GEORGETOWN UNIVERSITY

GUMSum: Multi-Genre Data and Evaluation for English Abstractive Summarization

Yang Janet Liu and Amir Zeldes
Department of Linguistics, Georgetown University
{yl879, amir.zeldes}@georgetown.edu

SCAN ME

ACL 2023 @
Toronto, Canada

GUMSum

- a challenge test containing 213 summaries covering 12 genres for English Abstractive Summarization
  - based on the English UD treebank, GUM (Zeldes, 2017), containing 200K tokens and gold entity annotations
- human-written summaries following **carefully defined guidelines** targeting the follow three goals:
  - **substitutive**: i.e., informative, functioning as a substitute for reading a text (c.f. Edmundson, 1969; Nenkova and McKeown 2011) rather than indicative (e.g., 'clickbait' designed to attract readership)
  - **faithful**: adhering to original formulations wherever possible
  - **hallucination-free**:
    - summaries make a strong effort not to add any information (even if it is likely to be true),
    - mentioning only entities and events actually contained in the text
- a double-annotated subset (24 docs in the GUM test of UD release) to support inter-annotator agreement and multiple-reference evaluation

Genres	Source	Docs	Toks	øSum.Len (sd)
Interviews	Wikinews	19	18,190	49 (6.3)
News stories	Wikinews	23	16,145	51 (9.0)
Travel guides	Wikivoyage	18	16,514	59 (8.9)
How-to guides	WikiHow	19	17,081	67 (6.5)
Academic	various	18	17,169	35 (11.2)
Biographies	Wikipedia	20	18,213	44 (9.8)
Fiction	various	19	17,510	47 (10.3)
Web forums	Reddit	18	16,364	50 (8.7)
Conversations	SBC	14	16,416	41 (13.7)
Speeches	various	15	16,720	46 (9.2)
Vlogs	YouTube	15	16,864	50 (11.8)
Textbooks	OpenStax	15	16,693	51 (8.9)
total / average		213	203,879	50 (12.2)

Table 1: Overview and Statistics of GUMSum.

General and Genre-specific Guidelines

General Guidelines: a domain-general, substitutive, maximally concise format constrained to:

[1] have at most one sentence / 380 characters

[2] have the goal of replacing reading the text

[3] give participants / time / place / manner of events

[4] form a sentence rather than a fragment

[5] Omit distracting information

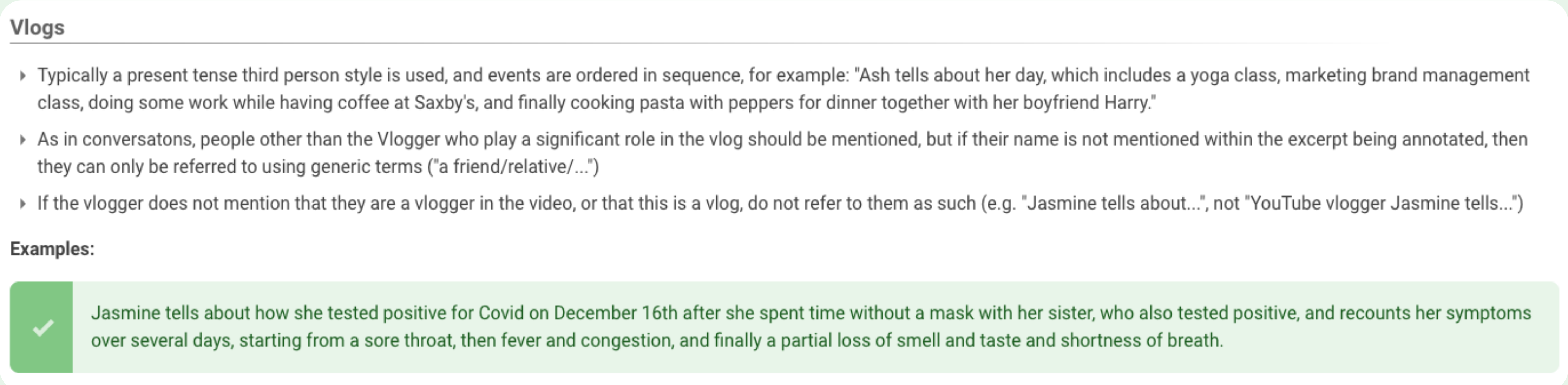
[6] avoid entities or information not present in the text, even if we are fairly sure it is true

[7] reject synonyms for words in the text

✅ On March 23, 1999, five bank robbers plundered the vault of First National Bank in Poughkeepsie, NY and escaped in a bus they had stolen. >> follows [1][3][4]

🙅 Bank robbers plundered a vault and escaped. >> violates [3]: underspecified

🙅 Bank robbers who robbed a bank in Poughkeepsie were never caught by police.
>> violates [6]: introducing an entity NOT in the original text
>> violates [7]: substituting 'robbed' for 'plundered', a deviation from the original text's style



Vlogs

- Typically a present tense third person style is used, and events are ordered in sequence, for example: "Ash tells about her day, which includes a yoga class, marketing brand management class, doing some work while having coffee at Saxby's, and finally cooking pasta with peppers for dinner together with her boyfriend Harry."
- As in conversatons, people other than the Vlogger who play a significant role in the vlog should be mentioned, but if their name is not mentioned within the excerpt being annotated, then they can only be referred to using generic terms ("a friend/relative/...")
- If the vlogger does not mention that they are a vlogger in the video, or that this is a vlog, do not refer to them as such (e.g. "Jasmine tells about...", not "YouTube vlogger Jasmine tells...")

Examples:

✓ Jasmine tells about how she tested positive for Covid on December 16th after she spent time without a mask with her sister, who also tested positive, and recounts her symptoms over several days, starting from a sore throat, then fever and congestion, and finally a partial loss of smell and taste and shortness of breath.

📹 ✅ Ash tells about her day, which includes a yoga class, marketing brand management class, doing some work while having coffee at Saxby's, and finally cooking pasta with peppers for dinner together with her boyfriend Harry.

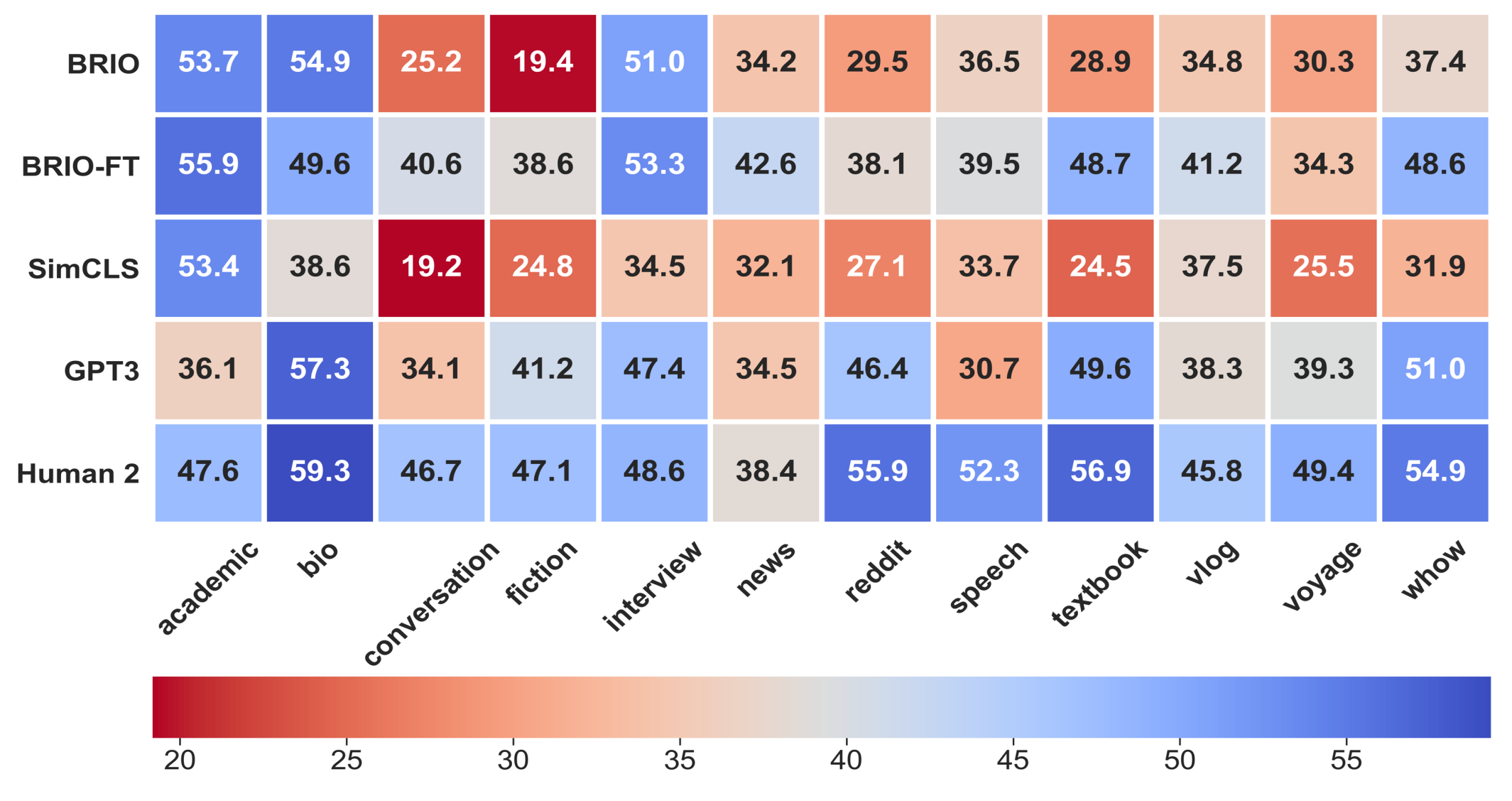
Automatic Evaluation

- system outputs from two recent supervised systems, BRIO (Liu et al., 2022) and SimCLS (Liu and Liu, 2021), as well as prompt-based outputs using a GPT3 model (Brown et al., 2020), GPT3-text-davinci-002 (GPT3-DV2), with the prompt 'Summarize the text above in one sentence.'
- BRIO-FT: fine-tune BRIO's trained-model on XSum

	R-1	R-2	R-L	BS	MS	METEOR	BLEU	BLEURT
SimCLS	23.1	6.2	17.2	86.0	12.1	13.4	2.1	31.9
BRIO	27.8	10.2	21.2	87.2	15.9	18.0	3.7	36.3
GPT3-DV2	31.1	**12.1**	25.1	88.5	21.1	20.8	3.8	42.2
BRIO-FT*	**37.3**	12.0	**27.1**	**88.7**	**27.4**	**27.6**	**6.1**	**44.3**
Human 2	**38.9**	**12.7**	**28.4**	**88.8**	**28.5**	**33.0**	**7.5**	**50.2**

Table 2: Automatic Evaluation Metrics of System Outputs and Human Agreement (* = 3 run average).

- All systems exhibited some hallucinations and unfaithfulness.
- Fine-tuning was particularly helpful for summarizing ***conversations***, ***Reddit***, and ***how-to guides***, on which all systems struggled.
- For humans, genre differences were much less pronounced, with lowest scores surprisingly for ***news***.



	academic	bio	conversation	fiction	interview	news	reddit	speech	textbook	vlog	voyage	whow
BRIO	53.7	54.9	25.2	19.4	51.0	34.2	29.5	36.5	28.9	34.8	30.3	37.4
BRIO-FT	55.9	49.6	40.6	38.6	53.3	42.6	38.1	39.5	48.7	41.2	34.3	48.6
SimCLS	53.4	38.6	19.2	24.8	34.5	32.1	27.1	33.7	24.5	37.5	25.5	31.9
GPT3	36.1	57.3	34.1	41.2	47.4	34.5	46.4	30.7	49.6	38.3	39.3	51.0
Human 2	47.6	59.3	46.7	47.1	48.6	38.4	55.9	52.3	56.9	45.8	49.4	54.9

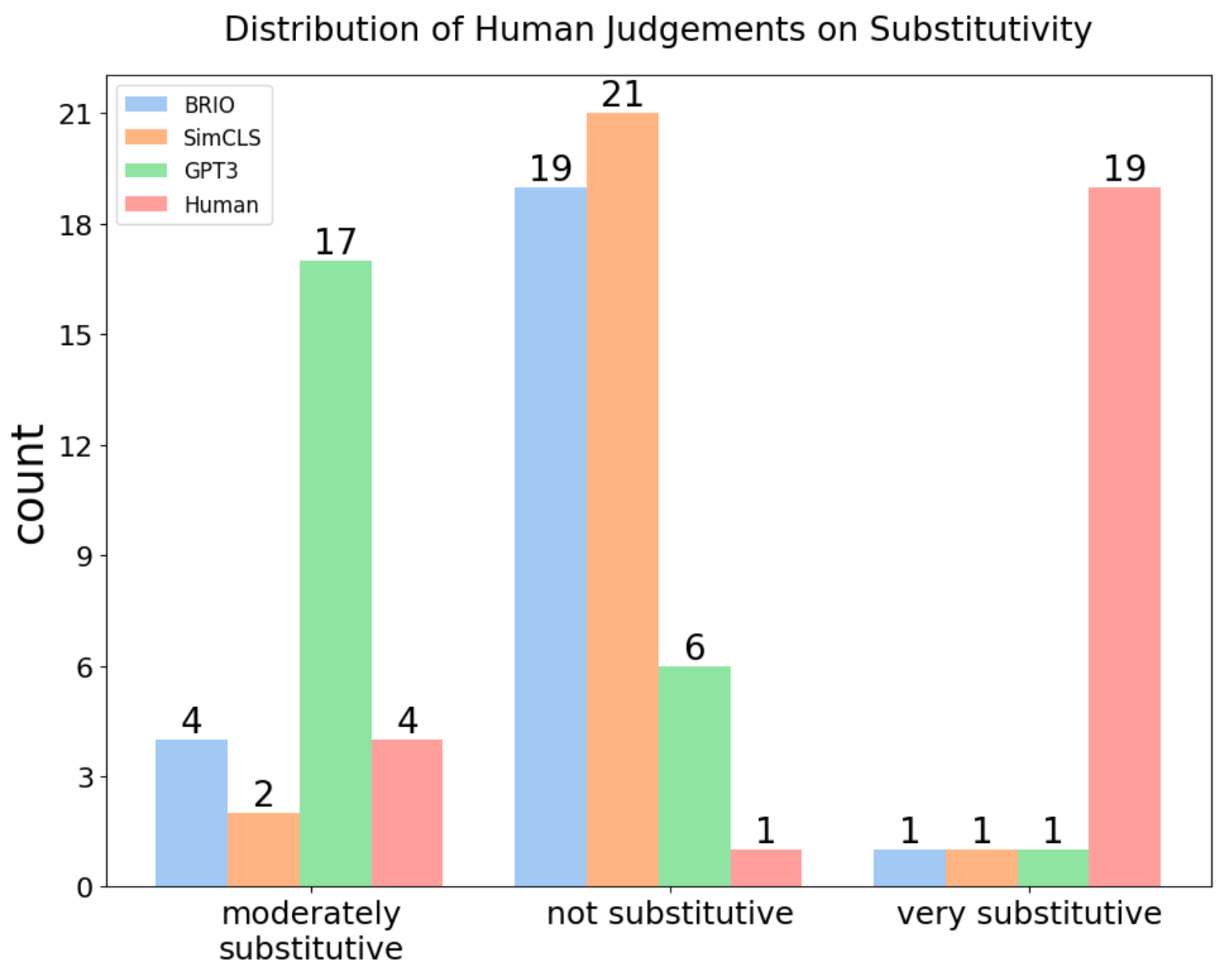
BLEURT Scores by Genre for Each Scenario (whow=how-to guides, voyage=travel guides).

References

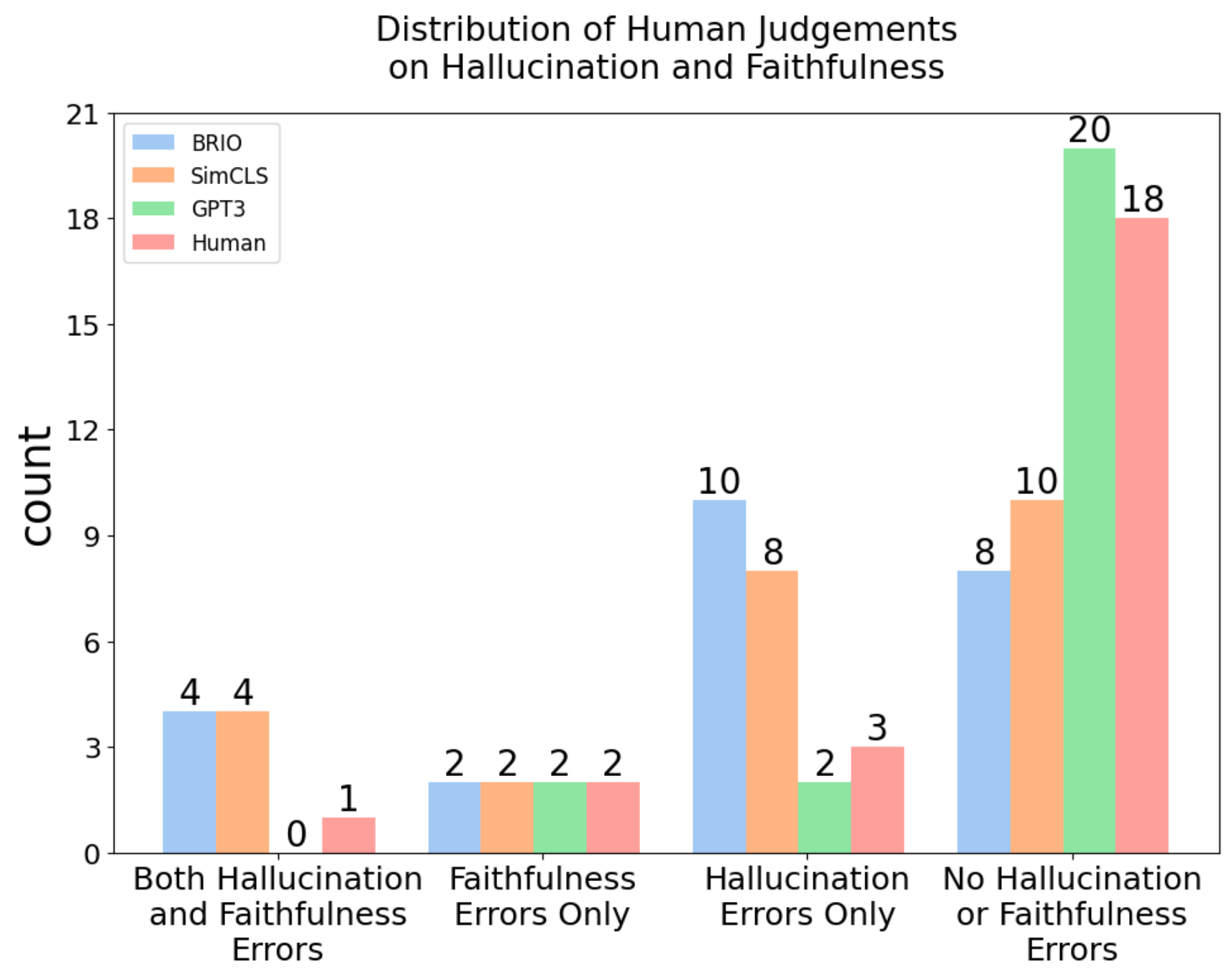
- H. P. Edmundson. 1969. New Methods in Automatic Extracting. *J. ACM*, 16(2):264–285.
- Ani Nenkova and Kathleen R. McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.
- Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.
- Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3): 581–612.

Human Evaluation & Qualitative Analysis

- asked 12 Linguistics students ($20.29/hour; 1.5 hours) to evaluate the full texts and the summaries of the 24 documents in the test set of GUM V9 (source) and to produce an additional summary for their assigned texts
  - substitutivity, faithfulness, hallucination, human preferences, and English quality
  - open-ended questions regarding difficulties or challenges encountered when writing a one-sentence summary for the assigned genres



Distribution of Human Judgements on Substitutivity



Distribution of Human Judgements on Hallucination and Faithfulness

conversation

BRIO In our series of letters from African journalists, novelist and writer Adaobi Tricia Nwaubani tells the story of a woman who went after a man in a dance class.

BRIO-FT Harold tells a story about a dance class he went to at Bahia on Sunday, in which a woman goes after a man who is also in the class, but does not want to dance with him.

SimCLS The following is the full text of a speech by Nigerian writer and novelist, Adaobi Tricia Nwaubani, at the opening of the new book, The God of Small Things.

GPT3-DV2 The speaker is telling a story about a woman who made sexual advances towards a man she didn't know at a club, and how other people in the club reacted.

Human 1 While discussing plans to go to a dance class in Bahia to learn lambada dancing, Miles tells his friends about an attractive woman he had seen dancing lewdly with a man there.

Human 2 Jamie is curious about going to a Lambada class Miles has been to and asks him what it's like, and the story he tells about the stupefying promiscuity he witnessed at a past class ends up convincing Jamie she shouldn't attend.

news

BRIO The US space agency has announced where the four remaining Space Shuttle orbiters will be permanently displayed following the end of the space shuttle program.

BRIO-FT on the eve of the 30th anniversary of the launch of the first space shuttle, NASA Administrator Charles Bolden announced the locations for the three remaining orbiters, Enterprise, Discovery and Atlantis, which will be placed in museums across the country.

SimCLS The US space agency has announced the locations that will house the three remaining Space Shuttle orbiters following their retirement later this year..,

GPT3-DV2 The locations of the three remaining space shuttle orbiters were announced by NASA Administrator Charles Bolden on the 30th anniversary of the first shuttle launch.

Human 1 On April 13, 2011, NASA celebrated the Space Shuttle program's 30th anniversary and upcoming final mission at the Kennedy Space Center in Cape Canaveral, Florida, and announced that the remaining three shuttles would be relocated to museums in New York, Washington and Los Angeles, prompting critcism from competing sites.

Human 2 NASA Administrator Charles Bolden announces where Space Shuttles Enterprise, Discovery, Endeavor, and Atlantis will be retired following the end of the Space Shuttle Program, in which some people were disappointed with the final location decisions.

- **Hallucinations from GPT3-DV2, BRIO, and SimCLS were more pronounced:**
  - designating a speaker mentioning retirement as an attendee of a seminar about retirement, which was not mentioned
  - adding to a textbook excerpt on the Civil War by calling it the longest, most expensive conflict in US history
- **Human violations in hallucination / faithfulness were rare and subtle, resulting from evaluators adhering to guidelines very literally:**
  - a human summary's use of the pronoun 'she' in reference to a vlogger whose pronouns had not been stated is a form of hallucination
  - a mention of 'Washington' in a news article was a faithfulness issue due to the ambiguity of the place resulting from without specifying 'DC'
- Typical hallucinated lead about journalists in the BRIO output (conversation, top)
- Systems failed to capture the site controversy in the 2nd half of the document while both humans did (news, bottom)