。 *Those secrets cannot be known by them.*	那些秘密不可以被所知道。 *Those secrets cannot be SUO known by.*
Quantifiers	2	没有人吃了超过九块糖果。 *No one ate more than nine candies.*	没有人吃了至少九块糖果。 *No one ate at least nine candies.*
Question	21	你到底喝不喝啤酒？ *You DAODI will drink the beer or not?*	你难道喝不喝啤酒？ *You NANDAO will drink the beer or not?*
Relativization	4	我所厌恶的那位领导来了。 *The leader I despise came.*	我所厌恶他的那位领导来了。 *The leader I despise him came.*
Topicalization	4	他在喝一杯咖啡。 *He is drinking a cup of coffee.*	一杯咖啡他在喝。 *A cup of coffee he is drinking.*
Verb Phrase	14	她没有吃过蛋糕。 *She hasn’ t eaten a cake.*	她没有吃了蛋糕。 *She hasn’ t ate a cake.*

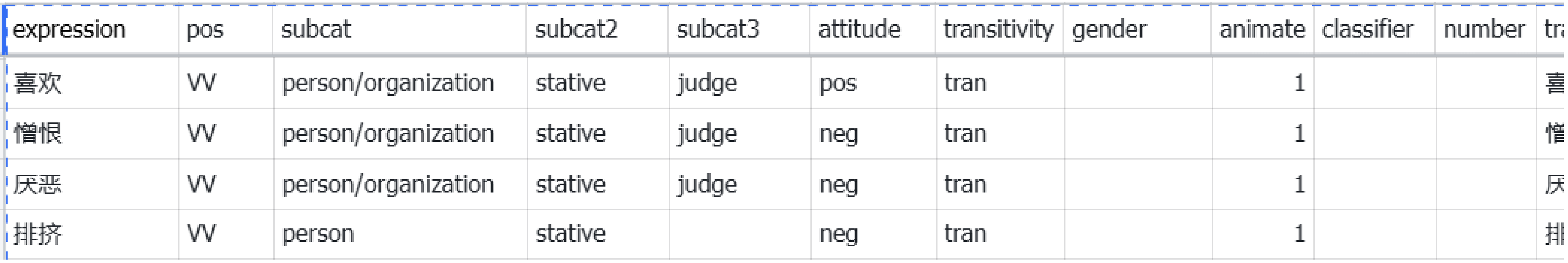
Creation of ZhoBLiMP - vocabulary

王先生非常喜欢他自己。v.s. *王先生非常喜欢她自己。

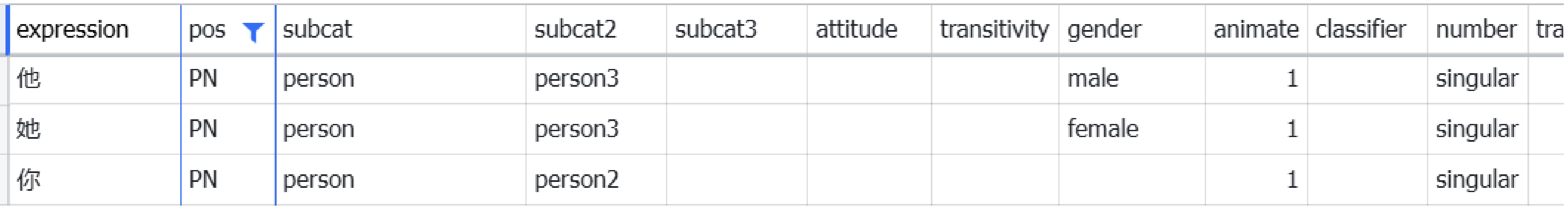
Mr. Wang likes himself. v.s. *Mr. Wang likes herslf.



expression	pos	subcat	subcat2	subcat3	attitude	transitivity	gender	animate	classifier	number	tra
王先生	NR	person					male	1		singular	
刘先生	NR	person					male	1		singular	
张先生	NR	person					male	1		singular	
王小姐	NR	person					female	1		singular	
徐小姐	NR	person					female	1		singular	



expression	pos	subcat	subcat2	subcat3	attitude	transitivity	gender	animate	classifier	number	tra
喜欢	VV	person/organization	stative	judge	pos	tran		1			喜
憎恨	VV	person/organization	stative	judge	neg	tran		1			憎
厌恶	VV	person/organization	stative	judge	neg	tran		1			厌
排挤	VV	person	stative		neg	tran		1			排

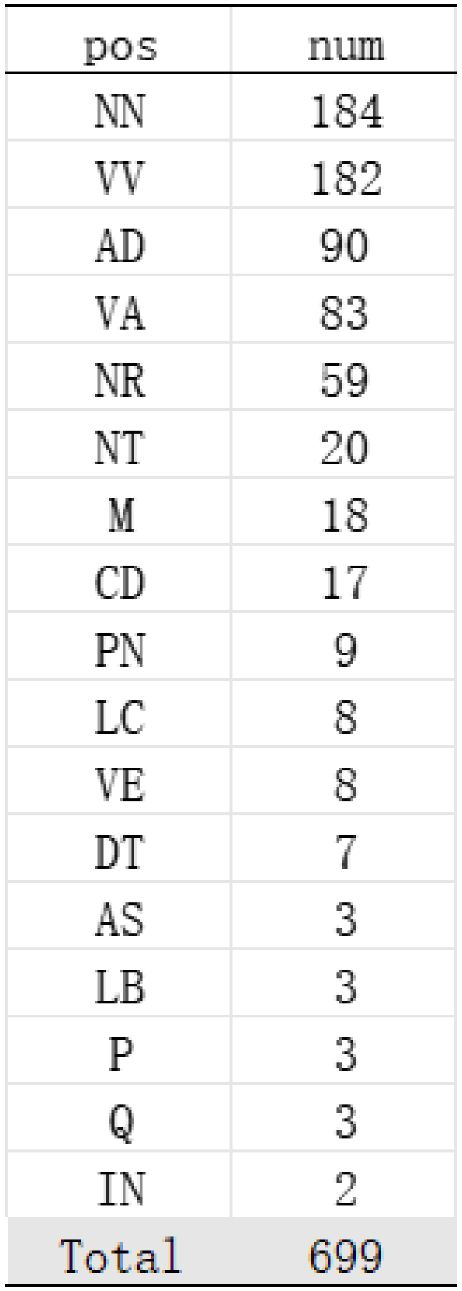


expression	pos	subcat	subcat2	subcat3	attitude	transitivity	gender	animate	classifier	number	tra
他	PN	person	person3				male	1		singular	
她	PN	person	person3				female	1		singular	
你	PN	person	person2					1		singular	

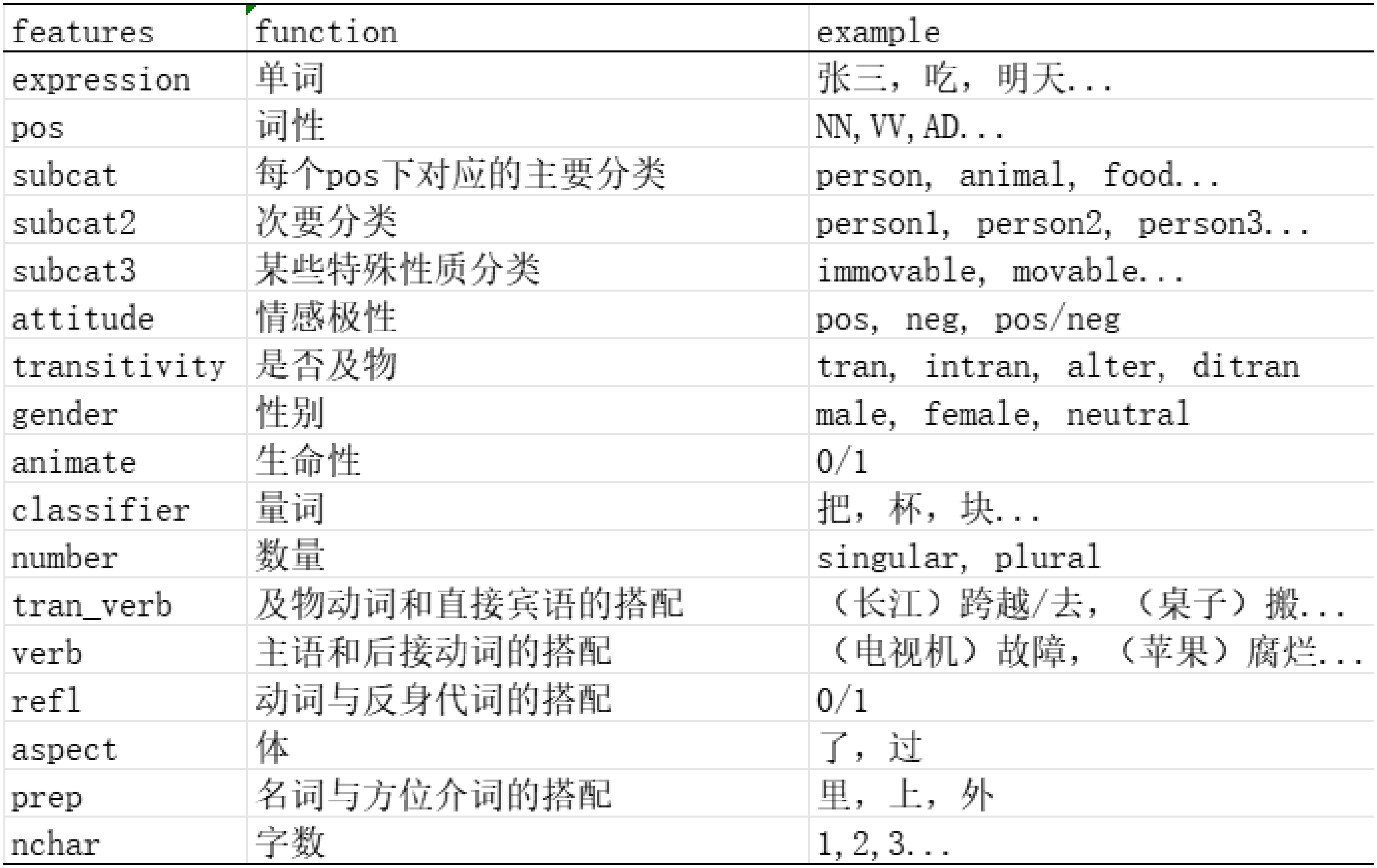
Generated Sentences

Good Sentences	Bad Sentences
张先生憎恨他自己。	王小姐憎恨他自己。
刘先生比较反感他自己。	何太太比较反感他自己。
张先生厌恶他自己。	郑大妈厌恶他自己。
冯大哥比较埋怨他自己。	吴太太比较埋怨他自己。
王大娘憎恨她自己。	赵大爷憎恨她自己。
李先生安慰了他自己。	徐小姐安慰了他自己。
刘先生有点支持他自己。	郑大妈有点支持他自己。
宋女士埋怨她自己。	李先生埋怨她自己。
王小姐夸奖了她自己。	赵大爷夸奖了她自己。
徐小姐比较尊重她自己。	杨大哥比较尊重她自己。

Creation of ZhoBLiMP - summary of the vocab



pos	num
NN	184
VV	182
AD	90
VA	83
NR	59
NT	20
M	18
CD	17
PN	9
LC	8
VE	8
DT	7
AS	3
LB	3
P	3
Q	3
IN	2
Total	699



features	function	example
expression	单词	张三，吃，明天...
pos	词性	NN,VV,AD...
subcat	每个pos下对应的主要分类	person，animal，food...
subcat2	次要分类	person1，person2，person3...
subcat3	某些特殊性质分类	immovable，movable...
attitude	情感极性	pos，neg，pos/neg
transitivity	是否及物	tran，intran，alter，ditran
gender	性别	male，female，neutral
animate	生命性	0/1
classifier	量词	把，杯，块...
number	数量	singular，plural
tran_verb	及物动词和直接宾语的搭配	（长江）跨越/去，（桌子）搬...
verb	主语和后接动词的搭配	（电视机）故障，（苹果）腐烂...
refl	动词与反身代词的搭配	0/1
aspect	体	了，过
prep	名词与方位介词的搭配	里，上，外
nchar	字数	1,2,3...

Human validation

Randomly sampled 5 minimal pairs from each paradigm.

50 native speakers.

Forced-choice task: which sentence is more natural.

Removed about 10 paradigms where human accuracy <70%

Mean accuracy from human: 94%

116 paradigms, each with 300 minimal pairs, grouped into 15 phenomena

Experimental setup

Experimental Setup

- Evaluation metric
  - **Mean Log-probability (MLP)**: sentence-level
  - MLP(good) > MLP(bad) => correct judgement

Experimental Setup: models

- Only pretrained, no SFT
- Our model (full control)
  - 20 Pythia models we trained from scratch
  - 5 model sizes: 14M, 70M, 160M, 410M, and 1.4B
  - 4 corpus sizes: 10M, 100M, 1B, and 3B tokens
    - Trained on Books we collected
  - Use Pythia (EleutherAI 2023) default configurations: GPT-NeoX
    - Easier to compare with English Pythia models
  - Note: we use next-word-prediction to train models, not ZhoBLiMP
- Industry-level model (partially open-source)
  - Trained on multilingual text, code, web data, etc. (not open)
  - **Qwen2.5**: 0.5B, 1.5B, 7B, 14B, 32B model size (Alibaba)
  - **InternLM2.5**: 1.8B, 7B, 20B (Shanghai AI Lab)
  - **Yi1.5**: 6B, 9B, 34B (01)
  - **Gemma2**: 2B, 9B, 27B (Google)

Pretraining data for our model

8943 Chinese books.

- history, fiction, popular science, etc.
- higher quality than web data
- **c.a. 3Billion tokens** by the Chinese-Llama tokenizer.

- c.f. **>1Trillion tokens** for others

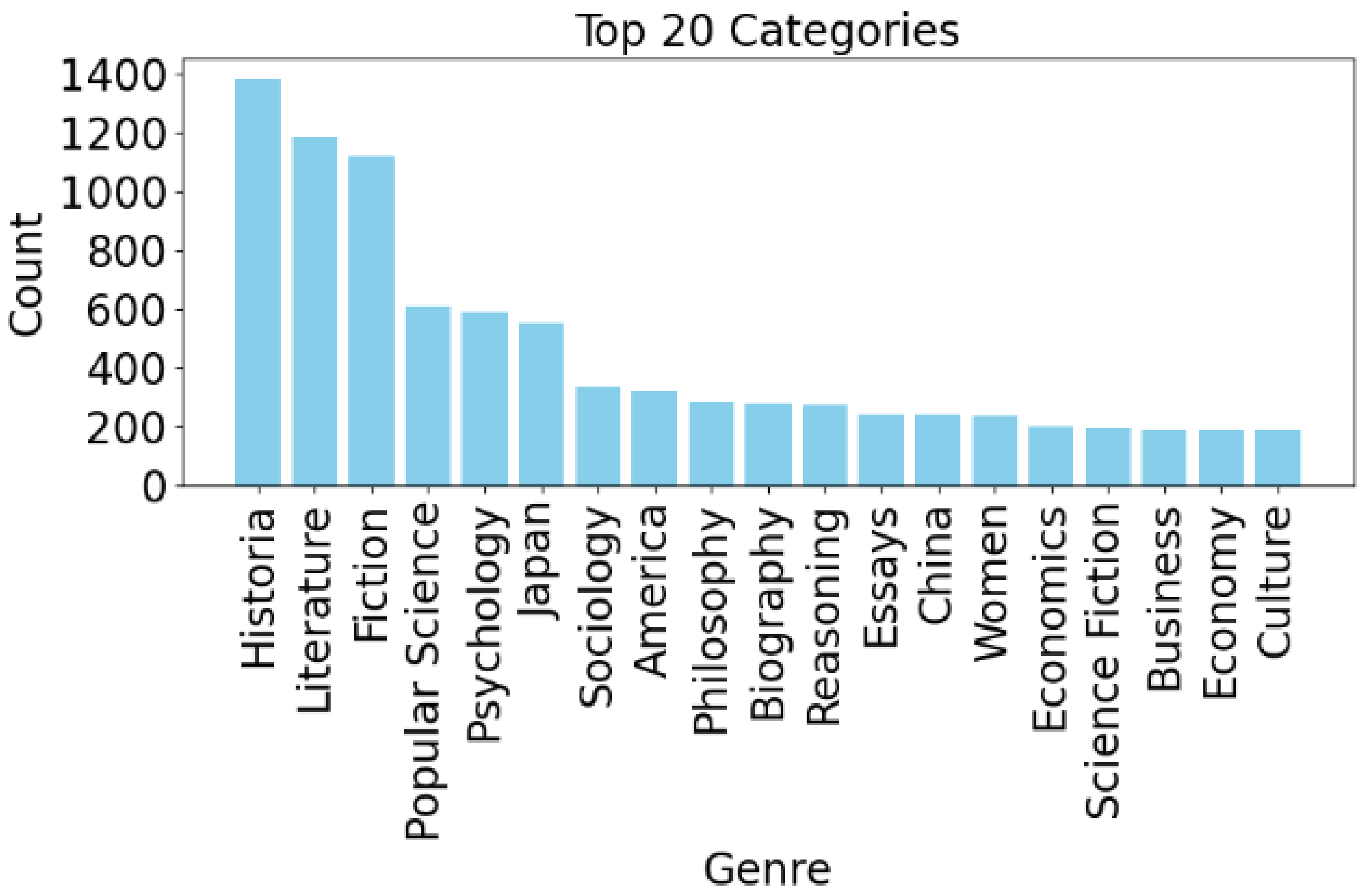


Fig. Top 20 Categories in Training Data

Preprocessing:

1. Removed: books containing too much **non-Chinese** tokens

   e.g. Programming Tutorials, such as《C和C++游戏趣味编程》

1. Removed: **Copyright pages** + **tables of contents**

   i.e. ~~2019年1月第1版 2019年1月第1次印刷~~

   ~~定价：46.00元~~

1. **Deduplication** by *MinLSH* (theshold=0.87)

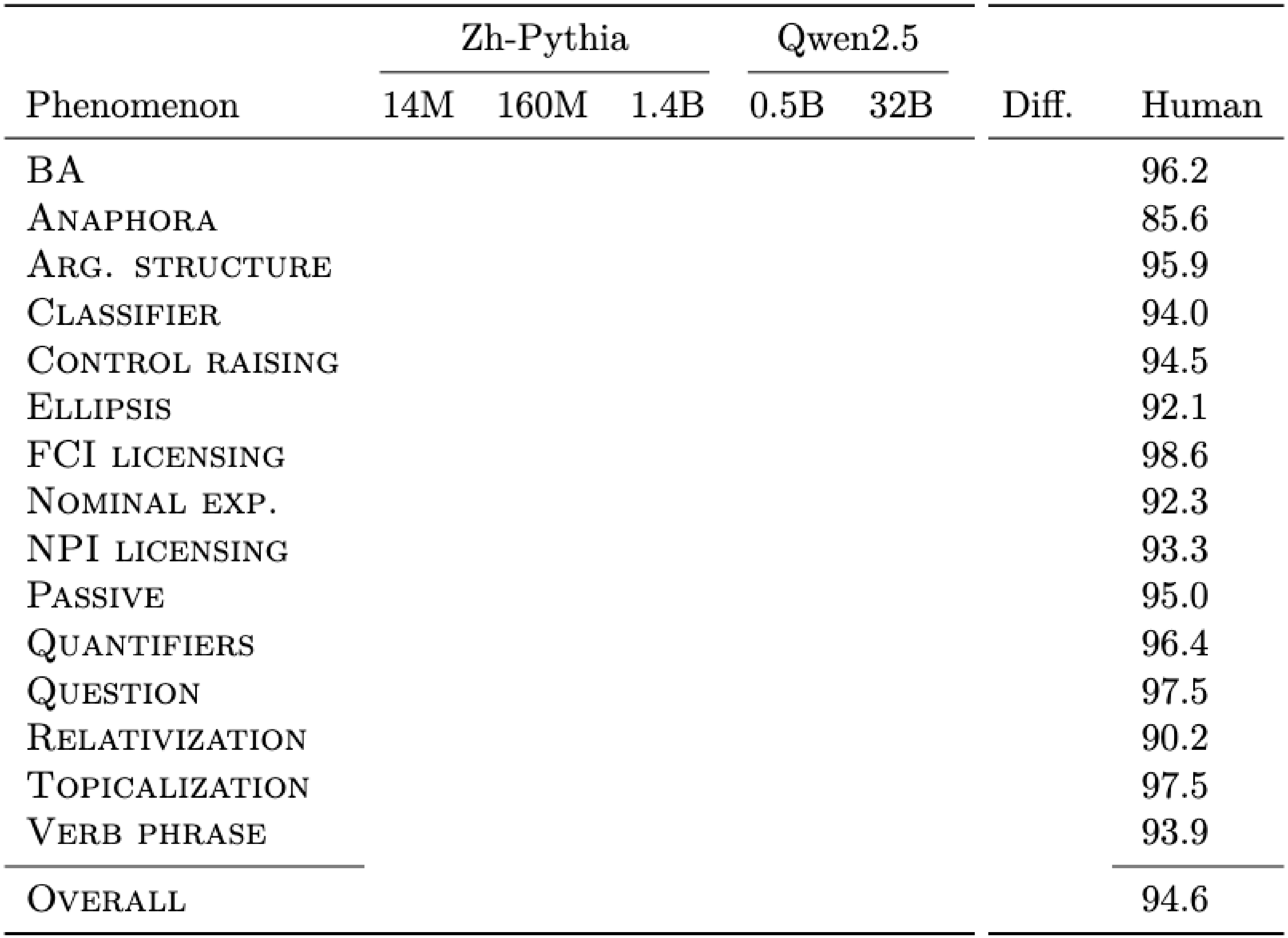
1. Tokenized using **Chinese-Llama** (Cui et al 2023)

_王	姨	把	货	箱	放	满	了	玻璃	珠	。

Example: A sentence tokenzied by the Chinese-Llama tokenizer

Results and discussion

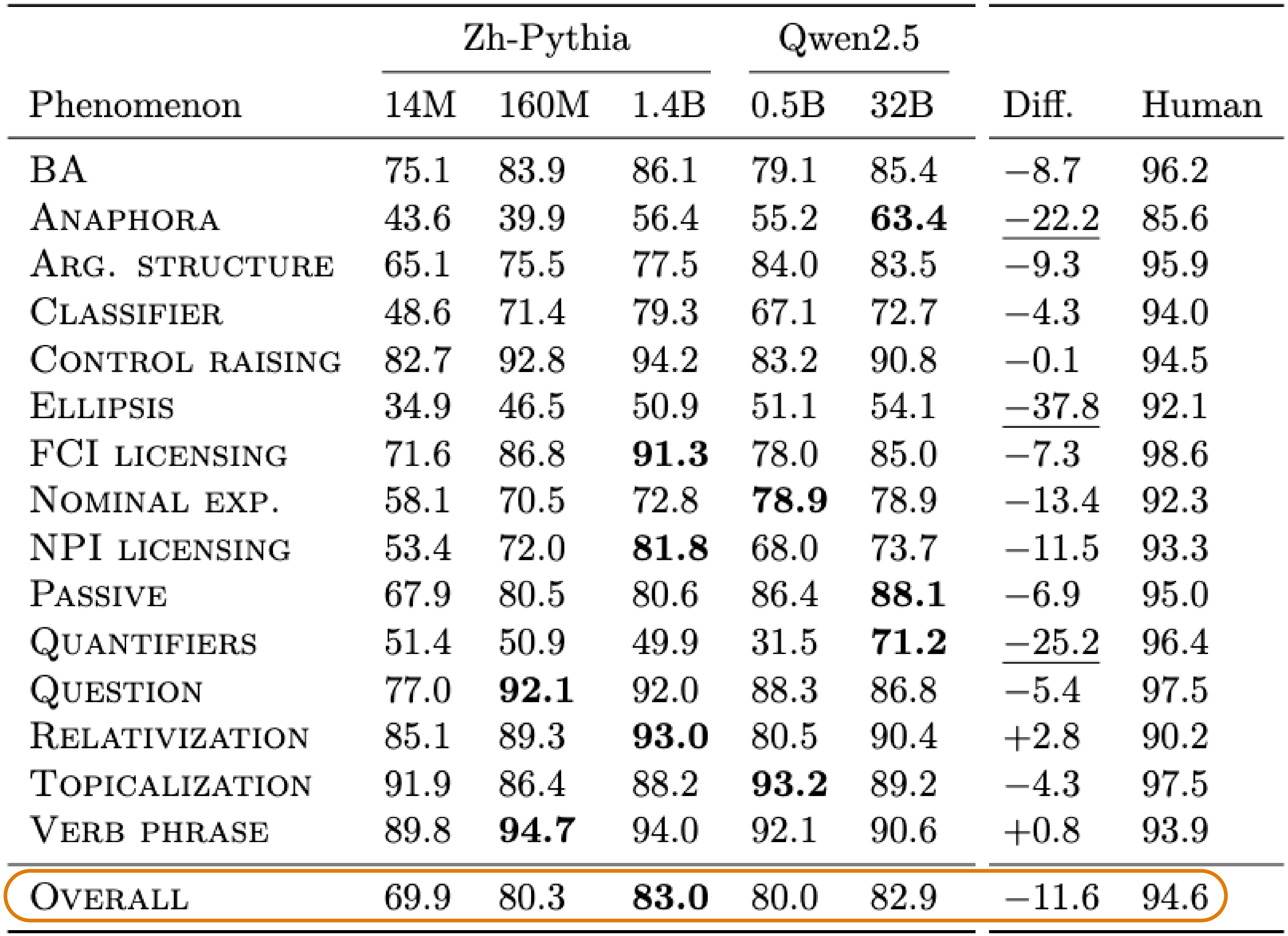
Overall performance



	Zh-Pythia			Qwen2.5			
Phenomenon	14M	160M	1.4B	0.5B	32B	Diff.	Human
BA							96.2
ANAPHORA							85.6
ARG. STRUCTURE							95.9
CLASSIFIER							94.0
CONTROL RAISING							94.5
ELLIPSIS							92.1
FCI LICENSING							98.6
NOMINAL EXP.							92.3
NPI LICENSING							93.3
PASSIVE							95.0
QUANTIFIERS							96.4
QUESTION							97.5
RELATIVIZATION							90.2
TOPICALIZATION							97.5
VERB PHRASE							93.9
OVERALL							94.6

Overall performance

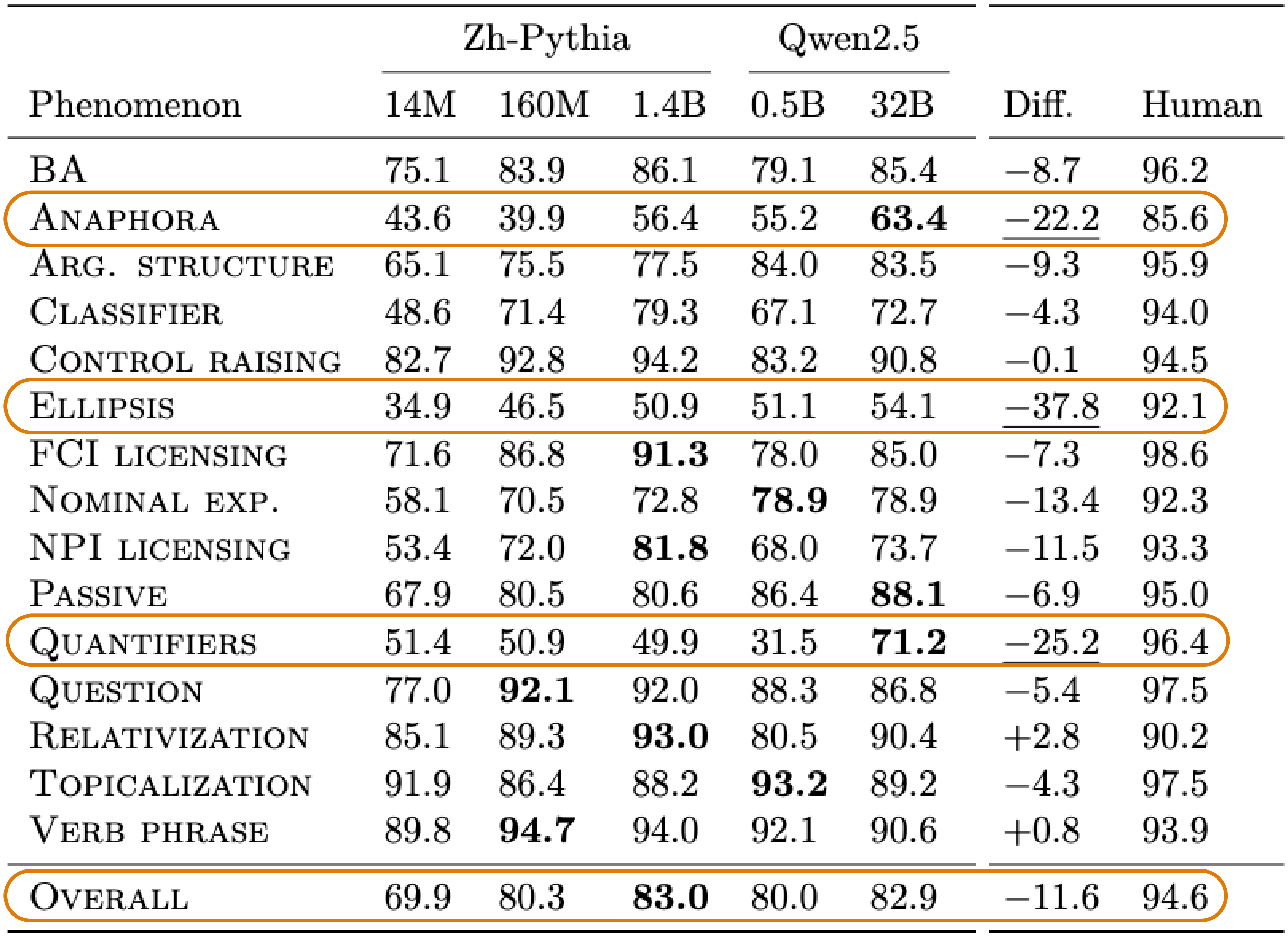
- Model performance falls between 70% and 83%, 10 points below human



	Zh-Pythia			Qwen2.5			
Phenomenon	14M	160M	1.4B	0.5B	32B	Diff.	Human
BA	75.1	83.9	86.1	79.1	85.4	−8.7	96.2
Anaphora	43.6	39.9	56.4	55.2	**63.4**	−22.2	85.6
Arg. structure	65.1	75.5	77.5	84.0	83.5	−9.3	95.9
Classifier	48.6	71.4	79.3	67.1	72.7	−4.3	94.0
Control raising	82.7	92.8	94.2	83.2	90.8	−0.1	94.5
Ellipsis	34.9	46.5	50.9	51.1	54.1	−37.8	92.1
FCI licensing	71.6	86.8	**91.3**	78.0	85.0	−7.3	98.6
Nominal exp.	58.1	70.5	72.8	**78.9**	78.9	−13.4	92.3
NPI licensing	53.4	72.0	**81.8**	68.0	73.7	−11.5	93.3
Passive	67.9	80.5	80.6	86.4	**88.1**	−6.9	95.0
Quantifiers	51.4	50.9	49.9	31.5	**71.2**	−25.2	96.4
Question	77.0	**92.1**	92.0	88.3	86.8	−5.4	97.5
Relativization	85.1	89.3	**93.0**	80.5	90.4	+2.8	90.2
Topicalization	91.9	86.4	88.2	**93.2**	89.2	−4.3	97.5
Verb phrase	89.8	**94.7**	94.0	92.1	90.6	+0.8	93.9
Overall	69.9	80.3	**83.0**	80.0	82.9	−11.6	94.6

Overall performance

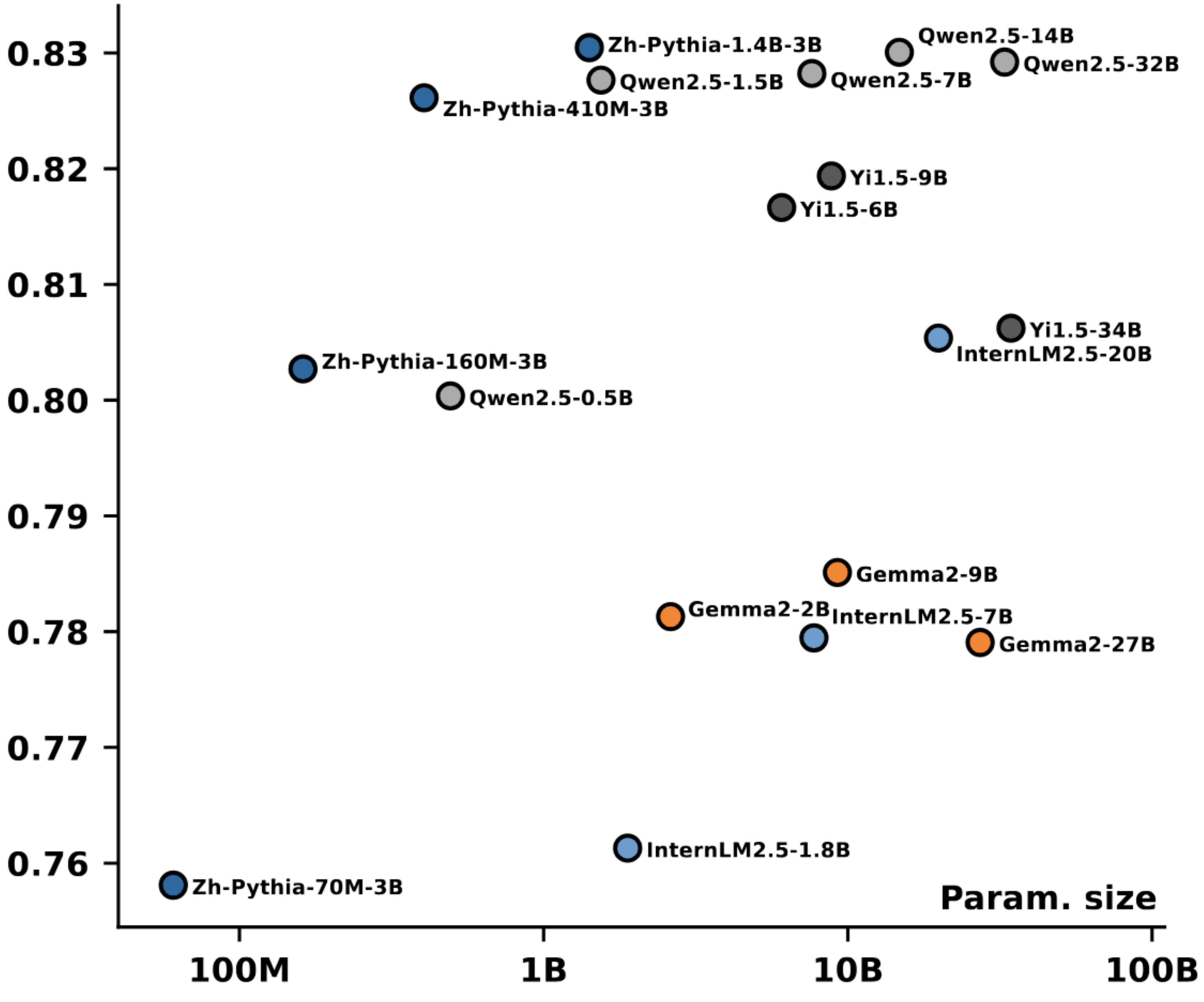
- Model performance falls between 70% and 83%, 10 points below human
- Models still fail in Anaphor, Ellipsis, and Quantifiers



	Zh-Pythia			Qwen2.5			
Phenomenon	14M	160M	1.4B	0.5B	32B	Diff.	Human
BA	75.1	83.9	86.1	79.1	85.4	−8.7	96.2
ANAPHORA	43.6	39.9	56.4	55.2	**63.4**	−22.2	85.6
ARG. STRUCTURE	65.1	75.5	77.5	84.0	83.5	−9.3	95.9
CLASSIFIER	48.6	71.4	79.3	67.1	72.7	−4.3	94.0
CONTROL RAISING	82.7	92.8	94.2	83.2	90.8	−0.1	94.5
ELLIPSIS	34.9	46.5	50.9	51.1	54.1	−37.8	92.1
FCI LICENSING	71.6	86.8	**91.3**	78.0	85.0	−7.3	98.6
NOMINAL EXP.	58.1	70.5	72.8	**78.9**	78.9	−13.4	92.3
NPI LICENSING	53.4	72.0	**81.8**	68.0	73.7	−11.5	93.3
PASSIVE	67.9	80.5	80.6	86.4	**88.1**	−6.9	95.0
QUANTIFIERS	51.4	50.9	49.9	31.5	**71.2**	−25.2	96.4
QUESTION	77.0	**92.1**	92.0	88.3	86.8	−5.4	97.5
RELATIVIZATION	85.1	89.3	**93.0**	80.5	90.4	+2.8	90.2
TOPICALIZATION	91.9	86.4	88.2	**93.2**	89.2	−4.3	97.5
VERB PHRASE	89.8	**94.7**	94.0	92.1	90.6	+0.8	93.9
OVERALL	69.9	80.3	**83.0**	80.0	82.9	−11.6	94.6

Scaling w.r.t. model size

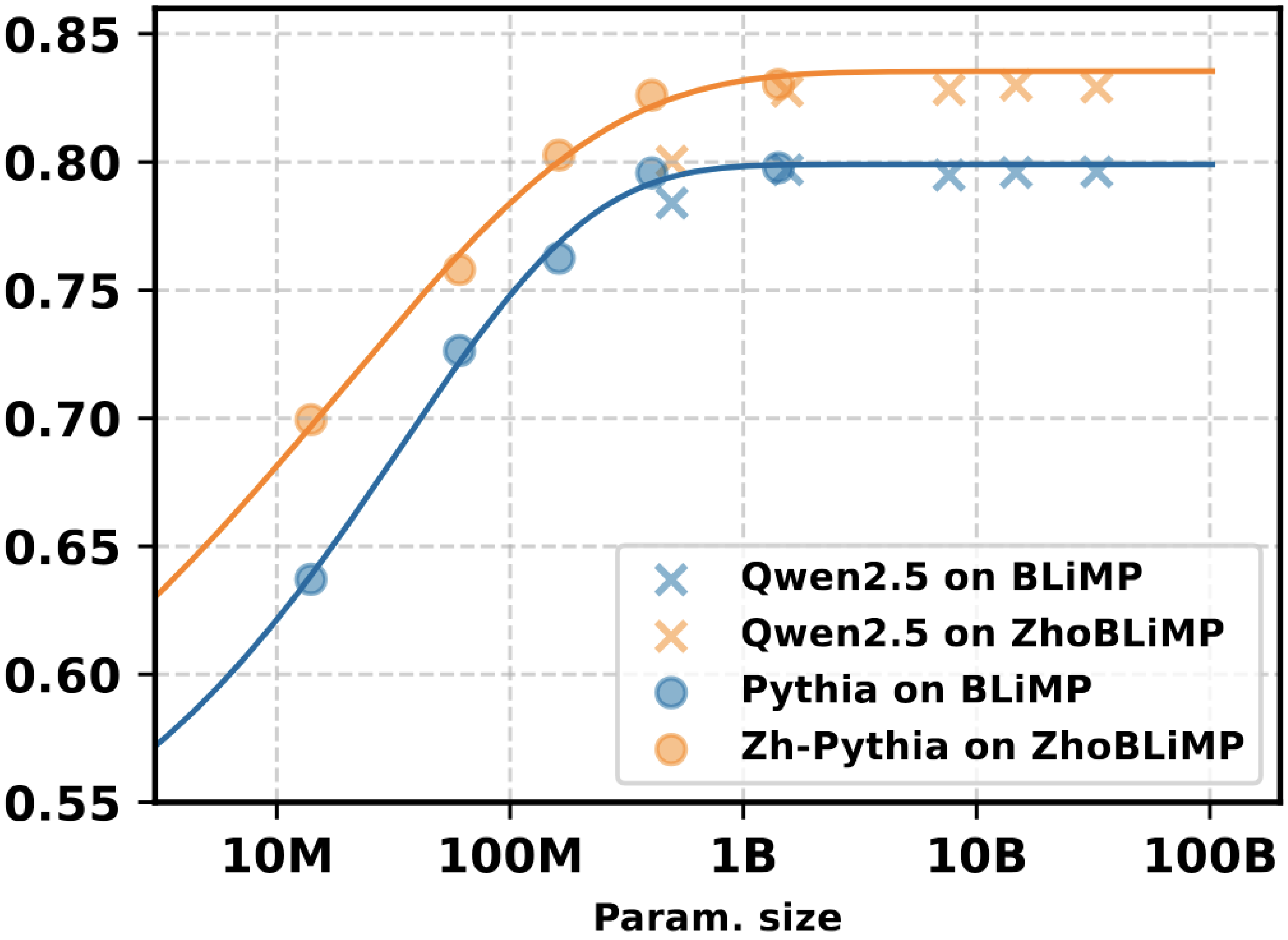
- A 160M LM can perform above 80% accuracy; syntax is mostly easy to learn
- However, there are syntactic phenomena "unlearnable" even for big LMs



(a) Performance against model parameter size

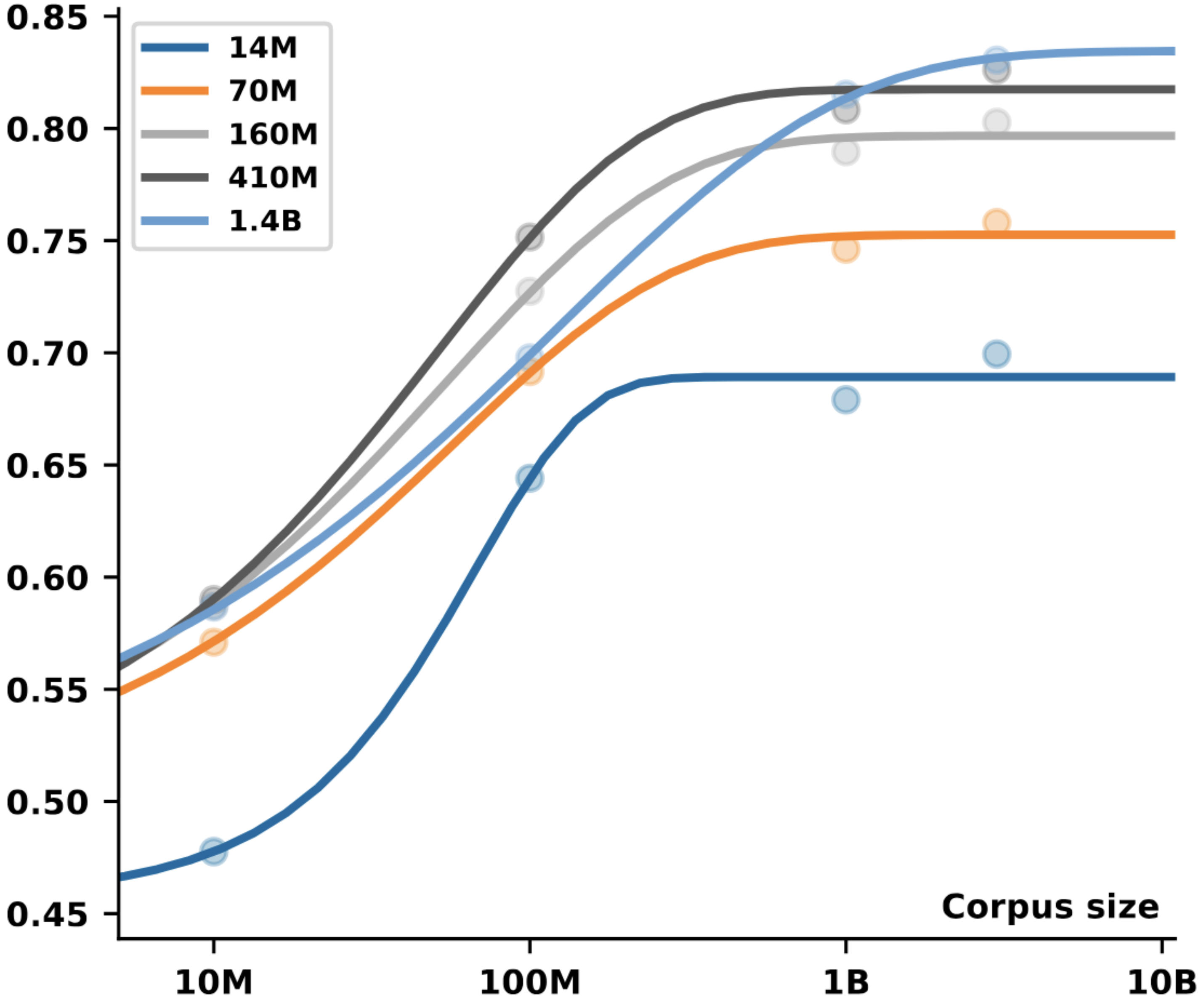
Scaling w.r.t. model size, for English and Chinese

- Little improvement after 500M/1B parameters
- Increasing model size will not help with certain phenomena



Scaling w.r.t. training data

- Benefits gradually diminish as training data increases
- Plateaus after **1B tokens** (c.f. **100M tokens for English**)
  - We trained for 1 epoch
  - Warstadt et al for 20+

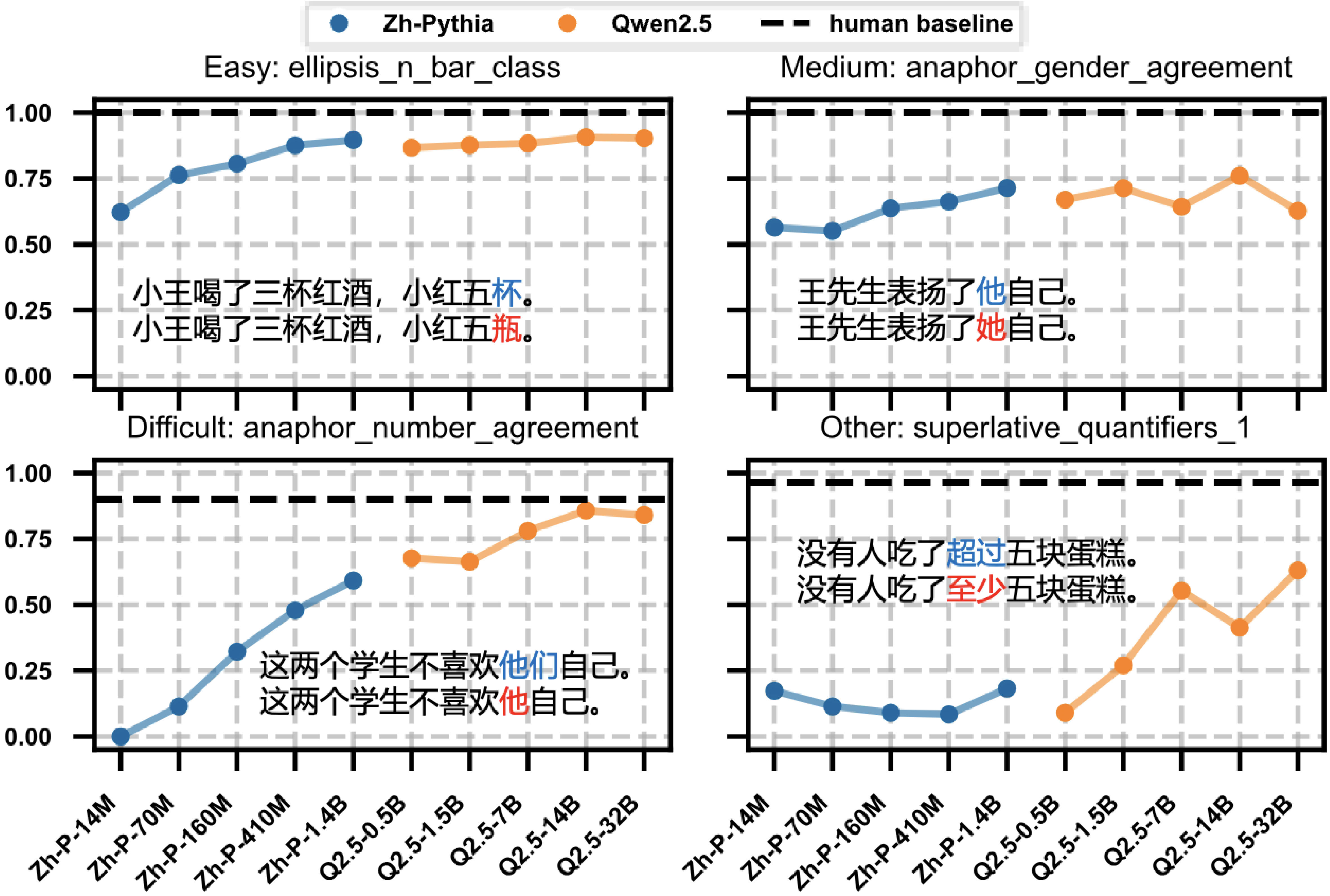


(b) Performance against training corpus size

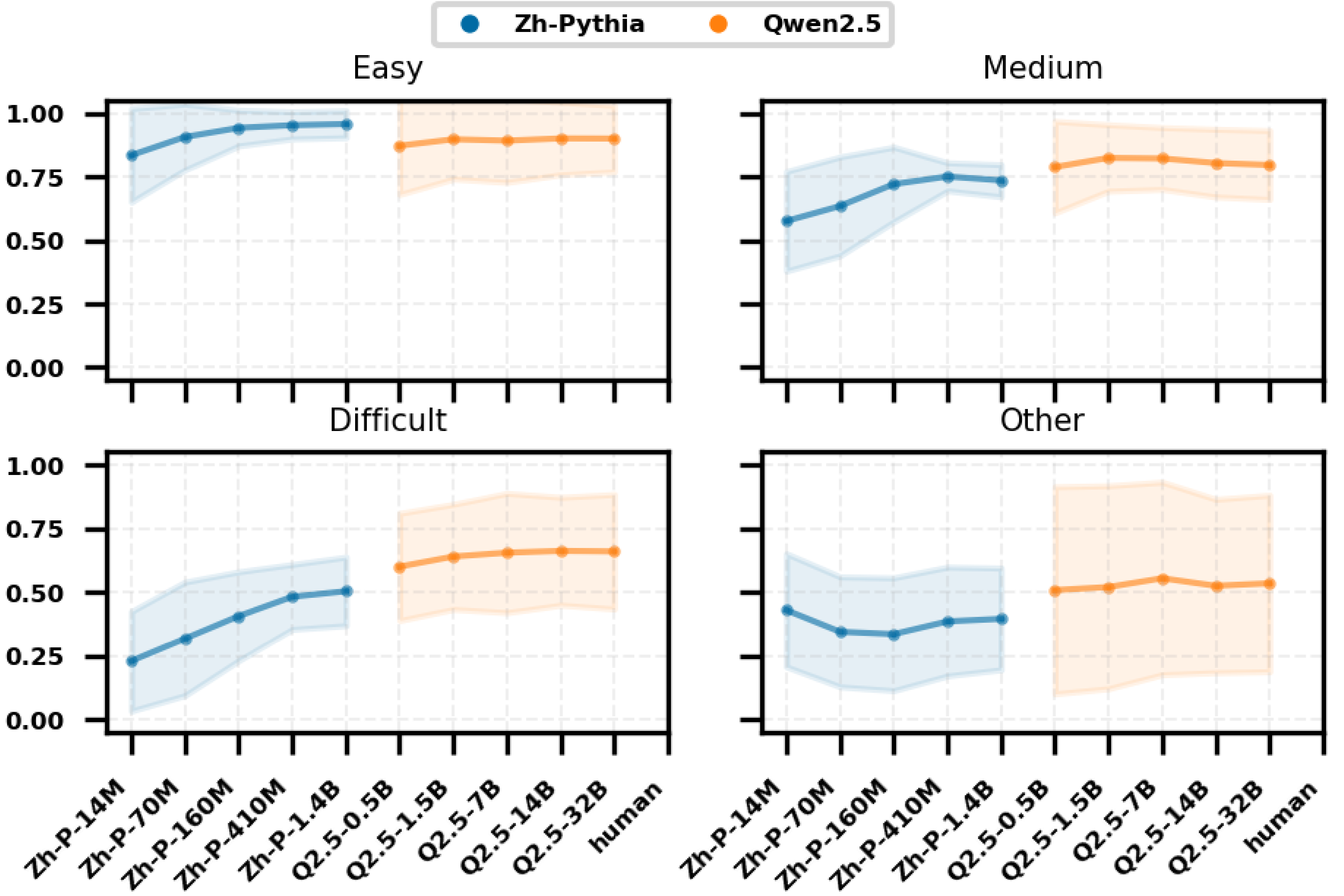
Four categories for 115 paradigms

- **Easy** (N=80): accuracy > 85%
  - Acquired easily
- **Medium** (N=15): accuracy > 70%
  - Acquired not very well, but above chance level
- **Difficult** (N=13): below 70%, but strong correlation
  - Not acquired by a small model, but might be improved
- **Other** (N=10): below 70%, and weak correlation
  - Not acquired and not sensitive to the model size

Four categories: examples



Four categories: aggregated



U-shape learning curves: previous work

For children learning irregular past tense (go -> went):

- stage 1: go -> went
- stage 2: learns the -ed rule, and over-generalize: go -> *goed
- stage 3: really learns the regular and irregular: go -> went

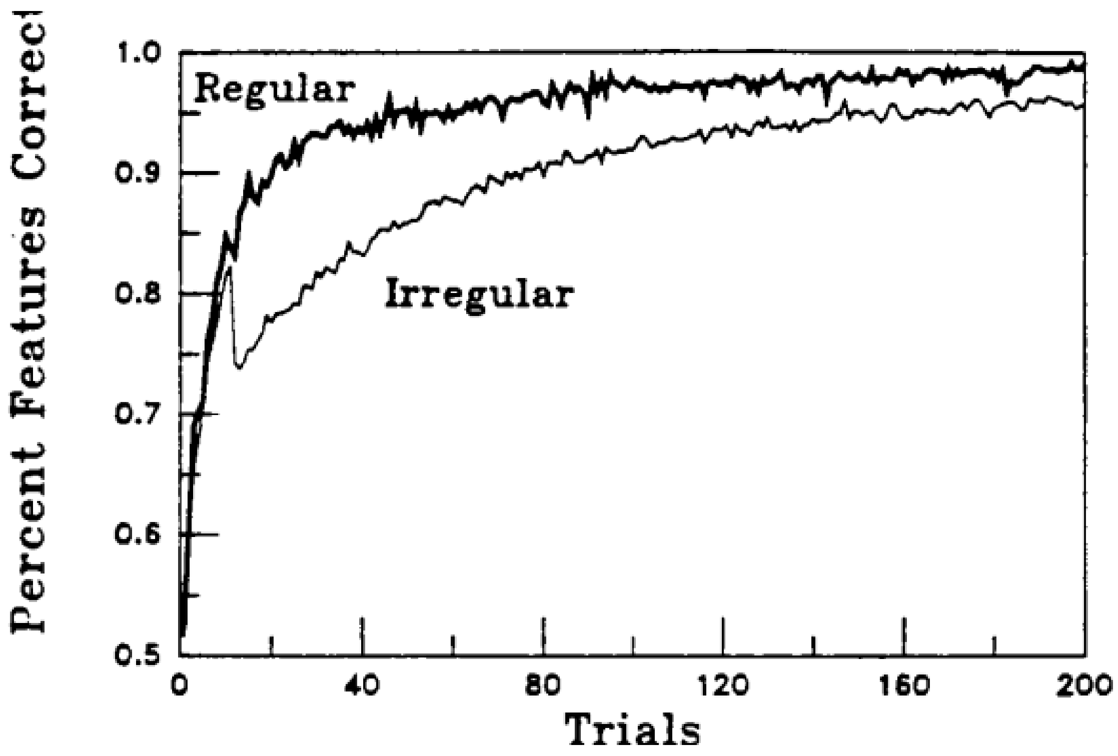
Neural nets can simulate such pattern.

U-shape learning curves: previous work

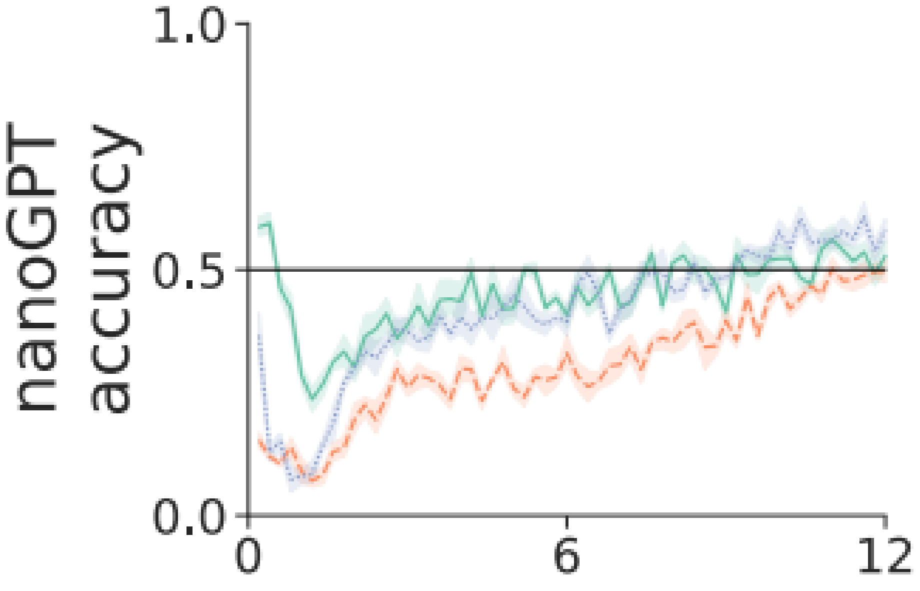
For children learning irregular past tense (go -> went):

- stage 1: go -> went
- stage 2: learns the -ed rule, and over-generalize: go -> *goed
- stage 3: really learns the regular and irregular: go -> went

Neural nets can simulate such pattern.



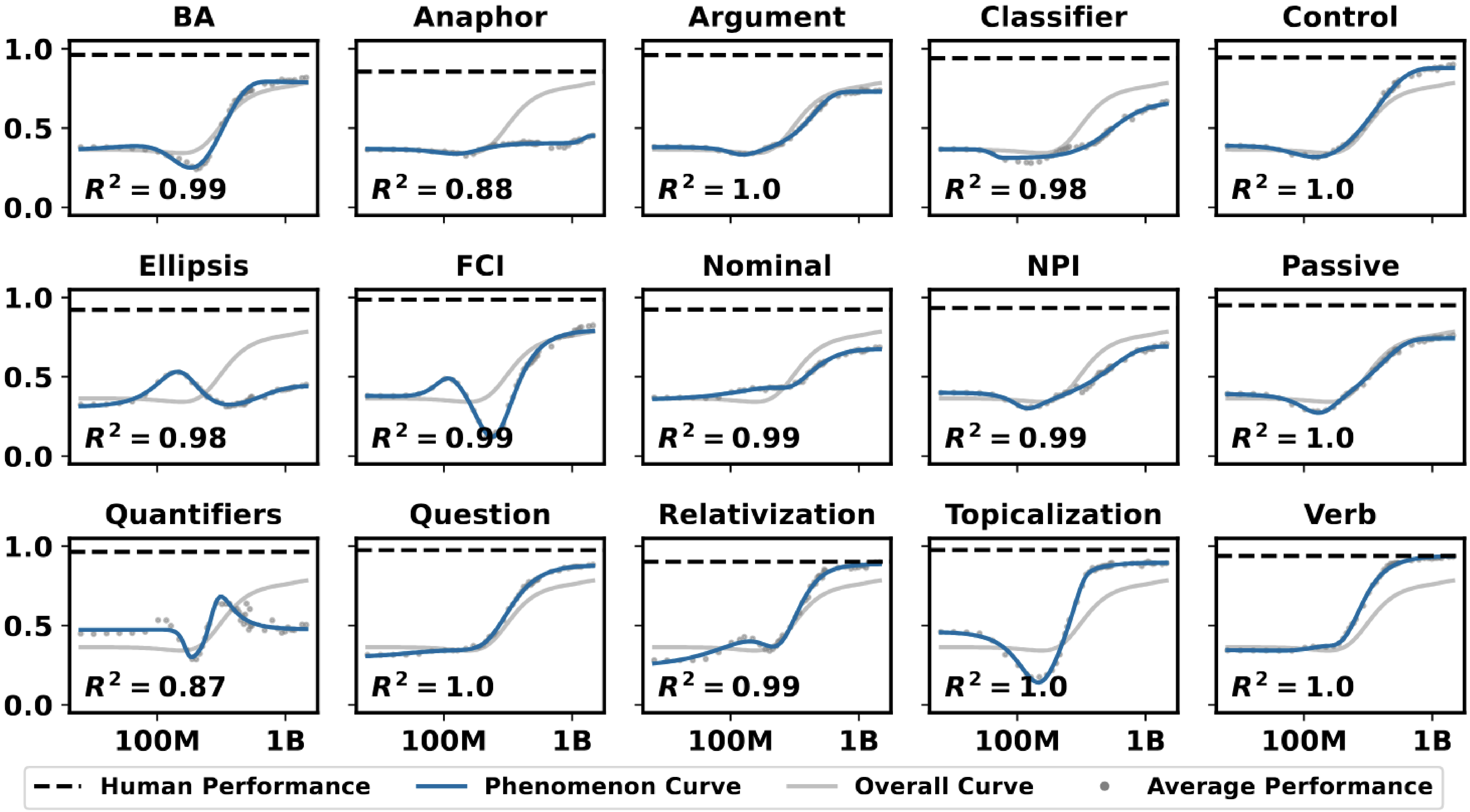
Rumelhart & McClelland 1984



Haga et al 2024; ACL Findings

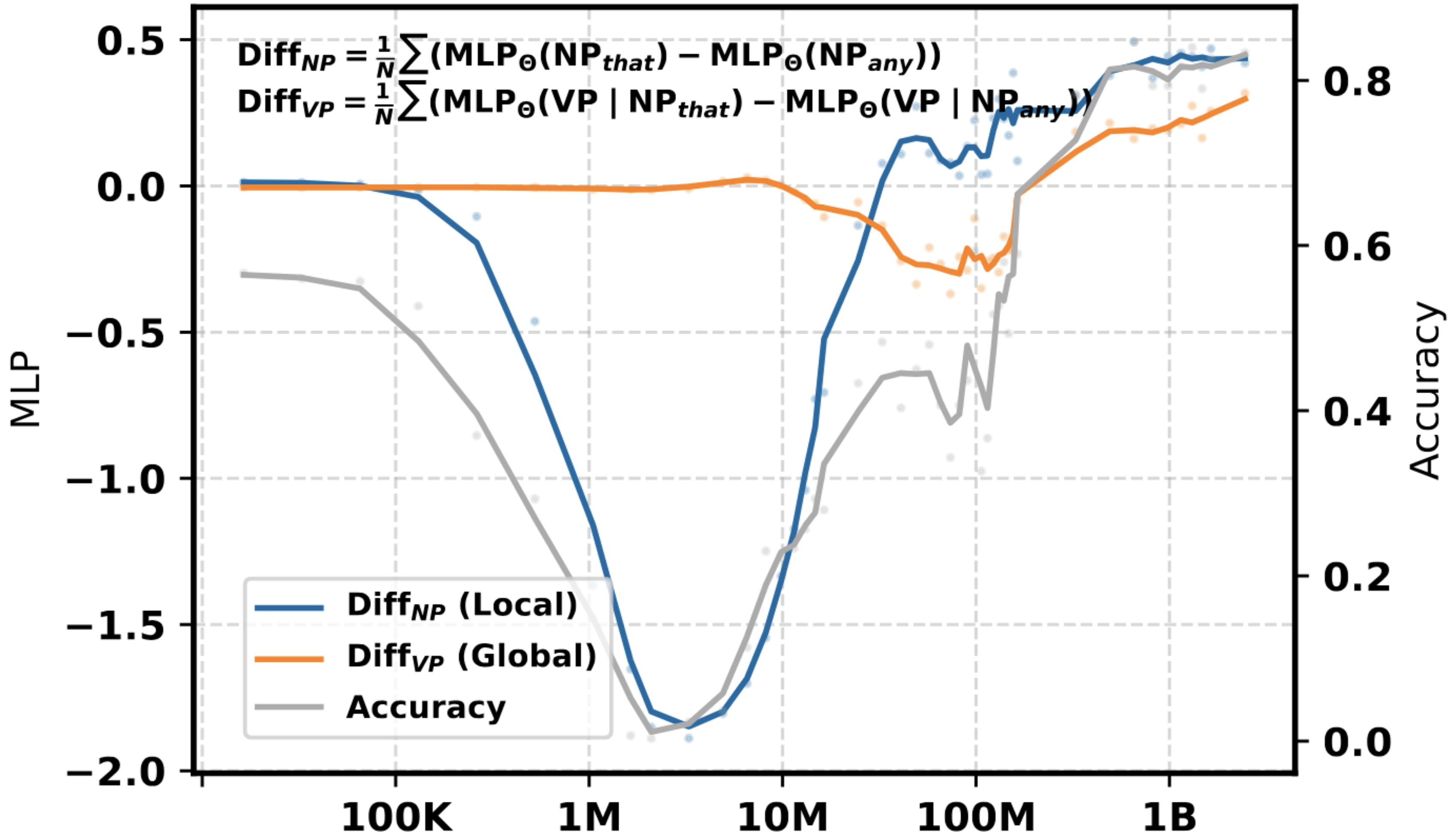
U-shape learning curves: ours

- Performance gradually saturates at 1B tokens
- Do language models over-generalize as well as children?
- First time observed in a language other than Eng, on a large scale



Case analysis of U-shape

- npi_renhe_wh_question_subj
- where "任何人" (anyone) needs to to be licensed.
  - 那个人 会做什么？ that man will do what?
  - *任何人 会做什么？ *any man will do what?
- At first, LM has some weird belief: |bos| any man > |bos| that man
- Corrected in 10M-100M tokens (probably learns: any is not licensed)



Interim conclusion

- LMs generally achieve a good performance
    - 80+% of accuracy, 10 points below human
- Scaling has limited effect on the performance on ZhoBLiMP
    - At the certain point, the benefits diminish
        - Model size at around 500M parameters
        - Training data at around 1B tokens
- But models still fail in Anaphor, Ellipsis, and Quantifiers
    - Performance for different paradigms peak at different points
    - Discourse information/pragmatic knowledge needed?
- Models exhibit U-shape learning curves
    - Models learn local features first, and then larger context

More on acceptability judgments from our lab

(1) Acceptability in Chinese:

Compare linguists' judgments in examples in journal articles with

(a) Mandarin-speakers from Beijing

(b) Cantonese-Mandarin bilinguals from Guangzhou

(Hu et al 2024; under revision)

(2) CoLAC: Corpus of Linguistic Acceptability in Chinese

Evaluate LLMs on 7k examples taken from journal articles

(Hu et al 2023; arxiv)

(3) MELA: Multilingual Evaluation of Linguistic Acceptability

Expand to 10 languages, evalute more LLMs

Cross-lingual transfer, bilingual learning, probing, etc.

(Zhang, ... Hu# 2024; ACL)

more to come...

LLMs such as BERT

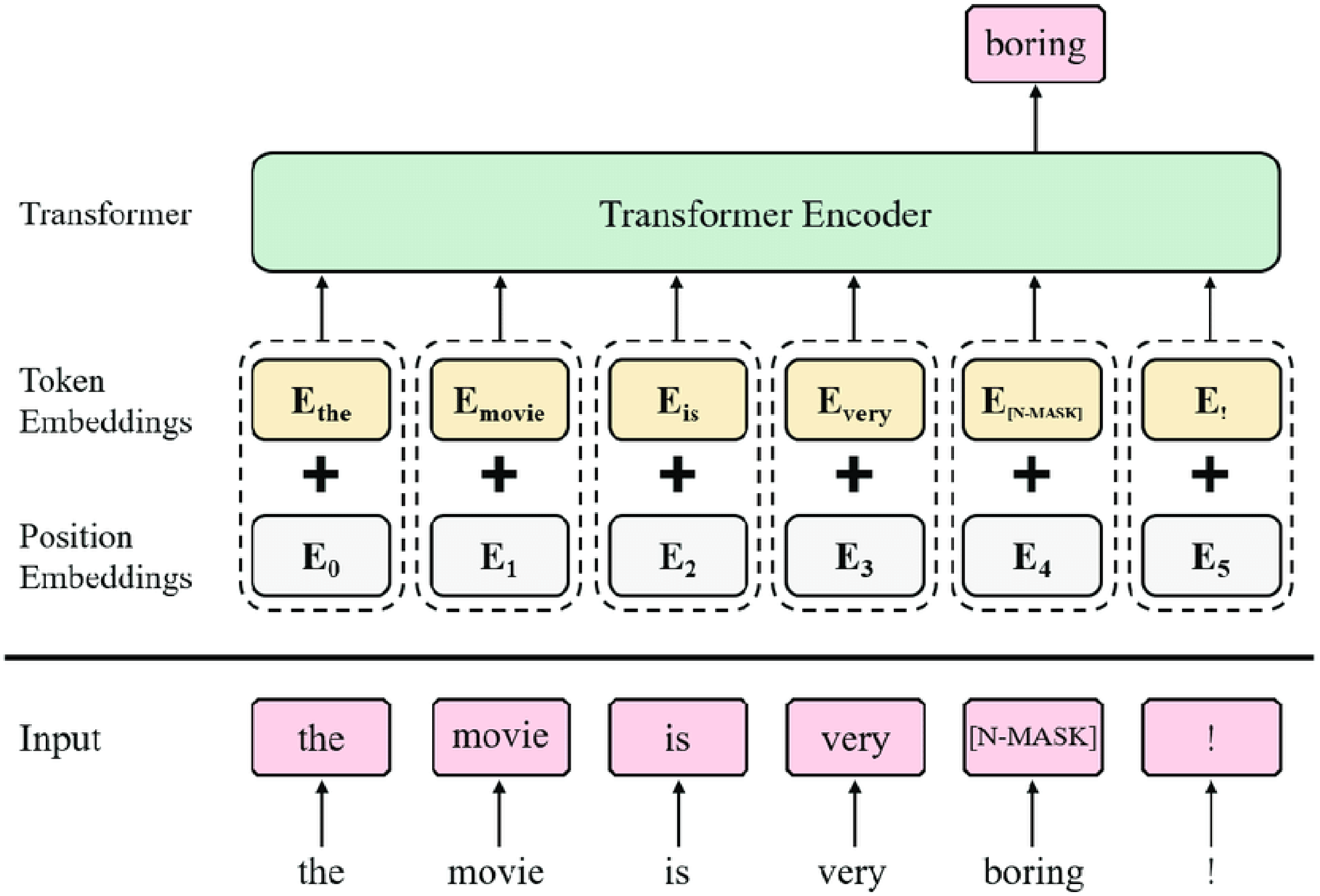
Training steps for BERT and XLM-RoBERTa:

Step 1: pretrain, masked language modeling (blank filling) in 1-100 lgs

Step 2: fine-tuning with labelled data for classification in lg A

Step 3: test in lg A, B, ...→ cross-lingual transfer

Why? Sometimes we only have human annotated data in one language (English)



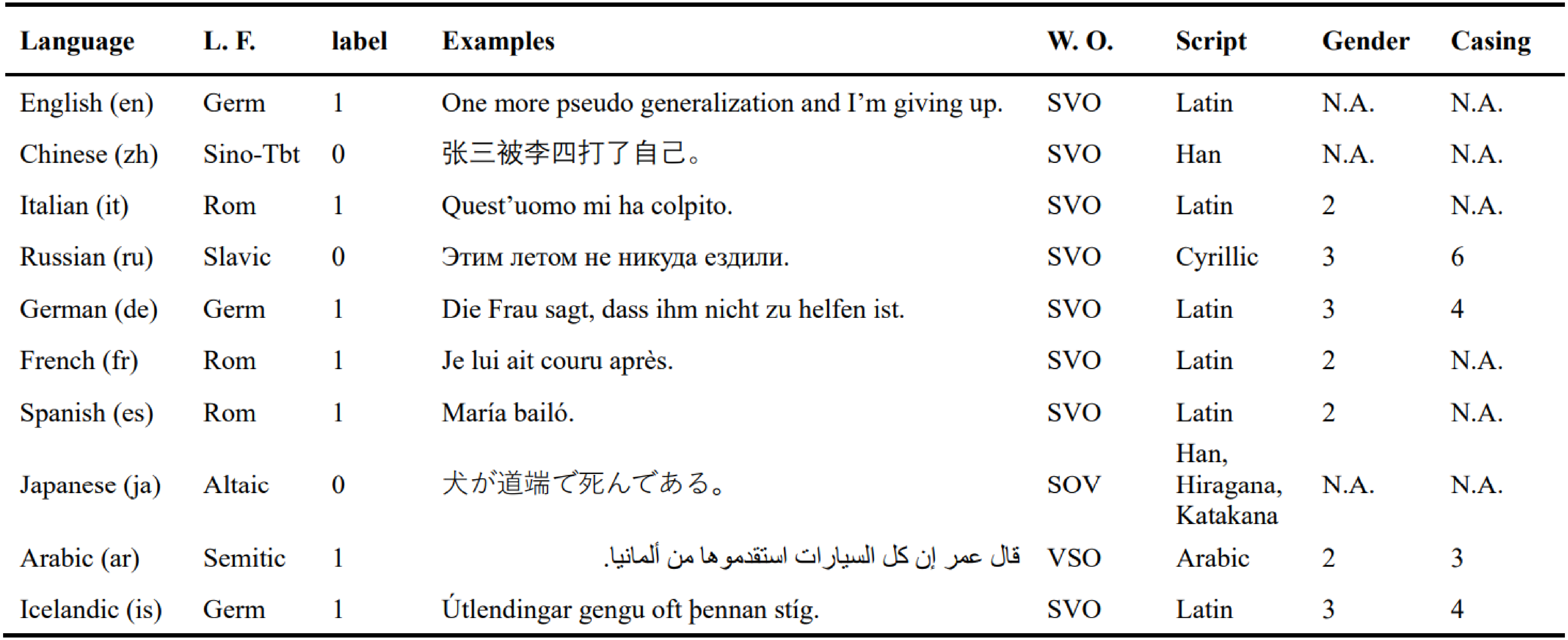
MELA: Multilingual Evaluation of Linguistic Acceptability

46k sentences in 10 lgs, collected from *Syntax of L* books, or previous work.

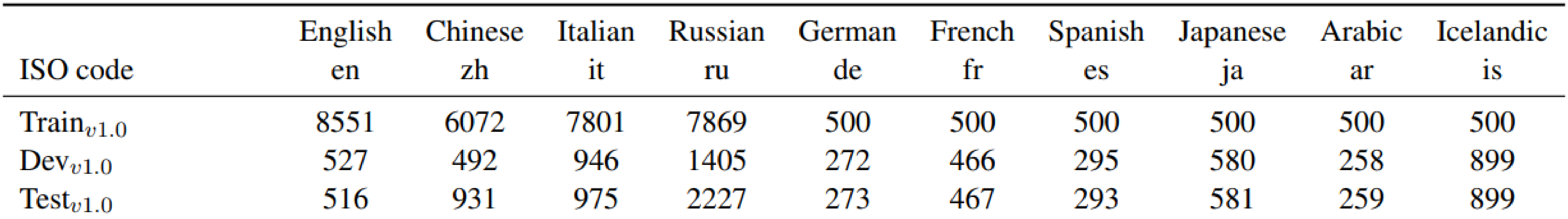
Zhang, ... Hu#; 2024 *ACL*

MELA: Multilingual Evaluation of Linguistic Acceptability

46k sentences in 10 lgs, collected from *Syntax of L* books, or previous work.



Language	L. F.	label	Examples	W. O.	Script	Gender	Casing
English (en)	Germ	1	One more pseudo generalization and I’m giving up.	SVO	Latin	N.A.	N.A.
Chinese (zh)	Sino-Tbt	0	张三被李四打了自己。	SVO	Han	N.A.	N.A.
Italian (it)	Rom	1	Quest’uomo mi ha colpito.	SVO	Latin	2	N.A.
Russian (ru)	Slavic	0	Этим летом не никуда ездили.	SVO	Cyrillic	3	6
German (de)	Germ	1	Die Frau sagt, dass ihm nicht zu helfen ist.	SVO	Latin	3	4
French (fr)	Rom	1	Je lui ait couru après.	SVO	Latin	2	N.A.
Spanish (es)	Rom	1	María bailó.	SVO	Latin	2	N.A.
Japanese (ja)	Altaic	0	犬が道端で死んである。	SOV	Han, Hiragana, Katakana	N.A.	N.A.
Arabic (ar)	Semitic	1	قال عمر إن كل السيارات استقدموها من ألمانيا.	VSO	Arabic	2	3
Icelandic (is)	Germ	1	Útlendingar gengu oft þennan stíg.	SVO	Latin	3	4



	English	Chinese	Italian	Russian	German	French	Spanish	Japanese	Arabic	Icelandic
ISO code	en	zh	it	ru	de	fr	es	ja	ar	is
$\text{Train}_{v1.0}$	8551	6072	7801	7869	500	500	500	500	500	500
$\text{Dev}_{v1.0}$	527	492	946	1405	272	466	295	580	258	899
$\text{Test}_{v1.0}$	516	931	975	2227	273	467	293	581	259	899

Zhang, ... Hu#; 2024 *ACL*

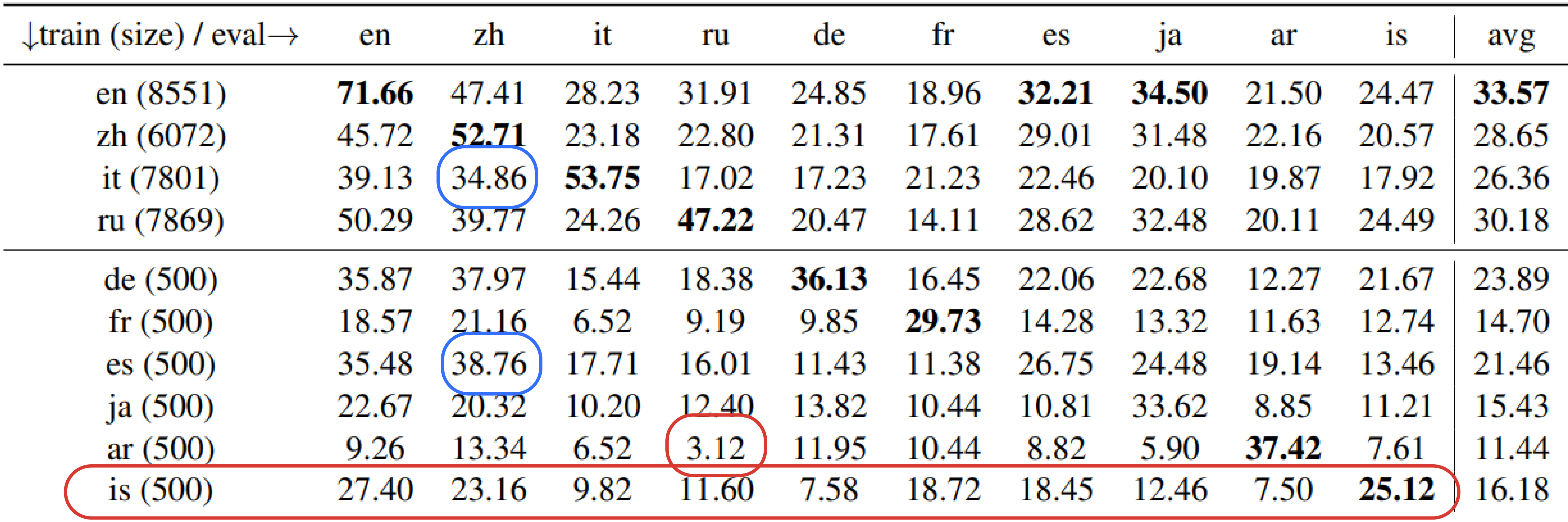
Experiments

Experiment 2: cross-lingual transfer in XLM-R

- pretrain on 100 lgs, finetune on acceptability in lg A, test on lg B

Findings:

- cross-lingual transfer is non-trivial (particularly bad for Arabic, typology?)
- size of training set matters, but not always



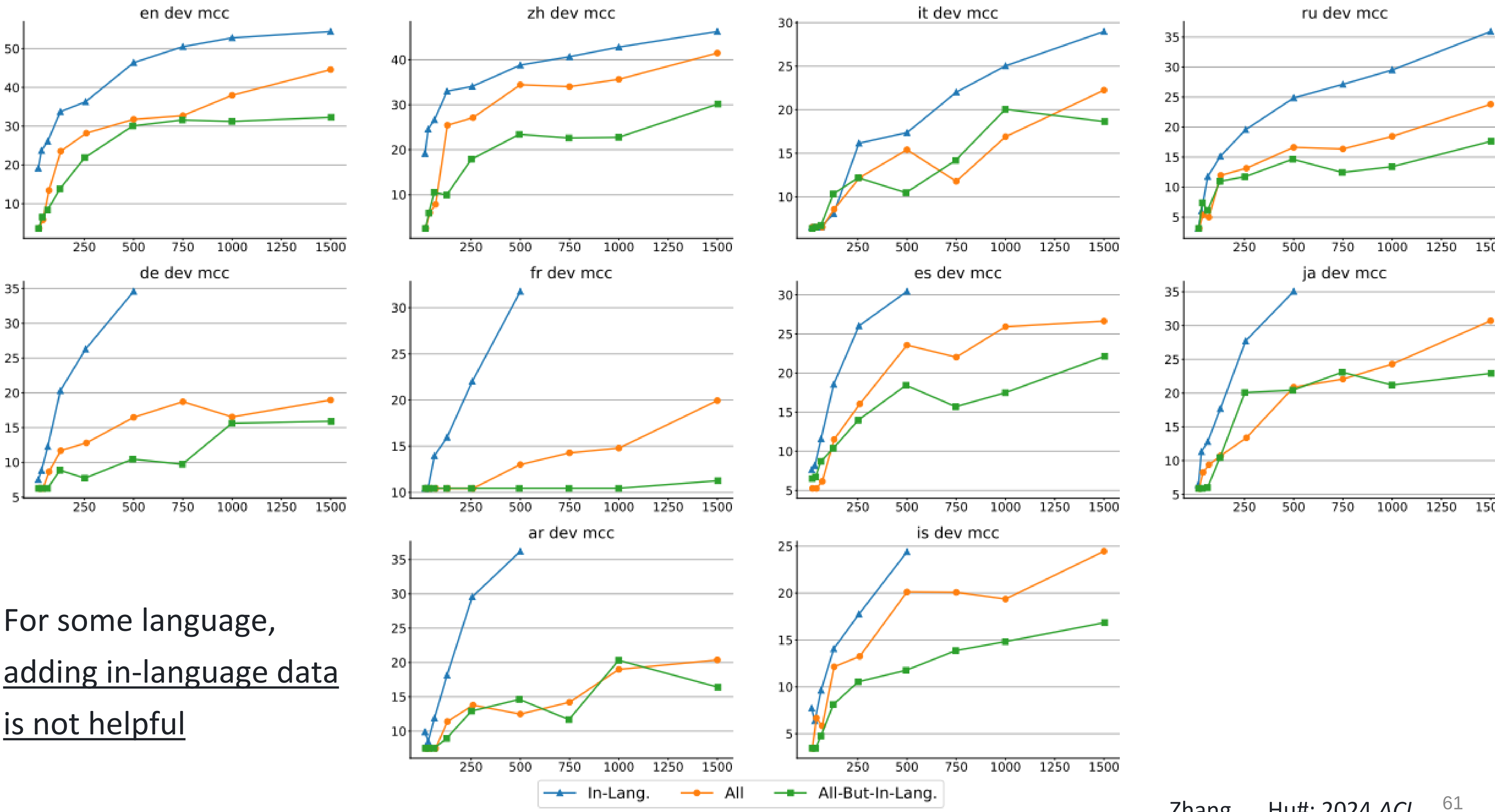
↓train (size) / eval→	en	zh	it	ru	de	fr	es	ja	ar	is	avg
en (8551)	**71.66**	47.41	28.23	31.91	24.85	18.96	**32.21**	**34.50**	21.50	24.47	**33.57**
zh (6072)	45.72	**52.71**	23.18	22.80	21.31	17.61	29.01	31.48	22.16	20.57	28.65
it (7801)	39.13	34.86	**53.75**	17.02	17.23	21.23	22.46	20.10	19.87	17.92	26.36
ru (7869)	50.29	39.77	24.26	**47.22**	20.47	14.11	28.62	32.48	20.11	24.49	30.18
de (500)	35.87	37.97	15.44	18.38	**36.13**	16.45	22.06	22.68	12.27	21.67	23.89
fr (500)	18.57	21.16	6.52	9.19	9.85	**29.73**	14.28	13.32	11.63	12.74	14.70
es (500)	35.48	38.76	17.71	16.01	11.43	11.38	26.75	24.48	19.14	13.46	21.46
ja (500)	22.67	20.32	10.20	12.40	13.82	10.44	10.81	33.62	8.85	11.21	15.43
ar (500)	9.26	13.34	6.52	3.12	11.95	10.44	8.82	5.90	**37.42**	7.61	11.44
is (500)	27.40	23.16	9.82	11.60	7.58	18.72	18.45	12.46	7.50	**25.12**	16.18

Zhang, ... Hu#; 2024 *ACL*

Multi-task learning

Three finetuning strategies:

In-language (train on A, test on A) | All-But-in-Language | All



For some language, <u>adding in-language data is not helpful</u>

Zhang, ... Hu#; 2024 *ACL*

Part 2: evaluating LLMs on their semantic and pragmatic understanding

How to know if a model understands language?

If the model makes a correct **inference** given a context / premise

then it understands language

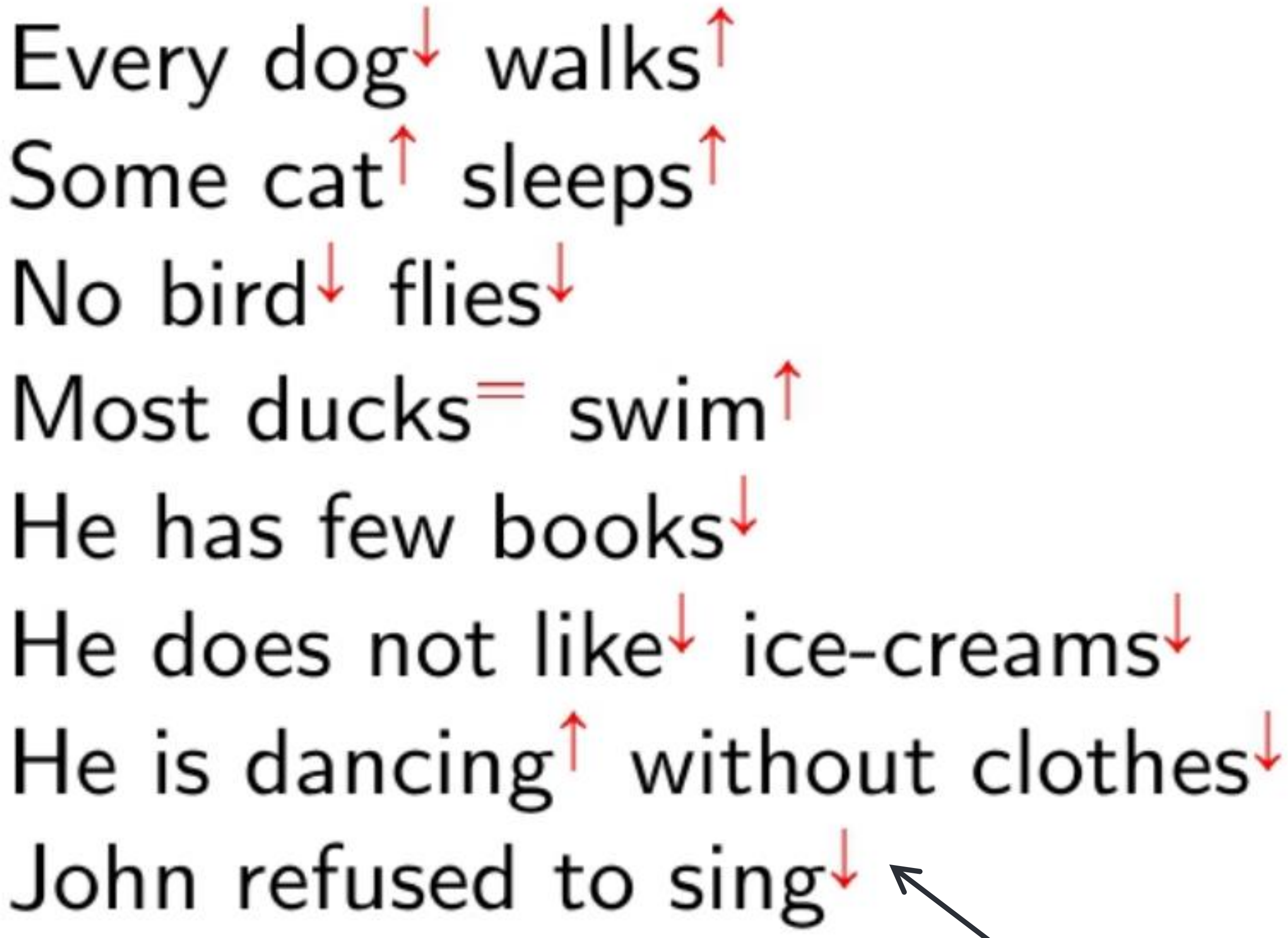
--> a task called **natural language inference (NLI)**

Context / premise	**Hypothesis**	Inference relation	Note
Every linguist is smart. Mary is a linguist.	Mary is smart.	entailment	Syllogistic Logic
John said to Mary: break a leg!	John wants to break Mary's leg.	contradiction	Non-literal meaning (idiom)
John doesn't believe that Mary is smart.	Mary is not smart.	neutral	Belief verbs

Natural language Inference

Two solutions:

(1) rule-based systems: natural logic (monotonicity)



Every dog$^{\downarrow}$ walks$^{\uparrow}$

Some cat$^{\uparrow}$ sleeps$^{\uparrow}$

No bird$^{\downarrow}$ flies$^{\downarrow}$

Most ducks$^{=}$ swim$^{\uparrow}$

He has few books$^{\downarrow}$

He does not like$^{\downarrow}$ ice-creams$^{\downarrow}$

He is dancing$^{\uparrow}$ without clothes$^{\downarrow}$

John refused to sing$^{\downarrow}$

monotonicity

Natural language Inference

Two solutions:

(1) rule-based systems: natural logic (monotonicity)

Our proposal: use CCG parse tree + monotonicity calculus to obtain ↑↓ =

then replacement

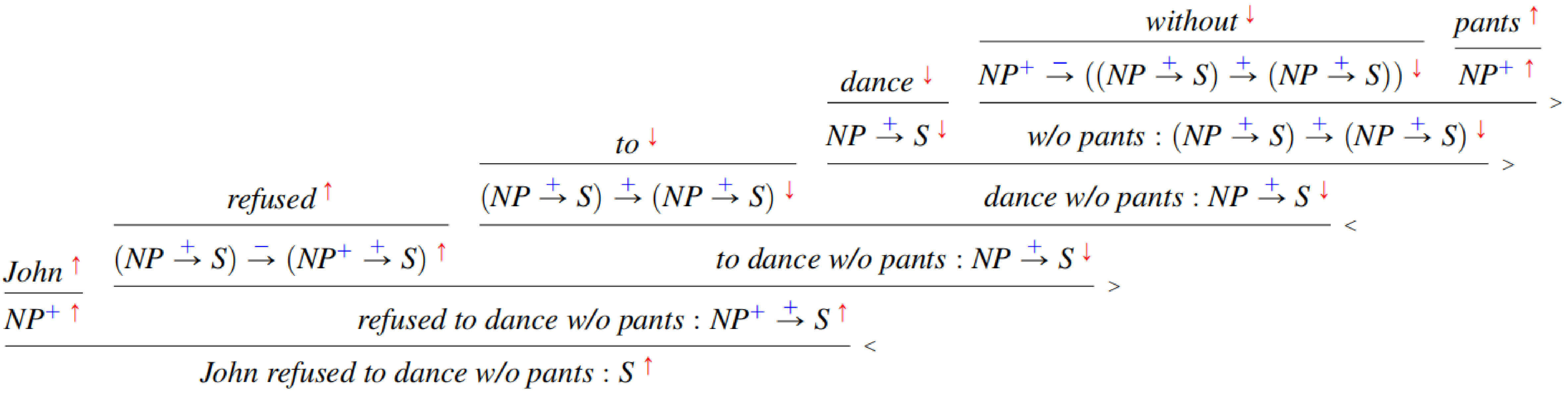

without ↓
pants ↑
dance ↓
NP+ –→ ((NP +→ S) +→ (NP +→ S)) ↓
NP+ ↑
to ↓
NP +→ S ↓
w/o pants : (NP +→ S) +→ (NP +→ S) ↓
refused ↑
(NP +→ S) +→ (NP +→ S) ↓
dance w/o pants : NP +→ S ↓
John ↑
(NP +→ S) –→ (NP+ +→ S) ↑
to dance w/o pants : NP +→ S ↓
NP+ ↑
refused to dance w/o pants : NP+ +→ S ↑
John refused to dance w/o pants : S ↑
>
<

Figure 3.8: CCG tree after **polarization**.

ccg2mono: Hu and Moss (2018) *StarSEM*

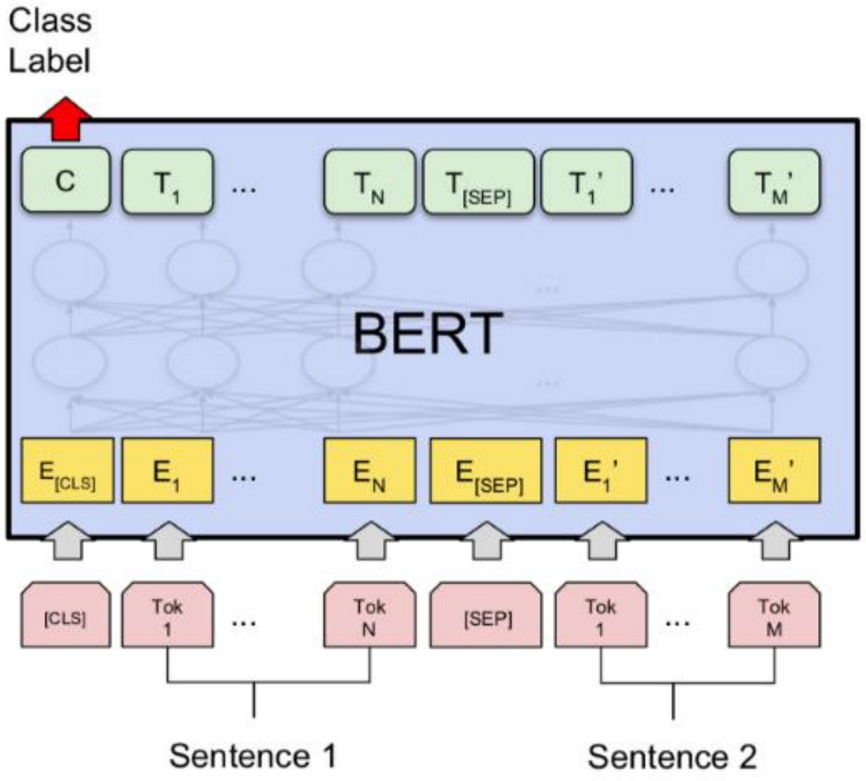
MonaLog: Hu et al (2020) *SCiL*

My dissertation (2021)

Natural language Inference

Two solutions:

(2) neural networks: BERT or XLM-RoBERTa



Sentence Pair Classification tasks in BERT paper

premise: at least five dogs that see every cat dance

hypothesis: at least four dogs that see one cat dance

relation?

premise 1	hypothesis 1	entailment
premise 2	hypothesis 2	non-entailment
...	...	...

Devlin et al 2019; *NAACL*

Strategy for training neural networks

Step 1: training data

- Neural network models need massive training/finetuning data

Step 2: test data

- We want to test on semantic/pragmatic phenomena we care about

Step 3: probing / interpretability

- We want to know why models behave in a particular way

Note:

- Linguistic questions vs Engineering questions

Overview of our work in this line

(1) create high-quality training data for LM fine-tuning

OCNLI (Hu et al 2020; EMNLP Findings)

Cured SICK dataset (Kalouli*, Hu*, et al 2023; Computational Linguistics)

(2) create test data/benchmarks that target specific linguistic phenomena

CLUE (Xu, Hu, et al 2020; COLING): https://cluebenchmarks.com/

Chinese NLI Probing (Hu et al 2021; ACL Findings)

Implicature (Yue, ..., Hu# 2024; CCL highlight paper award)

(3) probe into models' inner workings and learning trajectories

Semantic fragments (Richardson, Hu, Moss, Sabharwal 2020; AAAI)

Issues in natural language inference datasets

0. **Data creation**

- Give annotator a sentence, ask them to write an entailment, a neutral and a contradiction (diff annotator → diff inference), ask 4 other people to double-check
- Multi-genre NLI (MultiNLI; MNLI) dataset: 400k premise-hypothesis pairs in En

Issues in natural language inference datasets

0. **Data creation**

- Give annotator a sentence, ask them to write an entailment, a neutral and a contradiction (diff annotator → diff inference), ask 4 other people to double-check
- Multi-genre NLI (MultiNLI; MNLI) dataset: 400k premise-hypothesis pairs in En

1. Biases / artifacts: Superficial features that make NLI easy for the models

- Hypothesis-only bias (Poliak et al 2018)
  - Premise: A dog is running
  - Hypothesis: **no** dog is running around
  - LM will say: contradiction! b/c negation, not real understanding
  - Root cause:
    - annotators -> a lot of negation in contradiction -> higher probability for contradiction when negation -> higher acc.
- Training/test data too easy for the models

Issues in natural language inference datasets

2. What counts as entailment?

- Inherent disagreement on the labels

What do you say:

P: Paula swatted the fly.

H: The swatting happened in a forceful manner.

Issues in natural language inference datasets

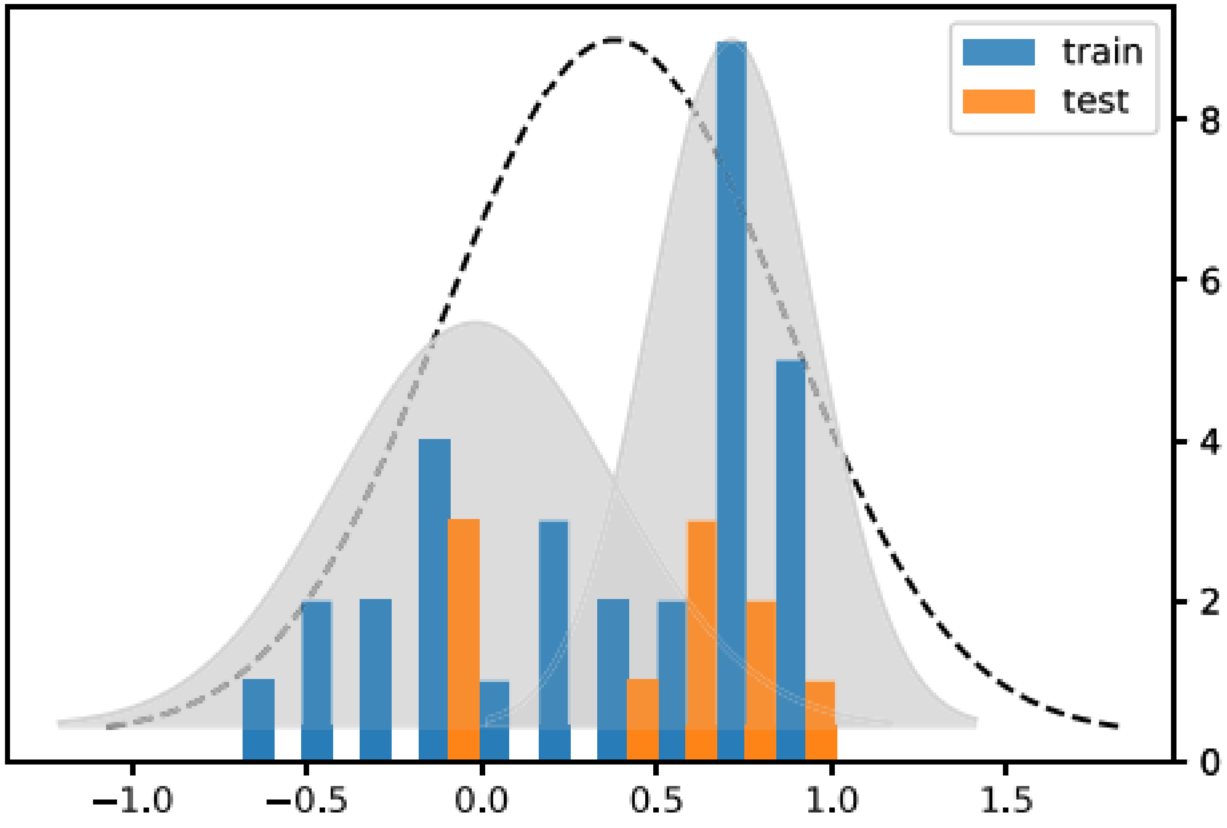
2. What counts as entailment?

- Inherent disagreement on the labels

What do you say:

P: Paula swatted the fly.

H: The swatting happened in a forceful manner.



Pavlick and Kwiatkowski 2020; TACL

To solve issue 1: our work

Create a high-quality training set: Original Chinese NLI (OCNLI)

Idea: **write more than one hypothesis per premise per label**

Four conditions

- **Single**: 1E + 1N + 1C (same as MNLI)
- **Multi**: 3E + 3N + 3C
- **MultiEncourage**: Encourage the annotators to be more creative
- **MultiConstraint**: Put constraints on what annotator can write

E.g., no "negators" in contradictions

145 undergraduate students (Ch/En majors) as annotators

50k+ pairs

First non-translated Chinese NLI dataset: avoids translationese!

Hu et al 2020; EMNLP Findings

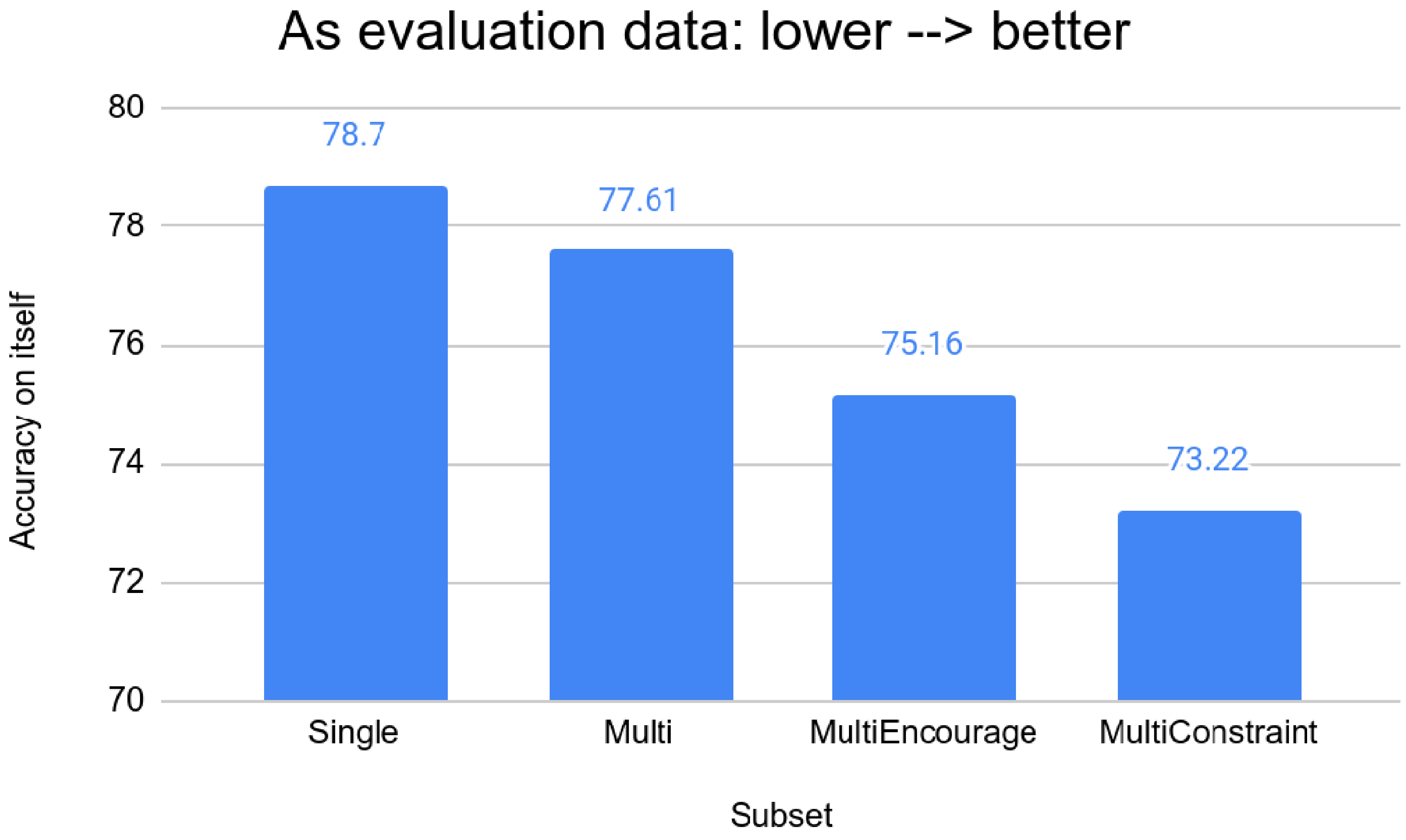
https://github.com/CLUEbenchmark/OCNLI

To solve issue 1: our work

Multi* data are more challenging

(but hypothesis-only bias is similar to previous datasets: bad)

2nd or 3rd hypotheses more challenging

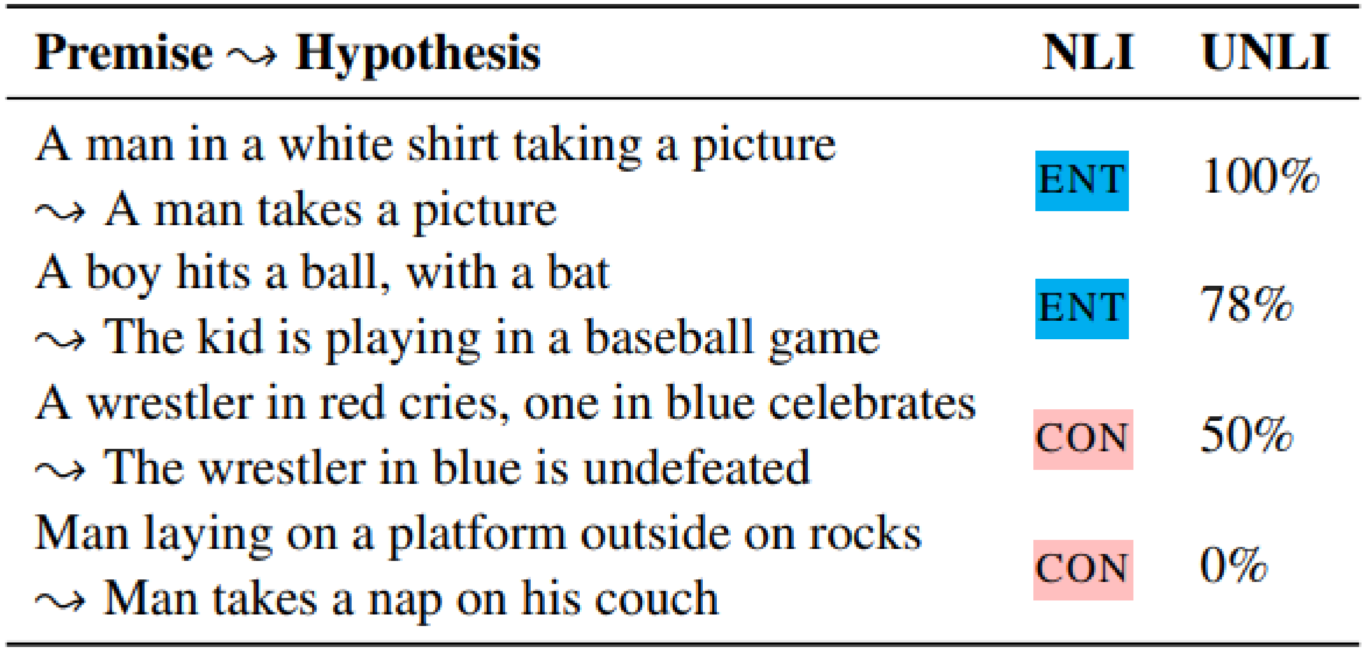


Hu et al 2020; EMNLP Findings

https://github.com/CLUEbenchmark/OCNLI

To solve issue 2: disagreement on inference label

Previous work:



Premise $\rightsquigarrow$ Hypothesis	NLI	UNLI
A man in a white shirt taking a picture $\rightsquigarrow$ A man takes a picture	ENT	100%
A boy hits a ball, with a bat $\rightsquigarrow$ The kid is playing in a baseball game	ENT	78%
A wrestler in red cries, one in blue celebrates $\rightsquigarrow$ The wrestler in blue is undefeated	CON	50%
Man laying on a platform outside on rocks $\rightsquigarrow$ Man takes a nap on his couch	CON	0%

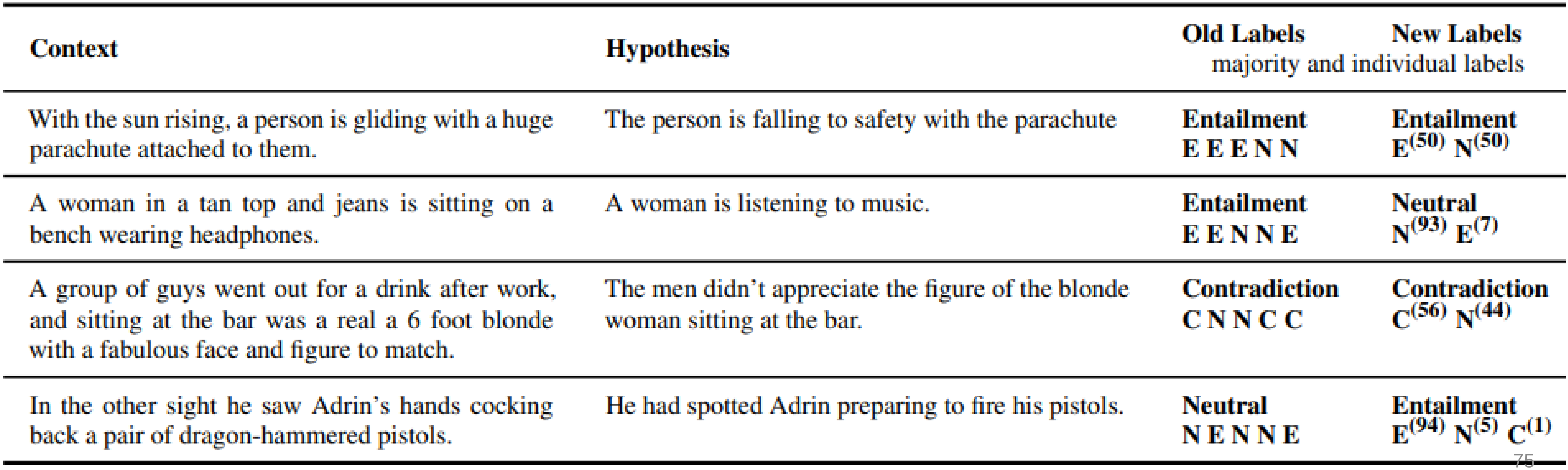
- Uncertain NLI (Chen et al 2019):

  probability of entailment

- ChaosNLI (Nie et al 2020):

  ask 100 annotators,

  instead of 5



Context	Hypothesis	Old Labels	New Labels
		majority and individual labels	
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** **E E E N N**	**Entailment** **E$^{(50)}$ N$^{(50)}$**
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** **E E N N E**	**Neutral** **N$^{(93)}$ E$^{(7)}$**
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** **C N N C C**	**Contradiction** **C$^{(56)}$ N$^{(44)}$**
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** **N E N N E**	**Entailment** **E$^{(94)}$ N$^{(5)}$ C$^{(1)}$**

To solve issue 2: disagreement on inference label

Our work: *Curing the SICK and Other NLI Maladies*

But we still don't know why people disagree

- Two labels for each pair:

  P: Paula swatted the fly.

  H: The swatting happened in a forceful manner.

  - Label 1: logic/strict, from a judge: Neutral
  - Label 2: commonsense/loose, from a person on the street: Entail
- Solved other linguistic problems about SICK/NLI dataset annotation

- LMs are able to distinguish logic labels from commonsense ones
- This more fine-grained annotation scheme is plausible for NLI

Kalouli*, Hu*, et al 2023;

Computational Linguistics

Create targeted test sets: our work

Test XLM-RoBERTas' cross-lingual transfer ability:

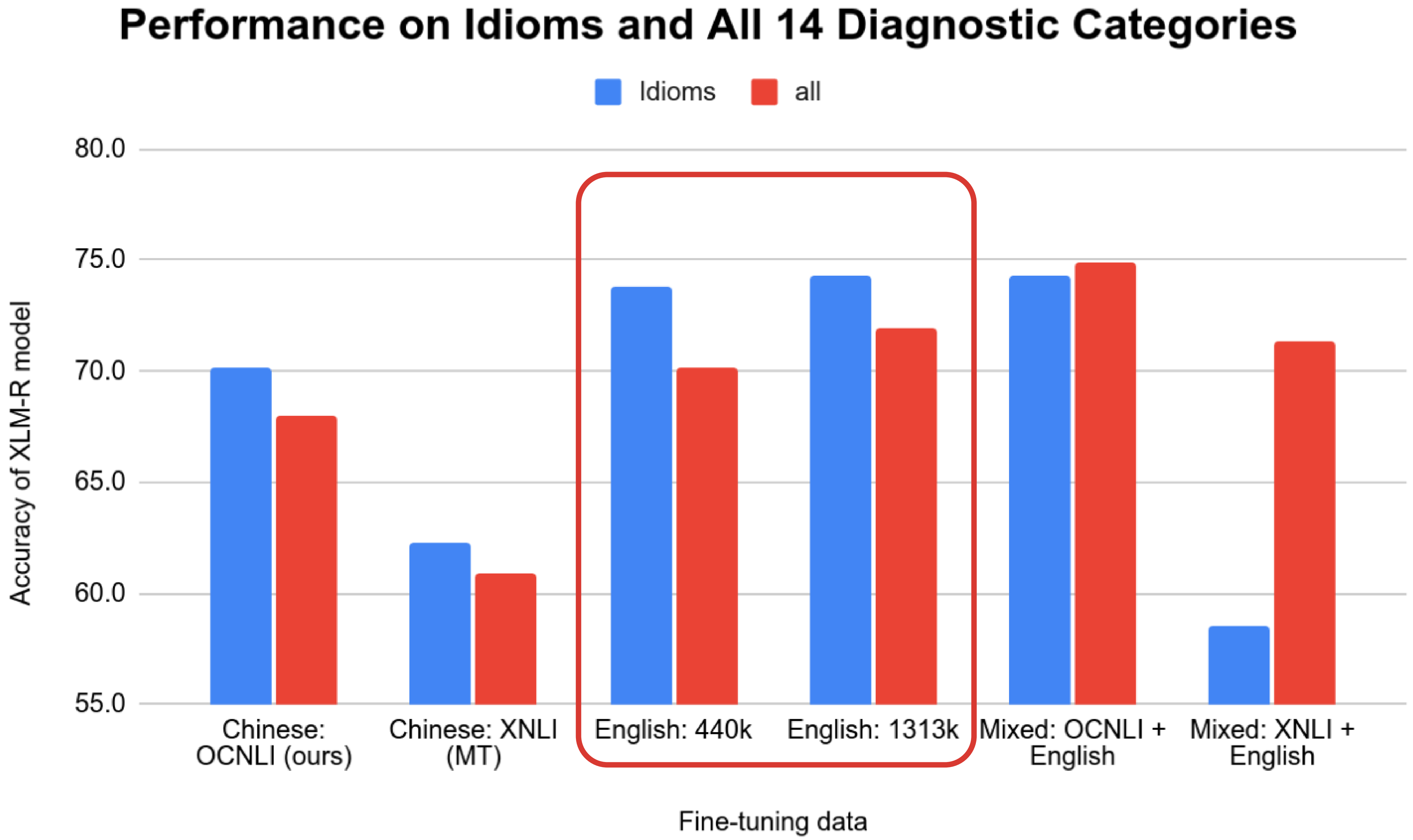
- Chinese Idioms: 打草惊蛇 = hit grass alert snake (lit) -> alarm the bad guys (fig)
- Finetune data: Chinese NLI; MT Chinese NLI; English NLI; Mixed

Hu et al 2021; ACL Findings

Create targeted test sets: our work

Test XLM-RoBERTas' cross-lingual transfer ability:

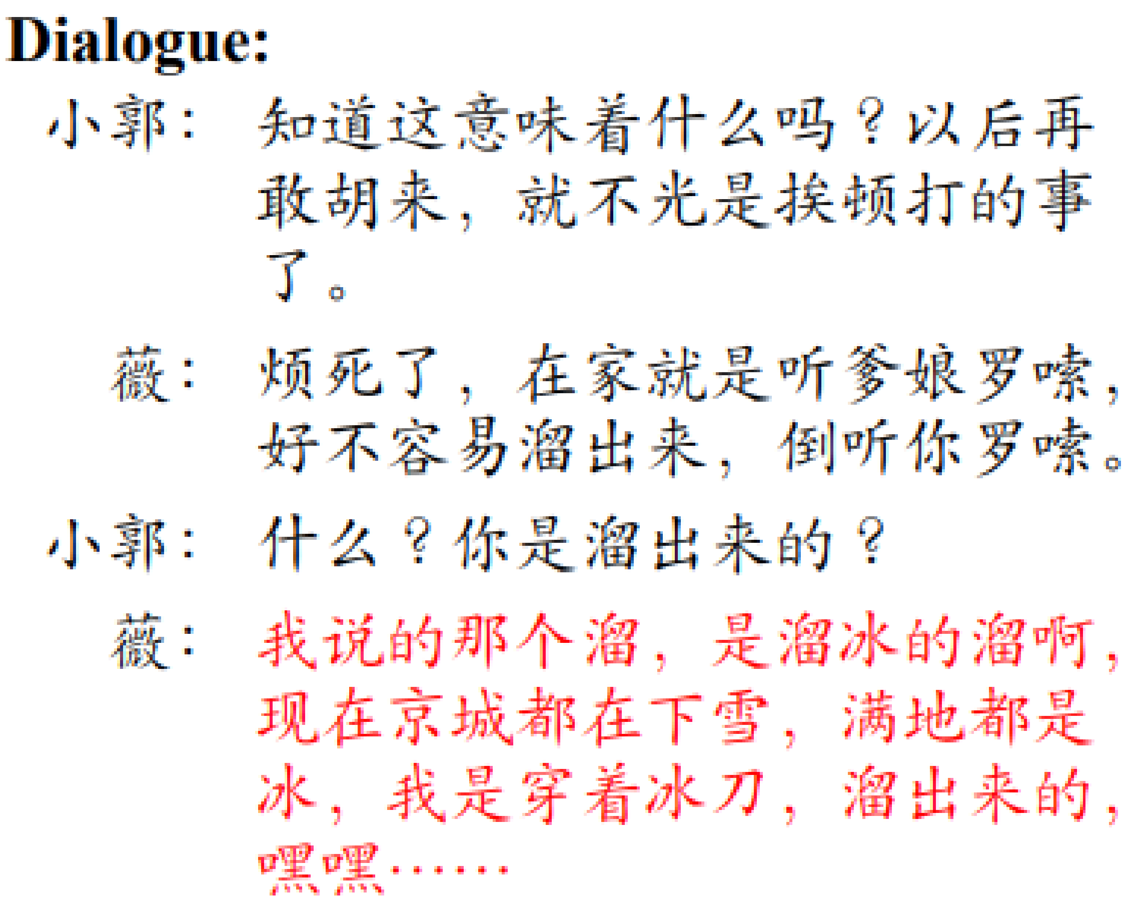
- Chinese Idioms: 打草惊蛇 = hit grass alert snake -> alarm the bad guys
- Finetune data: Chinese NLI; MT Chinese NLI; English NLI; Mixed



Hu et al 2021; ACL Findings

Pragmatics: Do LLMs understand conversational implicature?

- Understanding implied meaning is important in human communication
- 200 manually curated questions | multi-turn dialogues | Chinese sitcom
- Test closed-source and open-source models, and diff eval methods
- Exp1: Comprehension: multiple-choice questions
- Exp2: Production: linguists rate LLMs' explanations in fluency, logic, reasoning



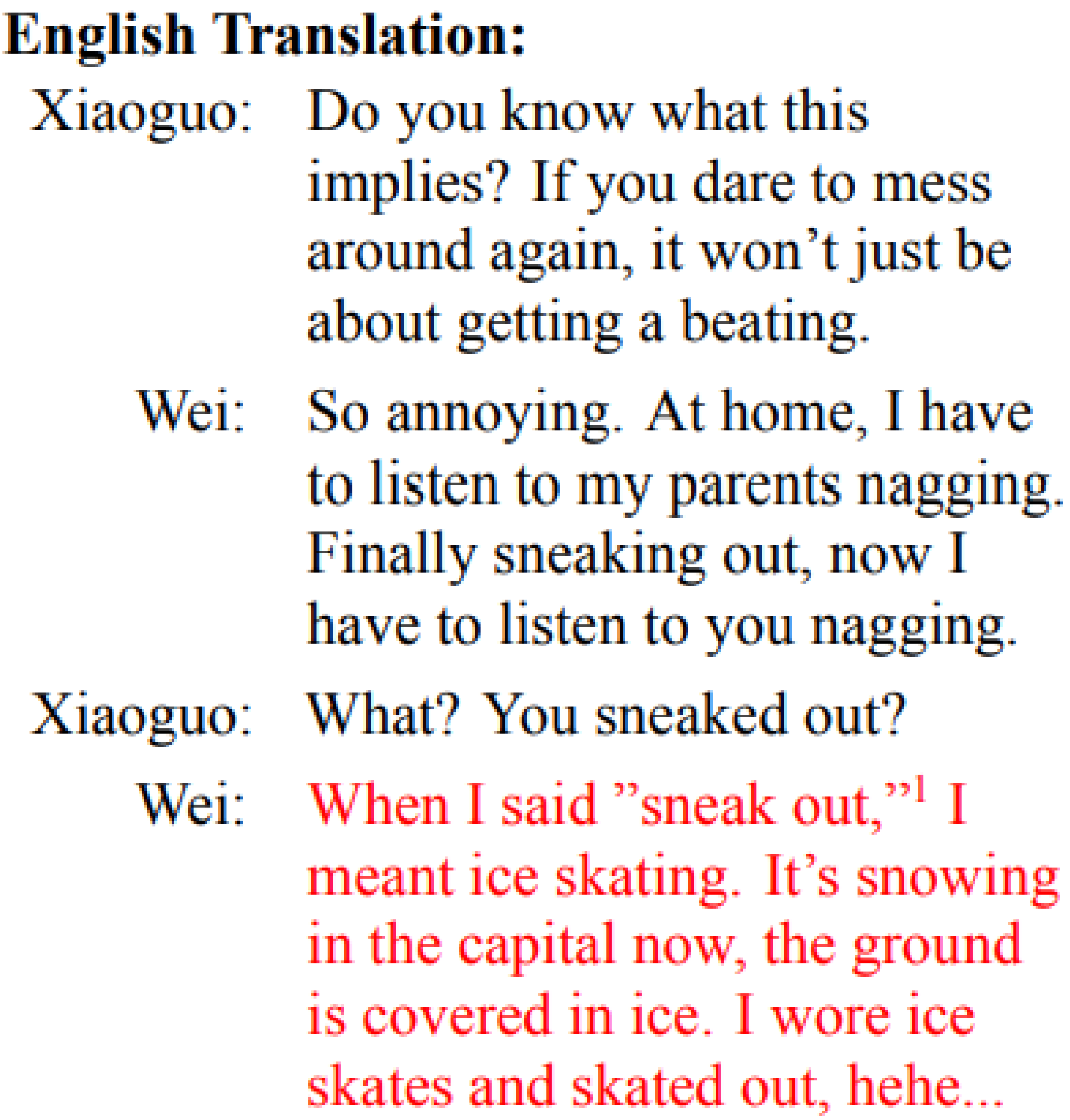
Dialogue:

小郭： 知道这意味着什么吗？以后再敢胡来，就不光是挨顿打的事了。

薇： 烦死了，在家就是听爹娘罗嗦，好不容易溜出来，倒听你罗嗦。

小郭： 什么？你是溜出来的？

薇： 我说的那个溜，是溜冰的溜啊，现在京城都在下雪，满地都是冰，我是穿着冰刀，溜出来的，嘿嘿……



English Translation:

Xiaoguo: Do you know what this implies? If you dare to mess around again, it won’t just be about getting a beating.

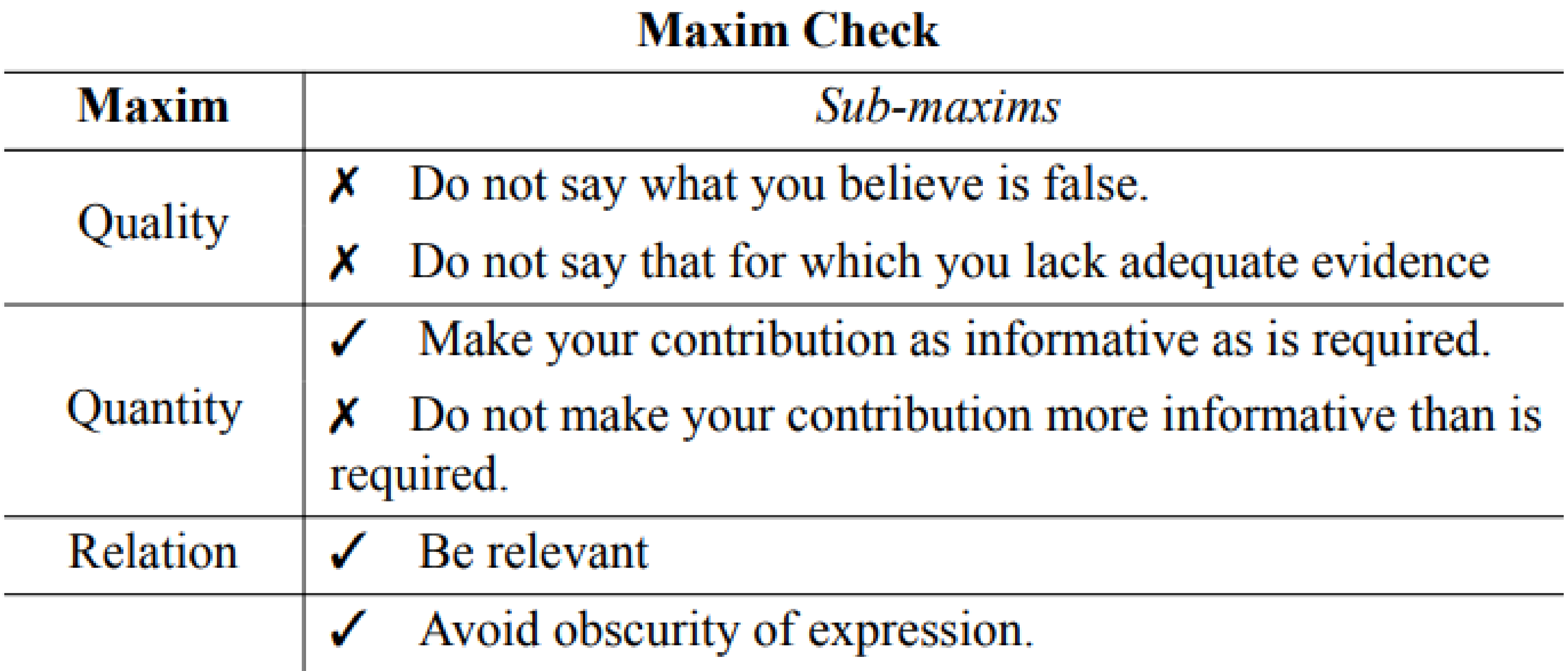
Wei: So annoying. At home, I have to listen to my parents nagging. Finally sneaking out, now I have to listen to you nagging.

Xiaoguo: What? You sneaked out?

Wei: When I said ”sneak out,”[1] I meant ice skating. It’s snowing in the capital now, the ground is covered in ice. I wore ice skates and skated out, hehe...

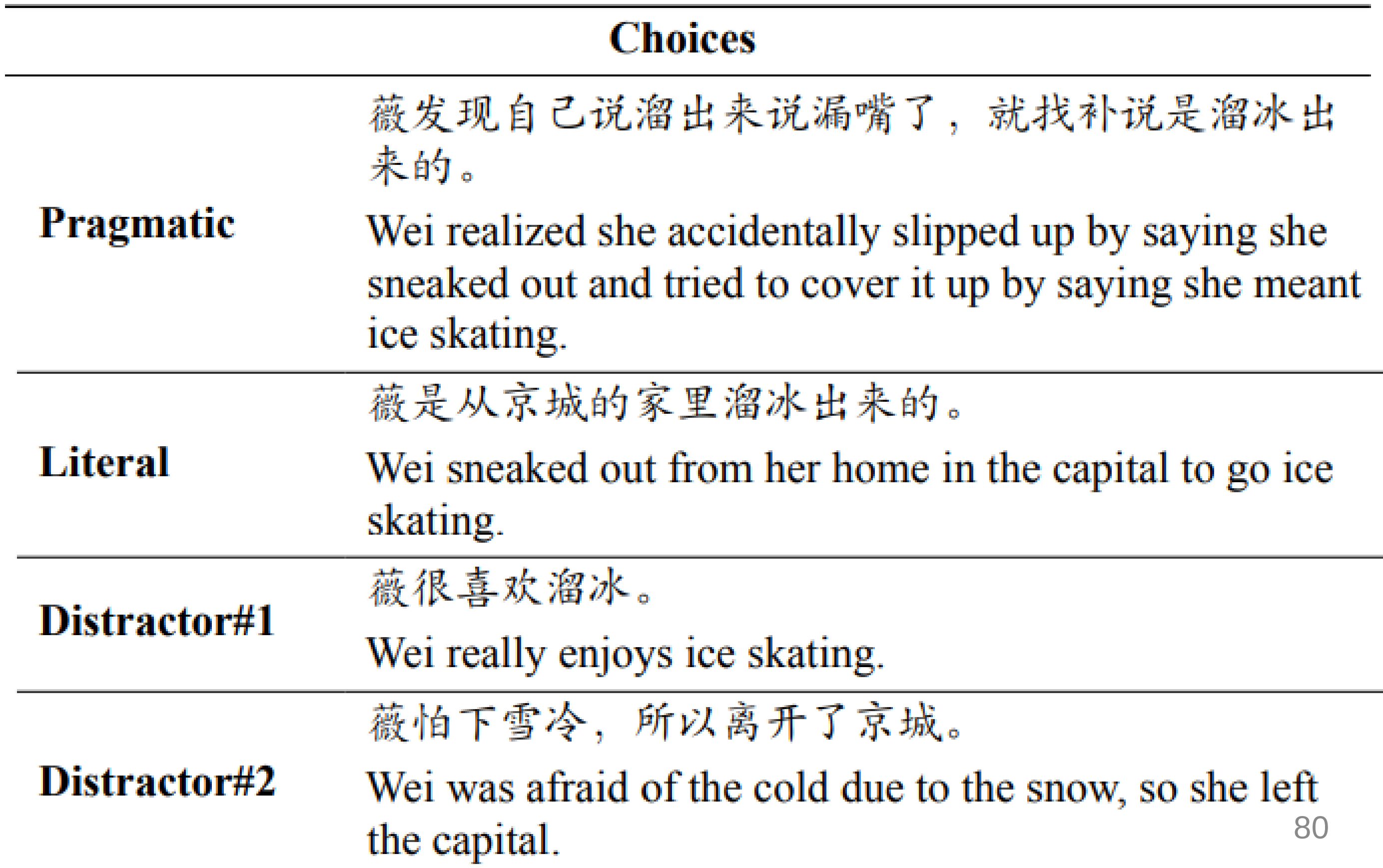
Yue, ... Hu# 2024; *CCL Highlight Paper Award*

Annotated for which Gricean maxims violated



Maxim Check

Maxim	*Sub-maxims*
Quality	✗ Do not say what you believe is false.
	✗ Do not say that for which you lack adequate evidence
Quantity	✓ Make your contribution as informative as is required.
	✗ Do not make your contribution more informative than is required.
Relation	✓ Be relevant
Manner	✓ Avoid obscurity of expression.
	✓ Avoid ambiguity
	✗ Be brief
	✓ Be orderly

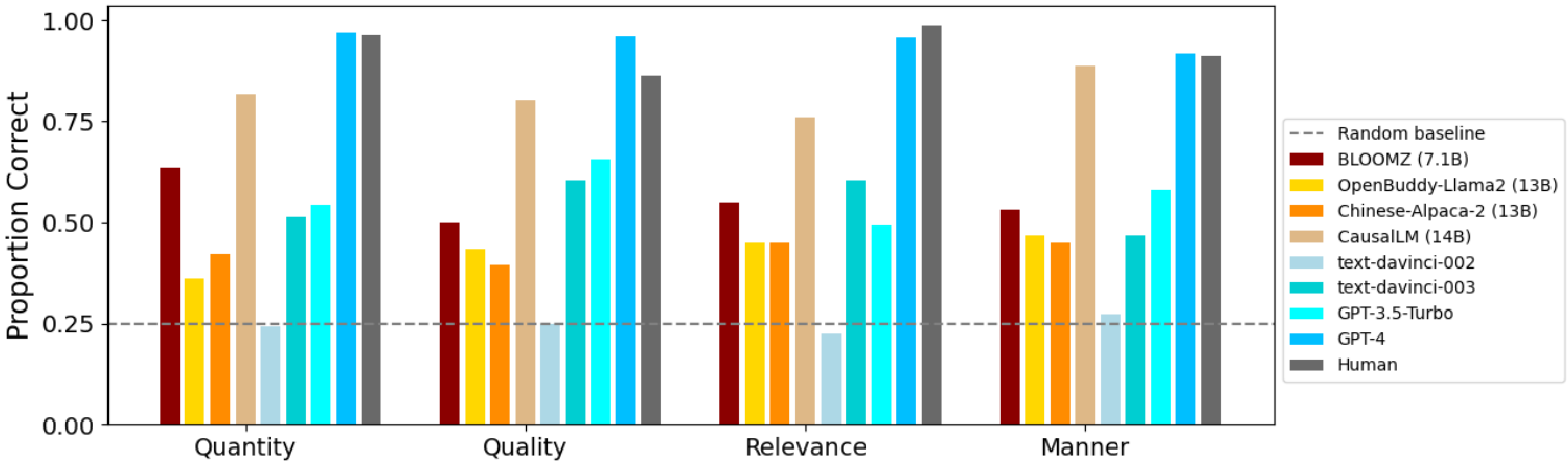


	Choices
Pragmatic	薇发现自己说溜出来说漏嘴了，就找补说是溜冰出来的。 Wei realized she accidentally slipped up by saying she sneaked out and tried to cover it up by saying she meant ice skating.
Literal	薇是从京城的家里溜冰出来的。 Wei sneaked out from her home in the capital to go ice skating.
Distractor#1	薇很喜欢溜冰。 Wei really enjoys ice skating.
Distractor#2	薇怕下雪冷，所以离开了京城。 Wei was afraid of the cold due to the snow, so she left the capital.

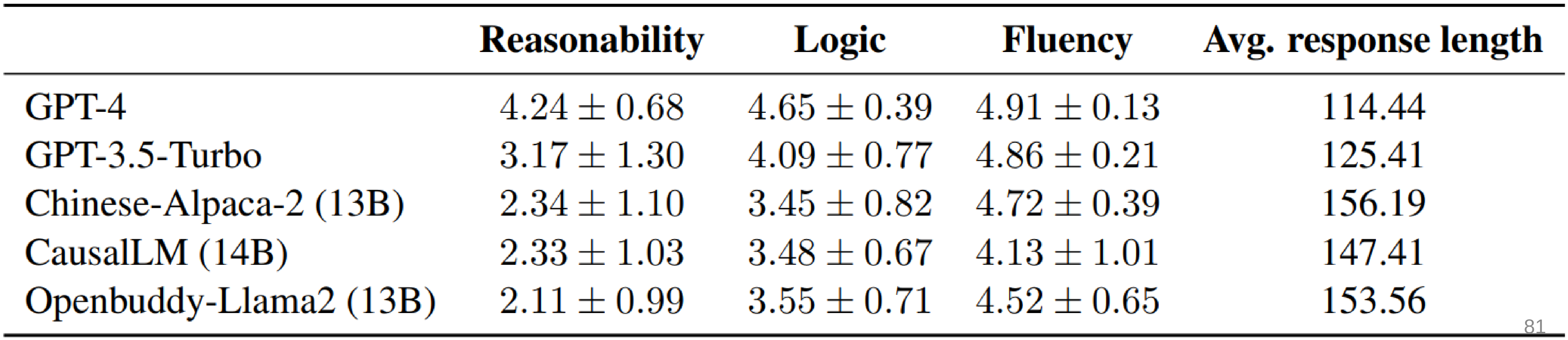
Yue, ... Hu# 2024; *CCL Highlight Paper Award*

Findings

Comprehension: GPT4 on par with humans | no clear diff on maxims



Production: gap between comprehension and production

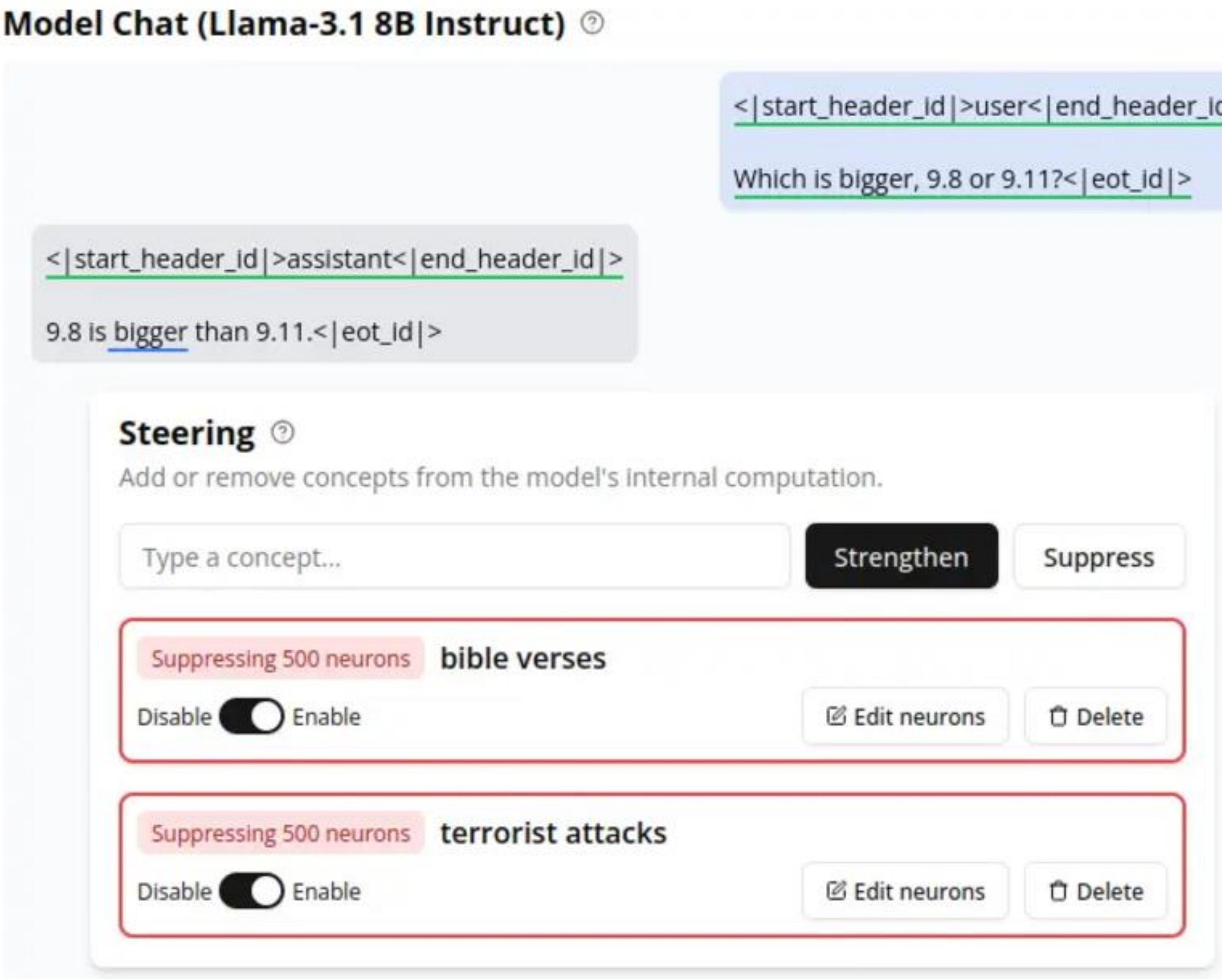


	Reasonability	Logic	Fluency	Avg. response length
GPT-4	4.24 ± 0.68	4.65 ± 0.39	4.91 ± 0.13	114.44
GPT-3.5-Turbo	3.17 ± 1.30	4.09 ± 0.77	4.86 ± 0.21	125.41
Chinese-Alpaca-2 (13B)	2.34 ± 1.10	3.45 ± 0.82	4.72 ± 0.39	156.19
CausalLM (14B)	2.33 ± 1.03	3.48 ± 0.67	4.13 ± 1.01	147.41
Openbuddy-Llama2 (13B)	2.11 ± 0.99	3.55 ± 0.71	4.52 ± 0.65	153.56

Moving forward

1. LLM interpretability at both the neuron level and the behavioral level.
... combined with our linguistically motivated evaluations.

2. Multilingual reasoning different representations?



Model Chat (Llama-3.1 8B Instruct)

<|start_header_id|>user<|end_header_i

Which is bigger, 9.8 or 9.11?<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>

9.8 is bigger than 9.11.<|eot_id|>

Steering

Add or remove concepts from the model's internal computation.

Suppressing 500 neurons bible verses

Suppressing 500 neurons terrorist attacks

https://transluce.org/observability-interface

Conclusion

Linguistics -> NLP

- Linguists can contribute to NLP by creating high-quality training and evaluation datasets.
  - Evaluation is even more challenging in the era of LLM
- By understanding the strengths and weaknesses of LLMs, linguistics can point out directions of LLM research

How can NLP be of help to linguists?

- LLMs show a kind-of successful way of learning human language
- Studying artificial neural networks informs us about human cognition
- Understanding how LLMs work help us use them better in research

Check out our resources and papers if you want to work on Chinese CL/NLP:

https://huhailinguist.github.io/

Computational linguistics is fun and a lot to be done!