Are Shortest Rationales the Best Explanations for Human Understanding?

Hua Shen♦, Tongshuang Wu✦, Wenbo Guo♦, Ting-Hao 'Kenneth' Huang♦

PennState · University of Washington

huashen218@psu.edu

ACL 2022 · 22ND – 27TH MAY | 60TH MEETING | DUBLIN

1. Motivation

- Self-explaining models typically extract **shortest possible rationales** — snippets of an input text "responsible for" corresponding output — **to explain the model prediction**.
- Based on the common assumption — **"shorter rationale is better for human understanding"**. However, this has **yet to be validated**.

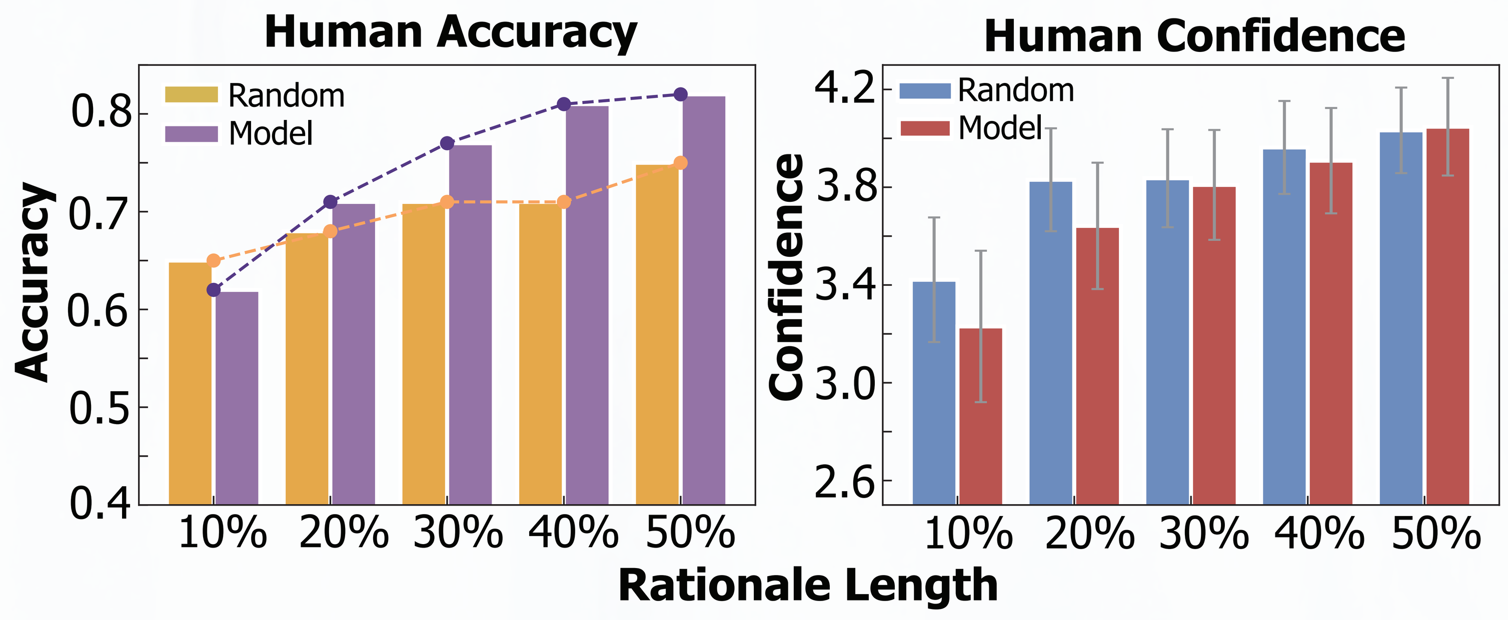
2. Research Object

Is the shortest rationale indeed the most human-understandable?

Our **goal** is to study the unexplored effect of **rationale length** on **human understanding**.

4. Results and Key Findings

We find that ***shortest rationales are largely NOT the best for human understanding***.



- Humans get worse **prediction accuracy** and **confidence** when rationales are **too short** (e.g., 10% length) than random baseline.
- The **eventually flattened slope** of model's accuracy potentially suggests a sweet spot to **balance human understanding on rationale length and model accuracy**.

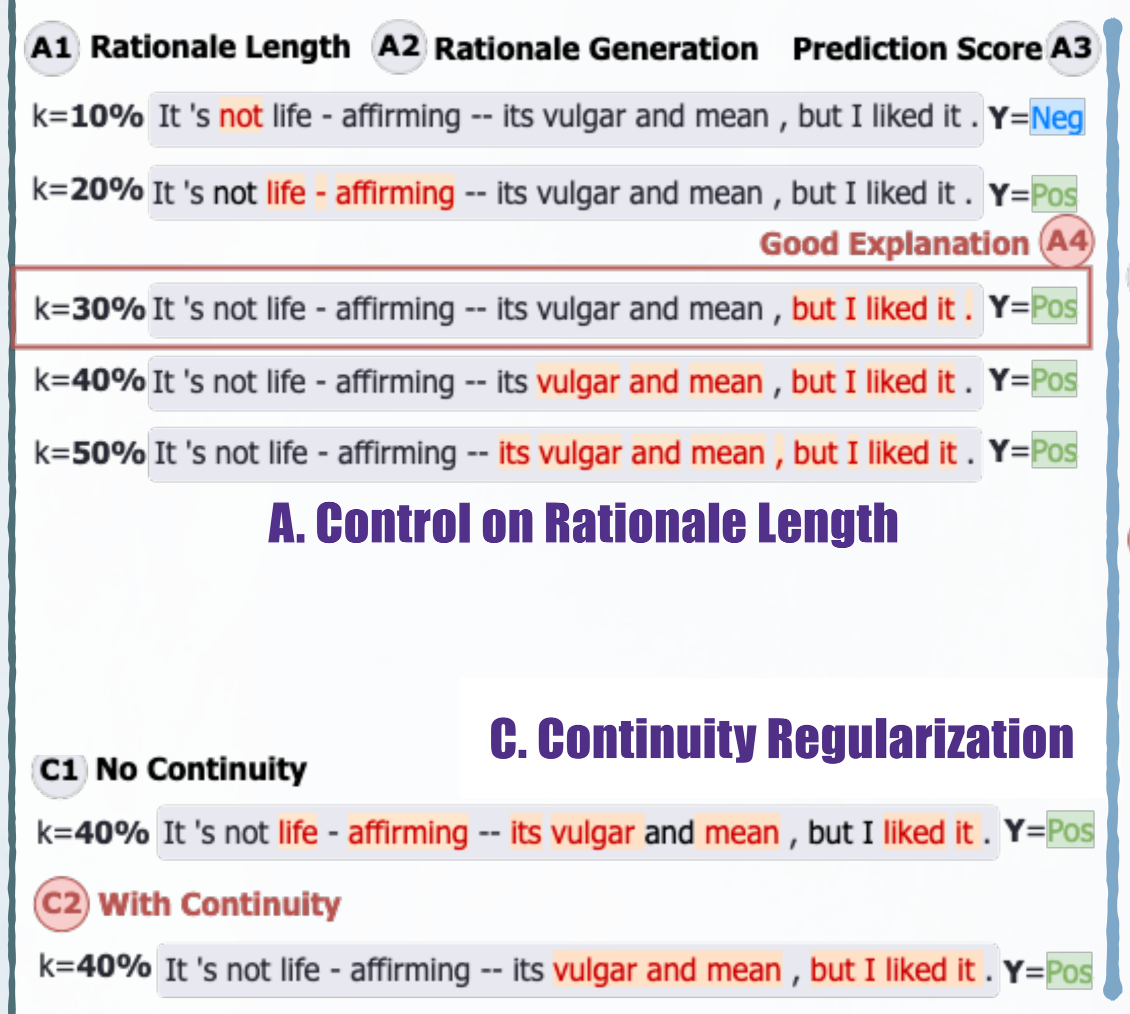
length level (%) & Extract. method		Negative P / R / F1	Positive P / R / F1
10%	LimitedInk	0.66 / 0.56 / / 0.61	**0.70** / 0.58 / 0.64
	Random	**0.67 / 0.57 / 0.62**	0.66 / **0.70 / 0.68**
20%	LimitedInk	**0.75 / 0.61 / 0.67**	**0.71 / 0.77 / 0.74**
	Random	0.69 / 0.60 / 0.64	0.68 / 0.74 / 0.71
30%	LimitedInk	**0.74 / 0.76 / 0.75**	**0.81 / 0.78 / 0.79**
	Random	0.72 / 0.61 / 0.66	0.72 / 0.78 / 0.75
40%	LimitedInk	**0.84 / 0.76 / 0.80**	**0.78 / 0.85 / 0.81**
	Random	0.79 / 0.63 / 0.70	0.65 / 0.79 / 0.71
50%	LimitedInk	**0.78 / 0.78 / 0.78**	**0.85 / 0.84 / 0.85**
	Random	0.77 / 0.63 / 0.70	0.75 / 0.84 / 0.79

Human performance on predicting model labels of **each category**, including **P**recision / **R**ecall / **F1** Score.

3. Methodology

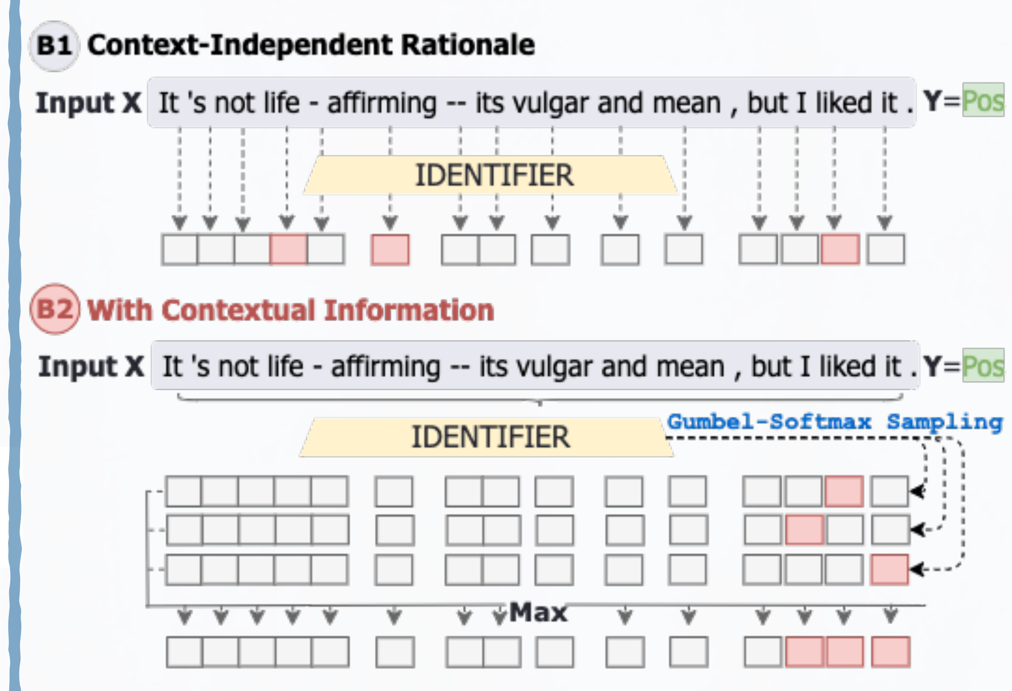
LimitedInk: A self-explaining model with Rationale Length Control

i) Design a novel self-explaining model, **LimitedInk**, to **control rationale length**.



A. Control on Rationale Length

C. Continuity Regularization



B. Contextual Rationale Generation

LimitedInk Performance

Method	Movies Task	Movies P	Movies R	Movies F1	BoolQ Task	BoolQ P	BoolQ R	BoolQ F1	Evidence Inference Task	Evidence Inference P	Evidence Inference R	Evidence Inference F1	MultiRC Task	MultiRC P	MultiRC R	MultiRC F1	FEVER Task	FEVER P	FEVER R	FEVER F1
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	**.41**	.22	.83	.35	.52	.42
Sparse-IB	.84	.21	.42	.28	.46	**.17**	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	**.37**	.50	**.43**
LimitedInk	**.90**	**.26**	**.50**	**.34**	**.56**	.13	**.17**	**.15**	**.50**	**.04**	**.27**	**.07**	**.67**	**.22**	.40	**.28**	**.90**	.28	**.67**	.39
Length Level	50%				30%				50%				50%				40%			

LimitedInk performs compatible with baselines in 5 ERASER text classification benchmark datasets: w.r.t. rationale metrics:

- end-task performance (**Task**, weighted average F1);
- human annotated rationale agreement (**P**recision, **R**ecall, **F1**).

Human Study with LimitedInk

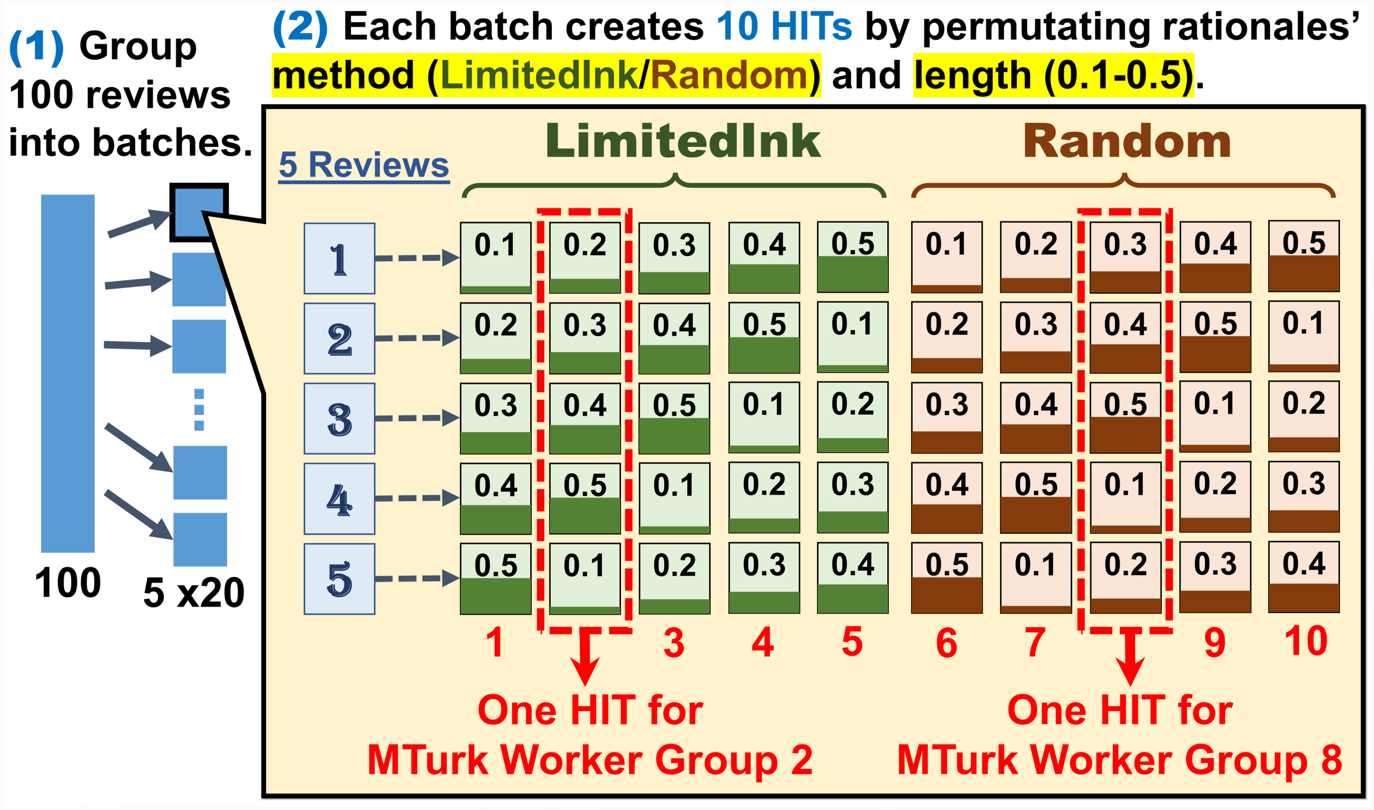
ii) Conduct **human study** to examine the **effect** of **rationale length** on **human understanding**.

Key Components of the User Interface

Ask MTurk workers:

- **predict movie reviews' sentiment** polarities
- based only on **rationales**.

- **Random Baseline:** randomly select rationale tokens of the same length with LimitedInk's rationale.
- **Participant Control:** to prevent workers from seeing same reviews repeatedly, we strictly control participant recruiting and grouping.



The Workflow of Human Evaluation

5. Key Insights

- ★ Future work could more cautiously define the best rationales for human understanding, then **find the right balance between model accuracy and rationale length.**
  - ✳ More concrete, one promising way could be to clearly define the **optimal human interpretability in a measurable way** and then learn to **adaptively select rationale with appropriate length.**

Open-source code: https://github.com/huashen218/LimitedInk.git