Pragmatic Reasoning Unlocks Quantifier Semantics of Foundation Models

Yiyuan Li, Rakesh R Menon, Sayan Ghosh, Shashank Srivastava

University of North Carolina at Chapel Hill

Quantifier Semantics

- Human uses quantifiers to address fuzziness between subsets of concepts or entities.
  - e.g. “*Some birds can fly.*” indicates at least one bird can fly.

Quantifier Semantics

- Human uses quantifiers to address fuzziness between subsets of concepts or entities.
  - e.g. “*Some birds can fly.*” indicates at least one bird can fly.
  - Universal/Existential Quantifiers **fixed scopes**
    - For all/any
  - Generalized Quantifiers (*Mostowski 1957*) **fuzzy scopes**
    - Few/some/most etc.
    - Indicate the proportion that predicates satisfy.

Quantifier Semantics

- Human uses quantifiers to address fuzziness between subsets of concepts or entities.
  - e.g. “*Some birds can fly.*” indicates at least one bird can fly.
  - Universal/Existential Quantifiers **fixed scopes**
    - For all/any
  - Generalized Quantifiers (*Mostowski 1957*) **fuzzy scopes**
    - Few/some/most etc.
    - Indicate the proportion that predicates satisfy.
  - Abundant in communications (*Joshi et al. 2020, Cui et al. 2022*).

Question to Answer

Some *birds can fly.* → ***X% (0 < X < 100)*** *birds can fly.*

Understanding/Reasoning

Implicit functionalities

Question to Answer

Some *birds can fly.* → ***X% (0 < X < 100)*** *birds can fly.*

❓

(Bommasani et al. 2021)

Contributions

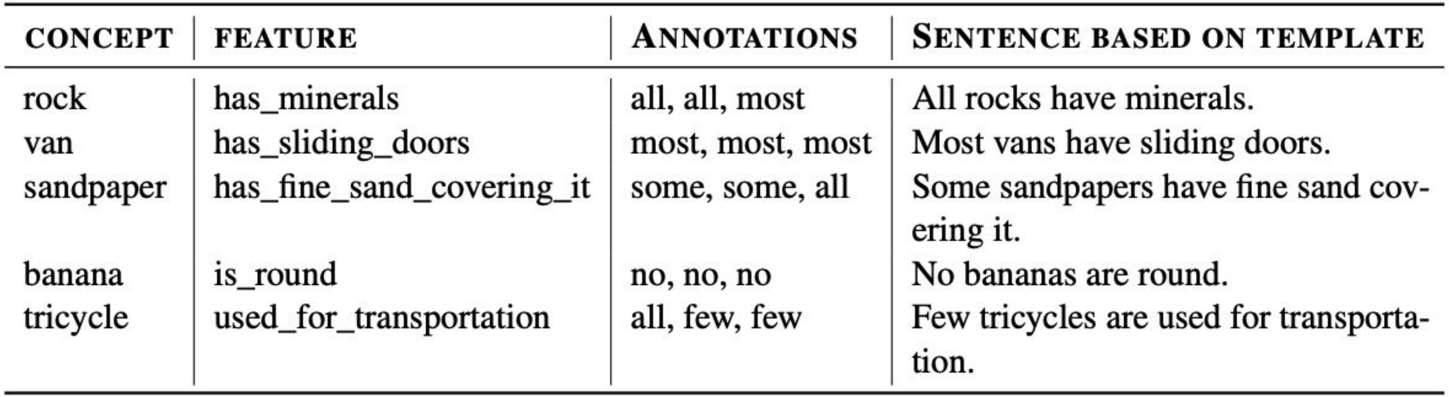
- An annotated dataset QuRe targeting real-world quantifier understanding.
- A pragmatic reasoning based framework PRESQUE for understanding quantifier semantics.

Task Definition

- Quantifier Understanding
    - Predicting the percentage scope (with an interval width) of a quantified sentence.
        - Spliting [0-1] into intervals W, e.g. {0%, 5%, 10%, …}
        - A quantifier understanding model predictes percentage scope from W that the predicate in the quantified sentence holds true (e.g. 5% - 30%).

Dataset

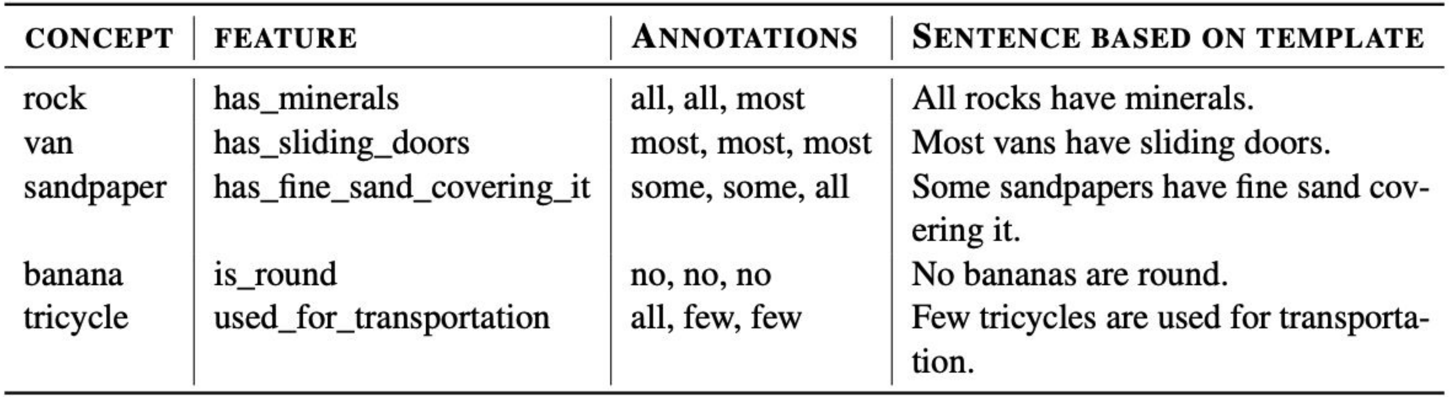
- Limited number of datasets with **human annotated quantifications**.
- HVD (*Herbelot and Vecchi 2015*)
    - quantifier annotation on the <concept, feature> pairs.



CONCEPT	FEATURE	ANNOTATIONS	SENTENCE BASED ON TEMPLATE
rock	has_minerals	all, all, most	All rocks have minerals.
van	has_sliding_doors	most, most, most	Most vans have sliding doors.
sandpaper	has_fine_sand_covering_it	some, some, all	Some sandpapers have fine sand covering it.
banana	is_round	no, no, no	No bananas are round.
tricycle	used_for_transportation	all, few, few	Few tricycles are used for transportation.

Dataset

- Limited number of datasets with **human annotated quantifications**.
- HVD (*Herbelot and Vecchi 2015*)
    - quantifier annotation on the <concept, feature> pairs.



CONCEPT	FEATURE	ANNOTATIONS	SENTENCE BASED ON TEMPLATE
rock	has_minerals	all, all, most	All rocks have minerals.
van	has_sliding_doors	most, most, most	Most vans have sliding doors.
sandpaper	has_fine_sand_covering_it	some, some, all	Some sandpapers have fine sand covering it.
banana	is_round	no, no, no	No bananas are round.
tricycle	used_for_transportation	all, few, few	Few tricycles are used for transportation.

No golden percentage scopes

QuRe Dataset

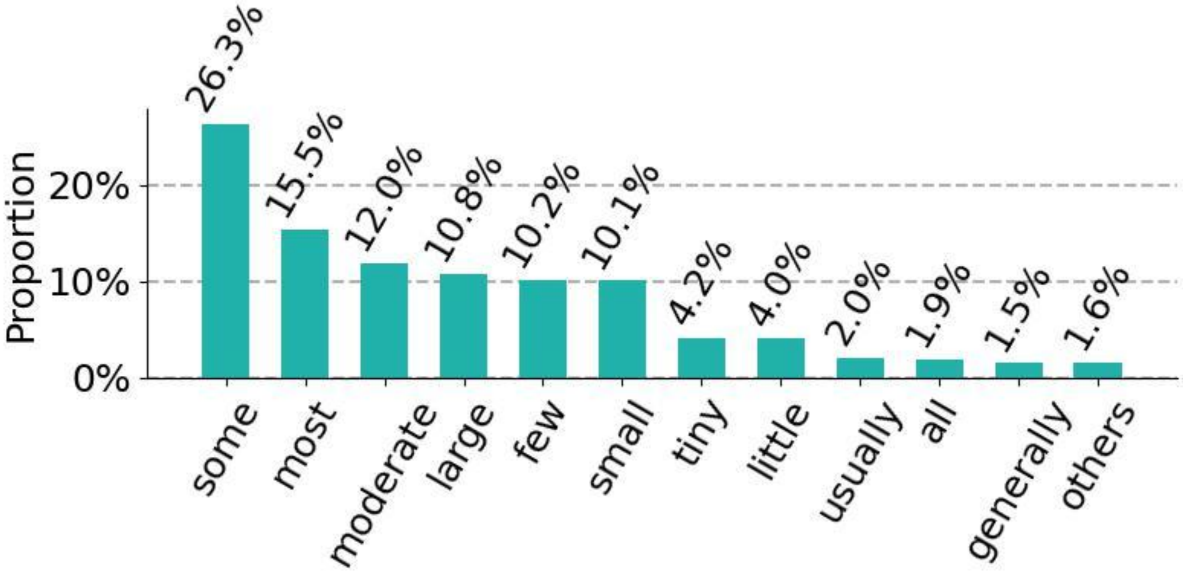
- QuRe
    - **More** generalized quantifiers.
        - {all, generally, most, usually, some, likely, few,

          little, occasionally, none, seldom,

          tiny, small, moderate, large}
    - **Specificity levels** in quantifier understanding
        - How hard it is to reason the percentage scope from the context.
    - **Golden percentage scopes** available.
        - The average age of the 304 drummers at Waterloo was 25, with some being boys under 16.
        - The average age of the 304 drummers at Waterloo was 25, with about 10% being boys under 16.
    - Sentence **topics**.

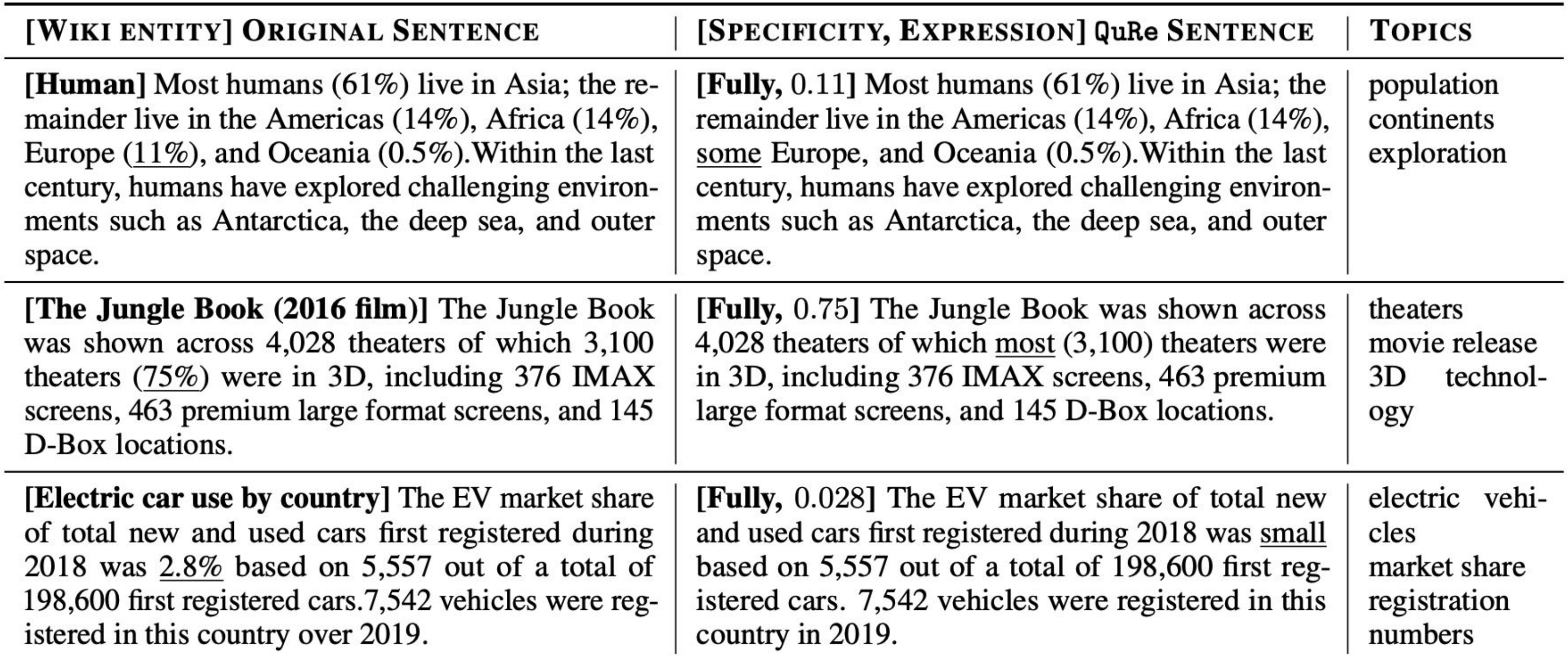
QuRe Dataset

- Quantifier distribution



QuRe Dataset

- Metadata examples



[Wiki entity] Original Sentence	[Specificity, Expression] QuRe Sentence	Topics
[Human] Most humans (61%) live in Asia; the remainder live in the Americas (14%), Africa (14%), Europe (11%), and Oceania (0.5%).Within the last century, humans have explored challenging environments such as Antarctica, the deep sea, and outer space.	**[Fully,** 0.11**]** Most humans (61%) live in Asia; the remainder live in the Americas (14%), Africa (14%), some Europe, and Oceania (0.5%).Within the last century, humans have explored challenging environments such as Antarctica, the deep sea, and outer space.	population continents exploration
[The Jungle Book (2016 film)] The Jungle Book was shown across 4,028 theaters of which 3,100 theaters (75%) were in 3D, including 376 IMAX screens, 463 premium large format screens, and 145 D-Box locations.	**[Fully,** 0.75**]** The Jungle Book was shown across 4,028 theaters of which most (3,100) theaters were in 3D, including 376 IMAX screens, 463 premium large format screens, and 145 D-Box locations.	theaters movie release 3D technology
[Electric car use by country] The EV market share of total new and used cars first registered during 2018 was 2.8% based on 5,557 out of a total of 198,600 first registered cars.7,542 vehicles were registered in this country over 2019.	**[Fully,** 0.028**]** The EV market share of total new and used cars first registered during 2018 was small based on 5,557 out of a total of 198,600 first registered cars. 7,542 vehicles were registered in this country in 2019.	electric vehicles market share registration numbers

Pragmatic Reasoning in Quantifier Understanding

- **P**ragmatic **Re**asoning for **S**emantics of **Qu**antifi**e**rs: **PRESQUE**
  - NLI backbone for text understanding.
  - Adoptation of Rational Speech Act (RSA).
  - **No training data needed**

Natural Language Inference (NLI)

- Find {*entailment, contradiction, neutrality*} relation between a *premise* sentence and *hypothesis* sentence (*Bowman et al. 2015*).

Natural Language Inference (NLI)

- Find {*entailment, contradiction, neutrality*} relation between a *premise* sentence and *hypothesis* sentence (*Bowman et al. 2015*).
- In PRESQUE, *quantified premises* to *percentaged hypotheses*.
    - e.g. few staircases have a spiral structure, 20% staircases have a spiral structure.

Natural Language Inference (NLI)

- Find {*entailment, contradiction, neutrality*} relation between a *premise* sentence and *hypothesis* sentence (*Bowman et al. 2015*).
- In PRESQUE, *quantified premises* to *percentaged hypotheses*.
    - e.g. few staircases have a spiral structure, 20% staircases have a spiral structure.

weak description (less exact)

strong description (more exact)

Limitations of NLI in Quantifier Understanding

- Implicit percentage value in quantifers (*Horowitz et al. 2018*)
- Sentence-level relation nature, impacts of linguistic and social clues (*Bergen et al. 2016*).
- Deficiencies in ambigous premises (*Thukral et al. 2021*) and quantative reasoning (*Naik et al. 2018; Ravichander et al. 2019*)

Beyond Direct Interpretation

- Pragmatic Theory (*Grice 1975*)
  - Locates the semantic meaning in interpretation considering the communication goal.
  - Reduced the complexity of semantic theories required for interpretation (*Bergen et al. 2016*)

Beyond Direct Interpretation

- Pragmatic Theory (*Grice 1975*)
    - Locates the semantic meaning in interpretation considering the communication goal.
    - Reduced the complexity of semantic theories required for interpretation (*Bergen et al. 2016*)
- Quantifier understanding through Rational Speech Act (RSA, *Frank and Goodman 2012*)

Rational Speech Act (RSA)

- World states *W* and utterances *U*.
- Lisenter *L* and speaker *S*.
- Bayesian approach of the pragmatic theory (iteratively modeling the mental state of *L* and *S*).

Quantifier Understanding through RSA

- W = {0%, 10%, 20%, …., 100%}   percentage value basis
- U = {no, few, some, most, all}   quantifier basis

Quantifier Understanding through RSA

- *W* = {0%, 10%, 20%, …., 100%}
- *U* = {no, few, some, most, all}
- premise $\bar{p}$: “All airplanes have engines.”
- hypothesis $\bar{h}$: “90% airplanes have engines.”

Quantifier Understanding through RSA

- premise $\bar{p}$: “All airplanes have engines.”
- hypothesis $\bar{h}$: “90% airplanes have engines.”
- Literal listener **baseline**

$$L_0(p|q) \propto \text{Entailment}(\tilde{p}, \tilde{h})$$

- Pragmatic speaker

$$S_0(q|p) \propto \text{Entailment}(\tilde{h}, \tilde{p})$$

- Pragmatic listener

$$L_1(p|q) \propto S_0(q|p)P(p)$$

$$P(p) = \sum_{q \in \mathcal{U}} P(p|q)P(q)$$

Model Choices of PRESQUE

- Foundation models
  - ALBERT-XXLarge (*Lan et al. 2020*)
  - XLNet-Large (*Yang et al. 2019*)
  - BART-Large (*Lewis et al. 2020*)
  - RoBERTa-Large (*Liu et al. 2019*)
- NLI finetuning datasets
  - SNLI (*Bowman et al. 2015*)
  - MNLI (*Williams et al. 2018*)
  - NLI-style FEVER (*Nie et al. 2019*)
  - Adversarial NLI (*Nie et al. 2020*)

Baselines

- Randomly ranking percentage values (Rnd)
- Literal listener (L_0): direct interpretation of NLI.

Evaluation Metrics

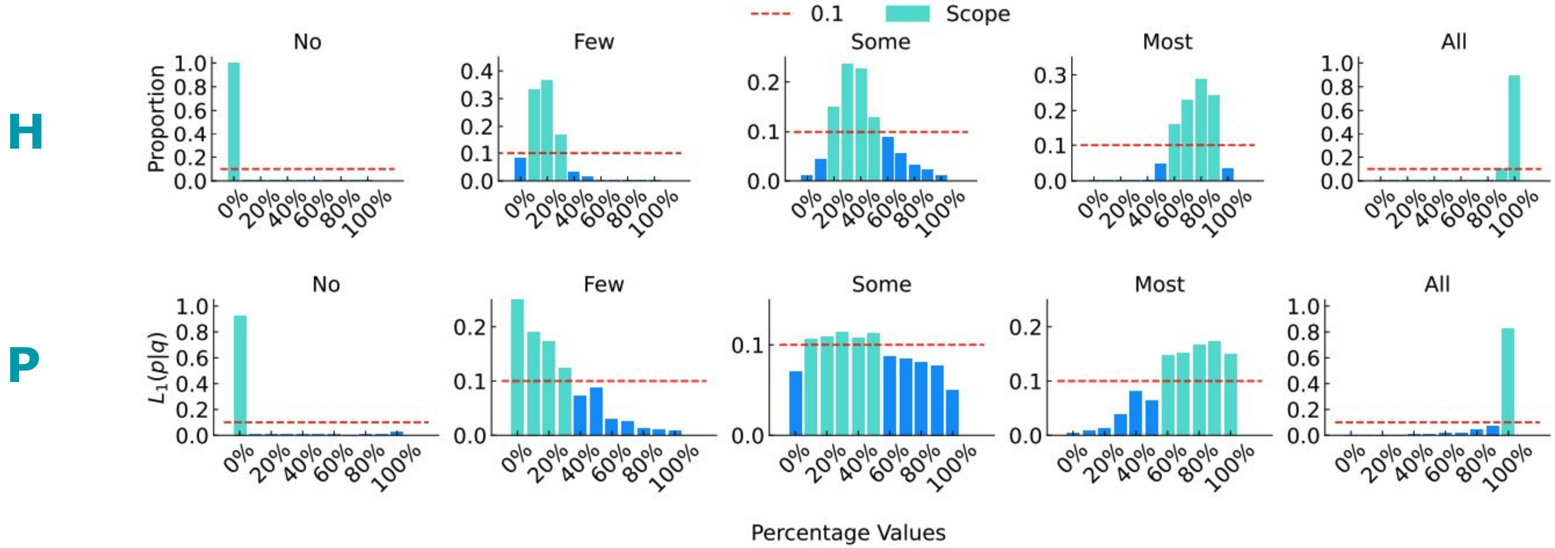
- HVD
  - Cross Entropy: The similarity between human and model perception of quantifier semantics.
- QuRe (starting from classification)
  - HIT@1: Topmost percentage value lies in the golden percentage scope.
  - Mean Reciporal Rank (MRR): The general ranking of the golden scope.
  - Cross Entropy: likelihood of the scope predictions.
  - Minimal Scope Distance (MSD@K): The distance of scope prediction of top K values and the golden scope.

Human Perception

- Instruct the annotator to define the percentage scope of the given quantifier (e.g. “Some stands for?”).

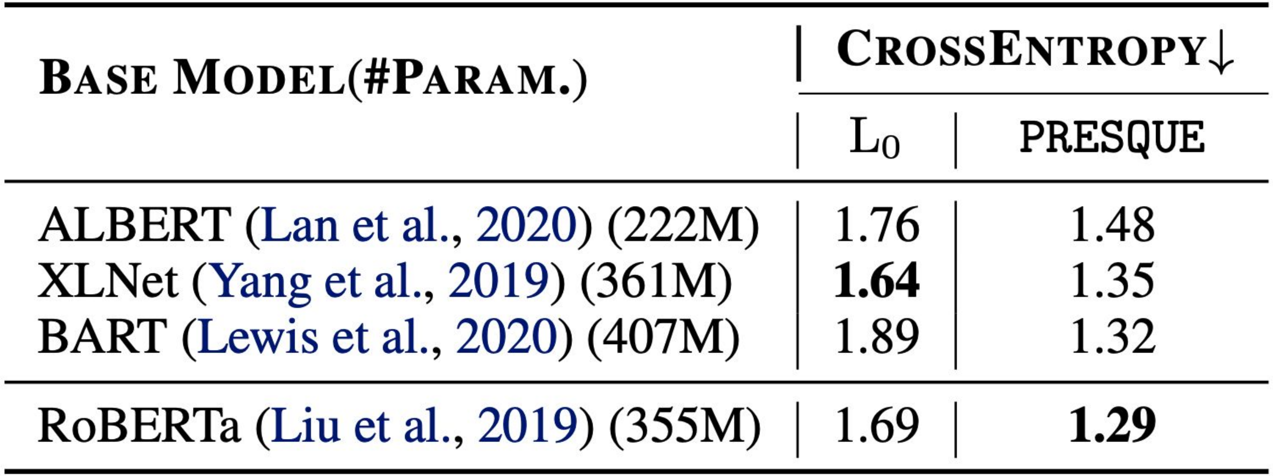
Model Perception - HVD

- Human perception (**H**) is similar to PRESQUE (**P**)



Result - HVD

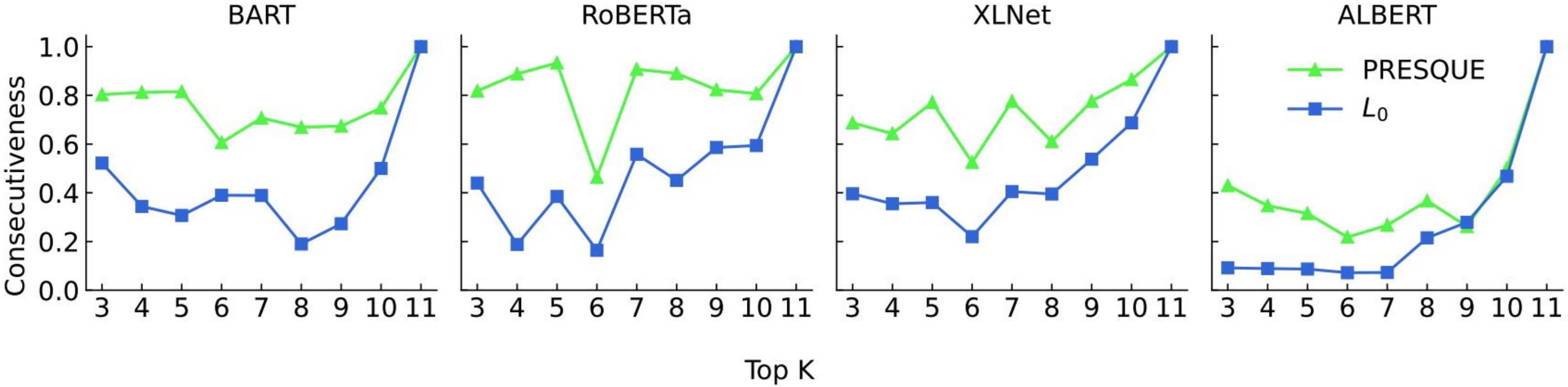
- PRESQUE is better than the literal listener (L_0).
- RoBERTa generally performs best among model choices.



Base Model(#Param.)	CrossEntropy↓	
	L_0	PRESQUE
ALBERT (Lan et al., 2020) (222M)	1.76	1.48
XLNet (Yang et al., 2019) (361M)	**1.64**	1.35
BART (Lewis et al., 2020) (407M)	1.89	1.32
RoBERTa (Liu et al., 2019) (355M)	1.69	**1.29**

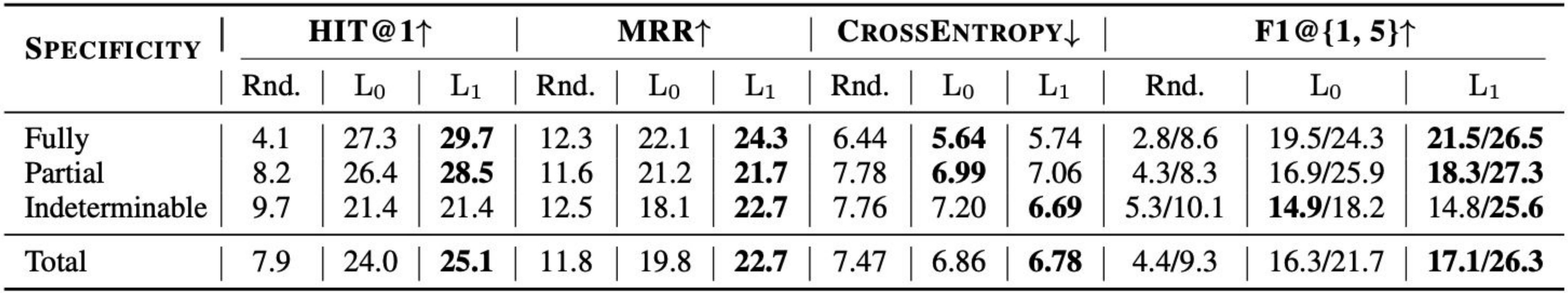
Consistency

- Consecutiveness of the Top K percentage inferences.
  - {10%, 20%, 30%}: consecutive (10%-30%)
  - {10%, 30%, 50%}: not consecutive
- PRESQUE has higher consecutiveness than L_0.



Result - QuRe

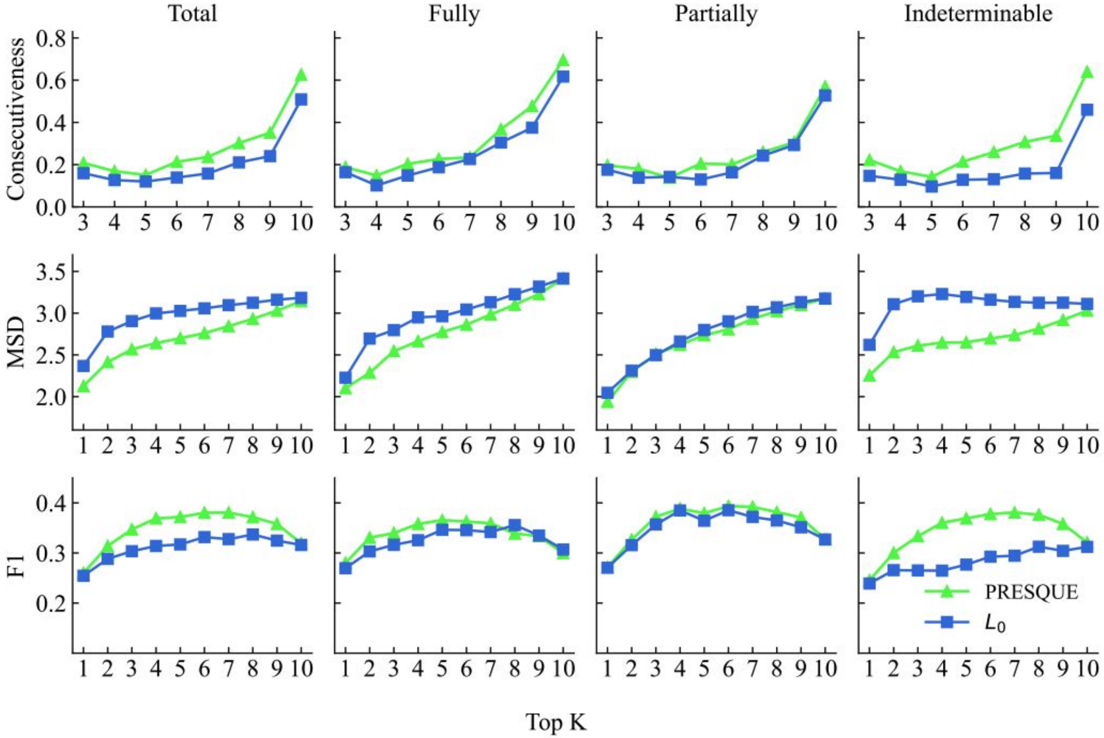
- PRESQUE generally performs better than L_0 among all specificity levels.



SPECIFICITY	HIT@1↑			MRR↑			CROSSENTROPY↓			F1@{1, 5}↑		
	Rnd.	L_0	L_1	Rnd.	L_0	L_1	Rnd.	L_0	L_1	Rnd.	L_0	L_1
Fully	4.1	27.3	**29.7**	12.3	22.1	**24.3**	6.44	**5.64**	5.74	2.8/8.6	19.5/24.3	**21.5/26.5**
Partial	8.2	26.4	**28.5**	11.6	21.2	**21.7**	7.78	**6.99**	7.06	4.3/8.3	16.9/25.9	**18.3/27.3**
Indeterminable	9.7	21.4	21.4	12.5	18.1	**22.7**	7.76	7.20	**6.69**	5.3/10.1	**14.9**/18.2	14.8/**25.6**
Total	7.9	24.0	**25.1**	11.8	19.8	**22.7**	7.47	6.86	**6.78**	4.4/9.3	16.3/21.7	**17.1/26.3**

Result - QuRe

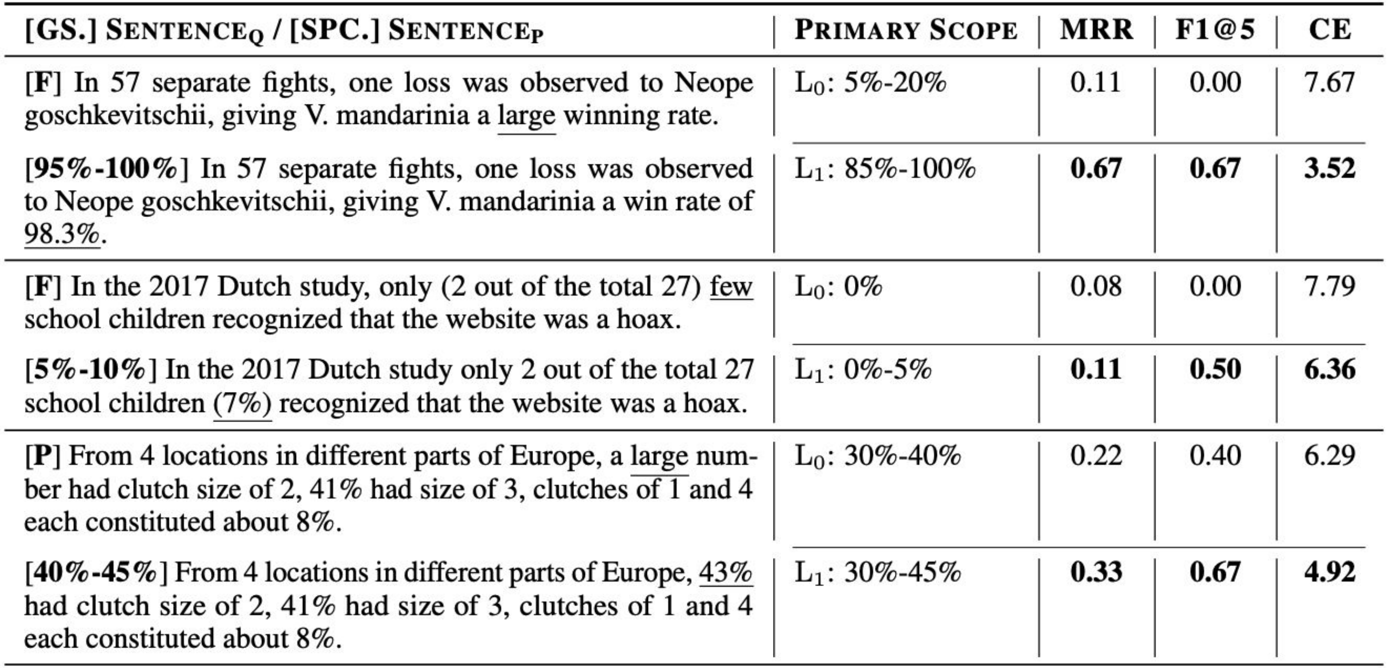
- Consistency + distance based scope evaluation



PRESQUE predictions has higher consecutiveness and are more similar to the golden percentage scopes than L_0.

Result - QuRe

- Examples



[GS.] SENTENCE$_Q$ / [SPC.] SENTENCE$_P$	PRIMARY SCOPE	MRR	F1@5	CE
[**F**] In 57 separate fights, one loss was observed to Neope goschkevitschii, giving V. mandarinia a large winning rate.	L$_0$: 5%-20%	0.11	0.00	7.67
[**95%-100%**] In 57 separate fights, one loss was observed to Neope goschkevitschii, giving V. mandarinia a win rate of 98.3%.	L$_1$: 85%-100%	**0.67**	**0.67**	**3.52**
[**F**] In the 2017 Dutch study, only (2 out of the total 27) few school children recognized that the website was a hoax.	L$_0$: 0%	0.08	0.00	7.79
[**5%-10%**] In the 2017 Dutch study only 2 out of the total 27 school children (7%) recognized that the website was a hoax.	L$_1$: 0%-5%	**0.11**	**0.50**	**6.36**
[**P**] From 4 locations in different parts of Europe, a large number had clutch size of 2, 41% had size of 3, clutches of 1 and 4 each constituted about 8%.	L$_0$: 30%-40%	0.22	0.40	6.29
[**40%-45%**] From 4 locations in different parts of Europe, 43% had clutch size of 2, 41% had size of 3, clutches of 1 and 4 each constituted about 8%.	L$_1$: 30%-45%	**0.33**	**0.67**	**4.92**

Paper: https://arxiv.org/pdf/2311.04659.pdf

Code: https://github.com/Nativeatom/PRESQUE

Thank you