Les Fondation Modèles comprennet **presque** les quantificateurs.

Research Question

Can foundation models understand (generalized) quantifiers like humans?

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

NLI Limitations

- Implicit percentage values of quantifiers.
- Sentence-level relation nature; impacts of linguistic and social clues.
- Deficiencies in ambiguous premises and quantitative reasoning.

Task

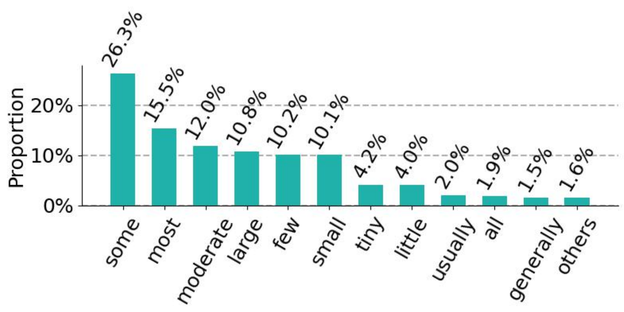
- Splitting [0, 1] in W with β: $W_{\beta=0.05} = \{0,5\,\%, 10\,\% \ldots\}$
- A model receives a quantified sentence and outputs $[p_1, p_2] \in W_\beta$ where the predicate in the quantified sentence holds true.

PRESQUE: **P**ragmatic **Re**asoning for **S**emantics of **Qu**antifi**e**rs

- NLI backbone
- RSA adoption
- No training, no bias

New Dataset: QuRe

- Existing datasets (e.g. HVD):
  - Limited quantifiers
  - **No golden percentage scope**
  - e.g. All rocks have minerals.
- Ours: QuRe
  - **More** generalized quantifiers



- **Specificity** levels
  - Difficulty to reason percentage scope from the non-quantification context.
  - Full/Partially/Indeterminable
- **Golden percentage scope**
- Sentence **topics**

[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
[List of blade materials] Prior to 2002, INFI contained 0.5% Carbon, 0.74% Nitrogen, about 1% Cobalt, and about 0.1% Nickel.	**[Partially,** 0.005] Prior to 2002, INFI contained tiny levels of Carbon, 0.74% Nitrogen, about 1% Cobalt, and about 0.1% Nickel.	chemical composition INFI elements
[List of blade materials] In order for a steel to be considered stainless it must have a Chromium content of at least 10.5%.	**[Indeterminable,** >= 0.105] In order for a steel to be considered stainless it must have some Chromium content.	steel metallurgy composition

Beyond Direct NLI Interpretation

- Pragmatic theory (*Grice 1975*)
- Rational Speech Act (RSA, *Goodman and Frank 2016*)
- Word state W and utterances U.
  - W: percentage value set {0, 10%, 20% …}
  - U: quantifier set {no, few, some …}
- The mental states of Listener L and speaker S are iteratively modeled with a Bayesian approach.

$$\begin{array}{lll} L_1 & \textit{pragmatic listener} & P_{L_1}(s|u) \propto P_{S_1}(u|s) \cdot P(s) \\ \downarrow & & \\ S_1 & \textit{pragmatic speaker} & P_{S_1}(u|s) \propto \exp(\alpha U_{S_1}(u;s)) \\ \downarrow & & \\ L_0 & \textit{literal listener} & P_{L_0}(s|u) \propto [\![u]\!](s) \cdot P(s) \end{array}$$

PRESQUE (L_1)

- Premise $\tilde{p}$: All airplanes have engines.
- Hypothesis $\tilde{h}$: 90% airplanes have engines.
- p: percentage values.
- q: quantifiers
- Entailment(·) comes from the NLI model.

- Literal listener

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

- Pragmatic speaker

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

- Pragmatic listener

$$L_1(p|q) \propto S_0(q|p)P(p) \qquad P(p) = \sum_{q \in \mathcal{U}} P(p|q)P(q)$$

Result (QuRe)

- PRESQUE generally performs better than L_0 among all

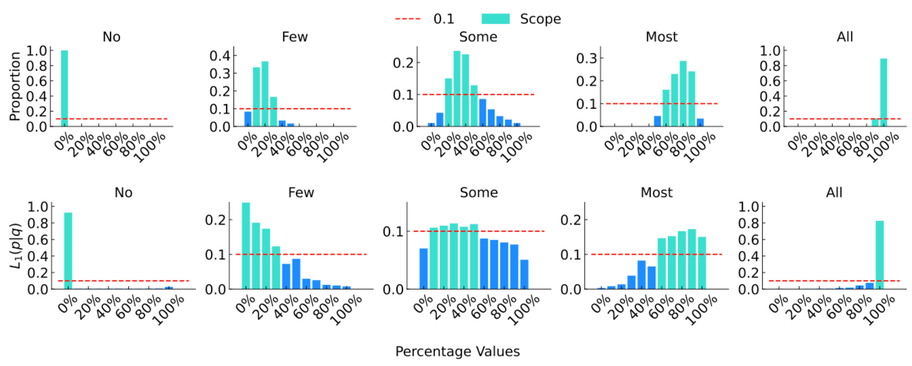
Specificity	HIT@1↑			MRR↑			CrossEntropy↓			F1@{1, 5}↑		
	Rnd.	L_0	L_1	Rnd.	L_0	L_1	Rnd.	L_0	L_1	Rnd.	L_0	L_1
Fully	4.1	27.3	**29.7**	12.3	22.1	**24.3**	6.44	**5.64**	5.74	2.8/8.6	19.5/24.3	**21.5/26.5**
Partial	8.2	26.4	**28.5**	11.6	21.2	**21.7**	7.78	**6.99**	7.06	4.3/8.3	16.9/25.9	**18.3/27.3**
Indeterminable	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**

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**
[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**
[I] It is typically made from rye bread, usually known as black bread, and is not classified as an alcoholic beverage in Poland, as its alcohol content usually is very little.	L_0: 60%-70%	0.06	0.00	6.97
[0-5%] It is typically made from rye bread, usually known as black bread, and is not classified as an alcoholic beverage in Poland, as its alcohol content usually ranges from 0% to 2%.	L_1: 0%-5%	**0.33**	**1.00**	**4.16**

Human & PRESQUE Perception

Human defined percentage scopes of quantifiers are similar to PRESQUE.



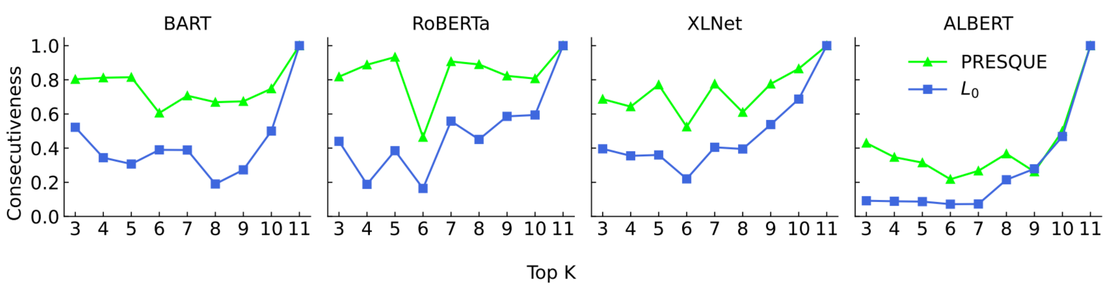
Result (HVD)

- PRESQUE has lower cross entropy than L_0 .
- RoBERTa generally performs

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**

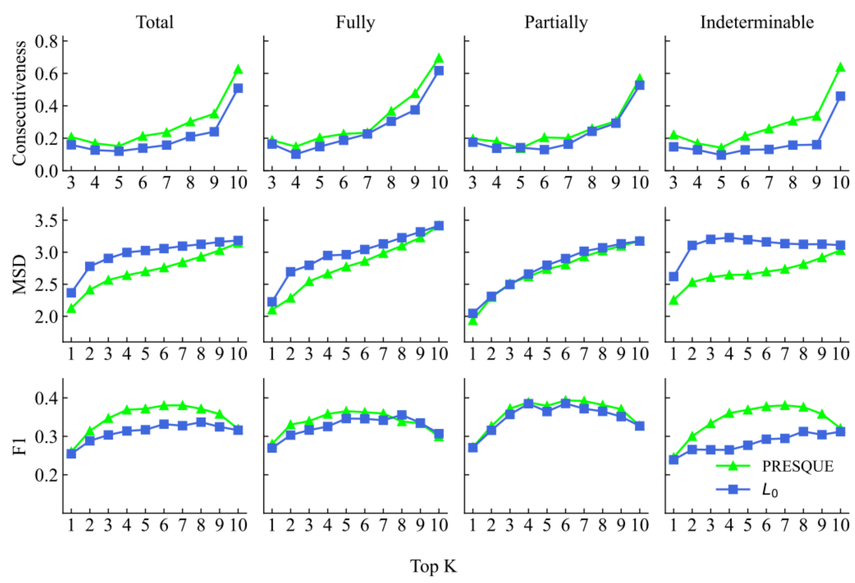
Consecutiveness

- If top 3 percentage predictions are
  - {10%, 20%, 30%}, gives a consecutive scope 10%-30%. **Consistent**
  - {10%, 30%, 50%}, no consecutive scope. **Not consistent**
- PRESQUE predictions has higher consecutiveness ratio than L_0 .



Consecutiveness & Distance Metrics

- MSD: the minimal distance between the consecutive scope of top K percentage predictions and the golden percentage scope.
- F1: the span overlap between the consecutive scope of top K percentage predictions and the golden percentage scope.
- PRESQUE has higher consecutiveness and lower MSD



Pragmatic Reasoning Unlocks Quantifier Semantics for Foundation Models

Yiyuan Li, Rakesh R Menon, Sayan Ghosh, Shashank Srivastava
yiyuanli@cs.unc.edu

Paper

EMNLP 2023