Dancing Between Success and Failure: Edit-level Simplification Evaluation using **SALSA**

David Heineman, Yao Dou, Mounica Maddela, Wei Xu

Georgia Tech

salsa-eval.com

① Linguistic Evaluation by the *Edit*

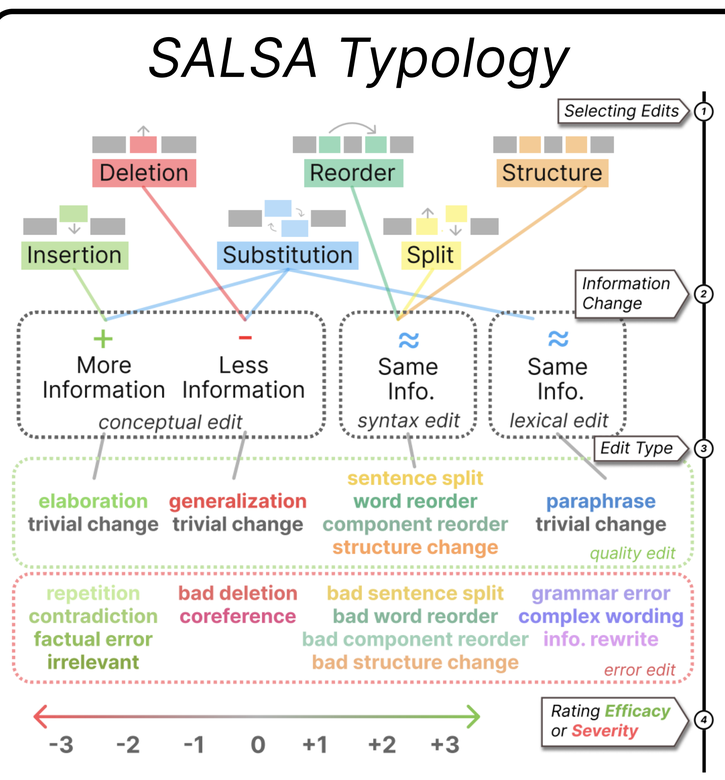
- We taxonomize language model *behavior* using span-level **error** *and* **quality** evaluation for text simplification
- Collect annotations with our resulting typology (**S**uccess and F**A**ilure **L**inguistic **S**implification **A**nnotation) across SOTA simplification systems
- We use SALSA for *model analysis*, *new automatic metrics*, and even *error identification*

SALSA Edit-level Dataset:

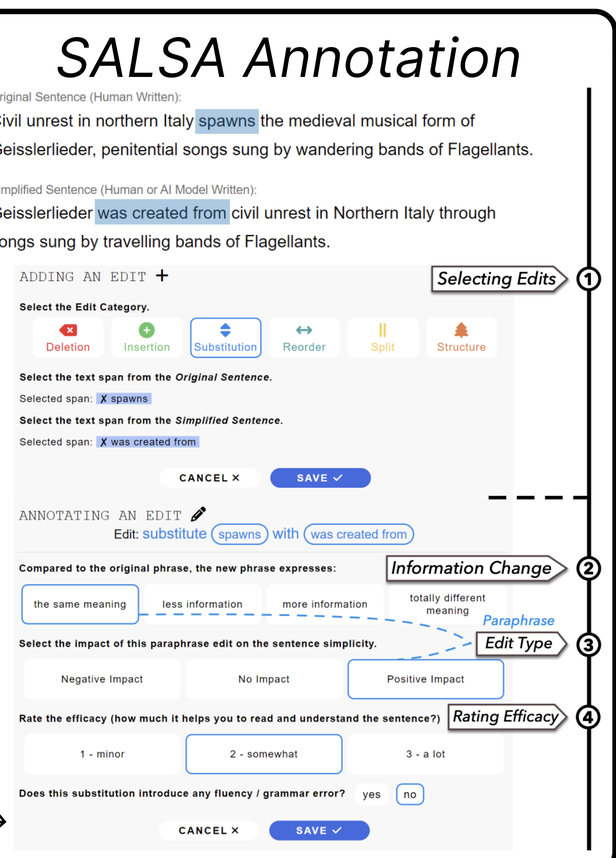
21	19K	840	8
quality / error edit types	annotations	simplifications	simplification systems

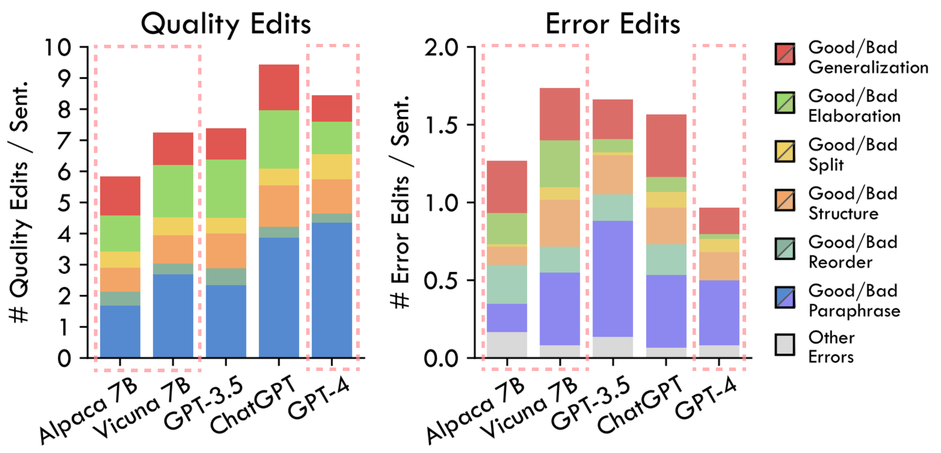
② The **SALSA** typology

- **Edit Selection.** Identify spans corresponding to six primitive operations: Insertion, Deletion, Substitution, Reorder, Structure and Split
- **Information Change**. Identify whether the simplification is *conceptual*, *syntactic* or *lexical*
- **Edit Type.** Classify edits into *21* fine-grained types by answering decision tree-style questions
- **Edit Rating.** Rate efficacy/severity [-3, 3]



Customize this interface at `thresh.tools/salsa` →





③ **Edit Analysis** of LLMs (GPT-4, Alpaca, ...)

- LLMs **elaborate** with world knowledge, while humans more often **generalize**
- Human simplification uses longer and more effective edits across all operation types
- Fine-tuned simplification models (e.g., T5-11B) write far too conservative simplifications

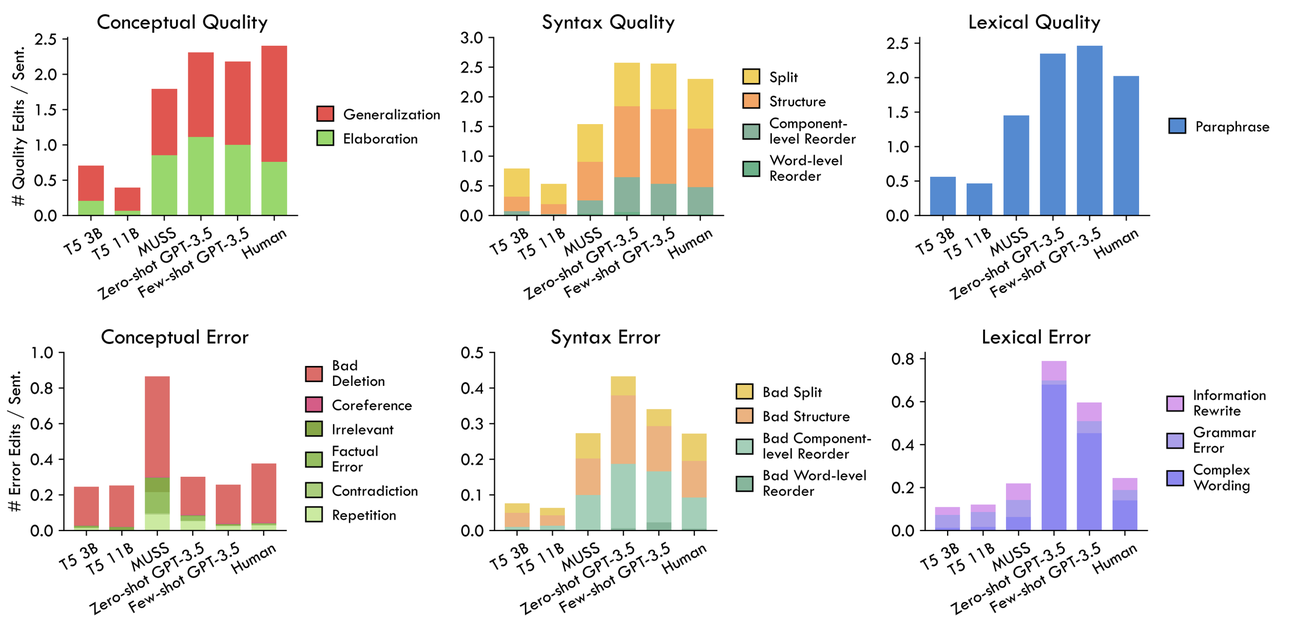
EXAMPLE *Few-shot GPT-3.5*
After defeating PSD candidate Viorica Dăncilă by a landslide in 2019, **his** second term..
In 2019, Klaus Iohannis defeated PSD candidate Viorica Dăncilă by a large margin. His second term..

EXAMPLE (*coreference error*) *ChatGPT*
The International Court of Justice (ICJ) rules that the United States violated its ...
The ICJ said that the US broke its ...

- **Open-source LLMs** perform a similar distribution of *quality* edits compared to GPT-3.5, but at the cost of *more conceptual errors*
- GPT-4 demonstrates a large error improvement in *content edits*, but still struggles with **paraphrasing**
- Demonstrates upper-limits on imitating simplification ability via synthetic instruction fine-tuning

EXAMPLE (*information rewrite*) *Alpaca 7B*
... justifies a runtime nearing 3 hours (with a post-credits scene, no less), and it already opened to over $100 million worldwide.
.. takes up almost 3 hours of the movie. The movie opened to over $100 million worldwide. A post-credits scene completes the story.

EXAMPLE (*component reorder*) *ChatGPT*
Poland announces the closure of a major border crossing with Belarus "until further notice" amid heightened tensions between the two countries.
Poland has closed a big border crossing with Belarus due to increased tensions between the two countries. The closure will remain in effect until further notice.



④ **LENS-SALSA**: Automatic Evaluation Metric

- Train on **dual span- / sentence-level** objective using SALSA quality and error annotations
- Evaluate metric across dimensions of held-out SALSA test set

		BLEU	SARI	BERTSCORE	COMET-MQM	LENS	LENS-SALSA
Quality	Lexical	-0.185	0.030	0.015	0.086	**0.289**	0.284
	Syntax	-0.117	0.097	0.008	0.024	0.206	**0.244**
	Conceptual	-0.240	-0.147	-0.325	-0.187	-0.006	**0.173**
Error	Lexical	-0.259	-0.162	-0.134	-0.004	-0.059	**0.015**
	Syntax	-0.147	-0.094	-0.136	-0.073	-0.042	**-0.013**
	Conceptual	-0.128	-0.099	-0.293	-0.169	-0.016	**0.062**
All	All Error	-0.263	-0.190	-0.329	-0.170	-0.035	**0.046**
	All Quality	-0.201	0.056	-0.018	0.033	0.304	**0.318**
	All Edits	-0.286	-0.035	-0.235	-0.129	0.266	**0.336**

Use LENS-SALSA for evaluation at:
`huggingface.co/davidheineman/lens-salsa`

Our reference-free metric LENS-SALSA!

⑤ Automatic **Edit Quality Estimation**

Task. Identify *quality* or *error* spans given a full simplification

Method. QA-based word aligner with FFNN for each span type

Promising baselines method for downstream tasks and future work!

Method	Quality	Error	Ok	Average
End-to-end				
Tag	67.00	28.24	92.88	62.71
Tag-ML	**70.73**	30.06	**93.09**	64.62
Two-stage (use word aligner to get edit information)				
Tag-EI	69.09	30.37	93.04	64.17
Ec-Sep	64.87	36.15	91.56	64.20
Ec-One	68.77	**39.50**	91.91	**66.73**
Oracle (Ec-One)	88.31	69.44	98.35	85.47