GUMsley: Evaluating Entity Salience in Summarization for 12 English Genres

Jessica Lin and **Amir Zeldes**
Department of Linguistics
Georgetown University
{yl1290, amir.zeldes}@georgetown.edu

Abstract

As NLP models become increasingly capable of understanding documents in terms of coherent entities rather than strings, obtaining the most salient entities for each document is not only an important end task in itself but also vital for Information Retrieval (IR) and other downstream applications such as controllable summarization. In this paper, we present and evaluate GUMsley, the first entity salience dataset covering all named and non-named salient entities for 12 genres of English text, aligned with entity types, Wikification links and full coreference resolution annotations. We promote a strict definition of salience using human summaries and demonstrate high inter-annotator agreement for salience based on whether a source entity is mentioned in the summary. Our evaluation shows poor performance by pre-trained SOTA summarization models and zero-shot LLM prompting in capturing salient entities in generated summaries. We also show that predicting or providing salient entities to several model architectures enhances performance and helps derive higher-quality summaries by alleviating the entity hallucination problem in existing abstractive summarization.

1 Introduction

The task of *salient entity extraction* (SEE) is to identify entities that are central to a document's overall meaning. Previous work on SEE has relied on crowdsourcing (Dojchinovski et al., 2016) or user statistics on the web (e.g. clickstream data, Gamon et al. 2013) to derive salience labels for entities. In this study, we extend an approach from Dunietz and Gillick (2014), who considered an entity salient if it also appears in a human-written summary or abstract of a news article, and we cover many further genres rather than just news. Figure 1 shows an example of salient entities in a *conversation* annotated according to our definition of salience.

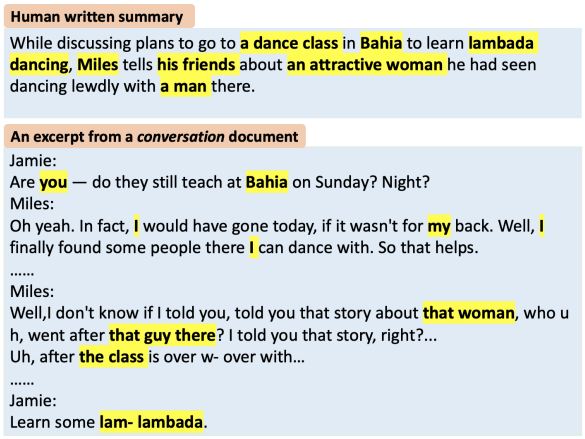

Figure 1: A salient entity example from our data. Salient entity mentions are highlighted in yellow.

SEE is increasingly important as NLP systems move from understanding document 'aboutness' at the word level (e.g. keyword extraction) (Tomokiyo and Hurst, 2003) to entity level document understanding (Maddela et al., 2022; Nan et al., 2021). Therefore, a dataset with SEE labels can benefit downstream applications such as information retrieval and summarization, which extract salient information from large documents and prioritize specific entities in controllable models.

Although several SEE datasets already exist (Dojchinovski et al., 2016; Dunietz and Gillick, 2014; Gamon et al., 2013; Trani et al., 2018; Wu et al., 2020), most are predominantly collected from news articles and derive labels using crowdsourcing or "found" information such as hyperlinks, which are not intended to annotate salience per se. This has two major limitations: First, crowdsourcing SEE without rigorous training and clear definitions of salience may be biased towards individuals and inconsistent interpretations of what is considered salient. Second, focusing on news limits system performance on more diverse data (e.g. conversation, vlogs, etc.).

To investigate the role of SEE in tasks such as summarization, previous entity-centric work (Fan et al., 2018; He et al., 2022; Xiao and

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2575–2588

Carenini, 2022) has compared summaries generated by entity-aware methods with generic summarization methods qualitatively. As part of our evaluation, we combine manual and automatic, qualitative and quantitative analyses to assess SEE impact on several approaches to summarization.

In this paper, we therefore present and evaluate a gold standard dataset manually annotated with SEE labels, by identifying all entities that appear in a human-written summary as salient, making the task less subjective. Our dataset, called GUMsley (**GUM** **s**alient **l**inked **e**ntit**y** corpus) is based on the existing UD English GUM corpus (Georgetown University Multilayer corpus, Zeldes 2017) and goes beyond other entity salience datasets in covering all named and non-named salient entities for 12 genres of English text. GUMsley also enables the evaluation of SEE annotations in a broad spectrum of genres and tasks, since the data contains Wikification identifiers for named entities, as well as comprehensive coreference resolution. Our results show that a significant amount of salient entities are not captured by SOTA abstractive summarization models or out-of-the-box LLMs (Section 5.1). We also conduct a quantitative analysis to show that providing gold or even predicted salient entities to models helps to generate a higher quality summary (Section 5.2).

2 Related Work

Entity Salience Datasets The growing interest in SEE is demonstrated by increasing numbers of annotated datasets, with different approaches to recognizing entities and assigning labels. The first step is usually entity identification. While some datasets (Dunietz and Gillick, 2014; Gamon et al., 2013) apply a multi-step NLP pipeline (NP extraction, coreference resolution, possibly a named entity resolver) to pinpoint entities, others (Dojchinovski et al., 2016; Trani et al., 2018; Wu et al., 2020) have done so manually. Since pipelines may propagate errors to later steps, full manual annotation is used in our study to avoid such issues. To collect salience labels, most studies (Dojchinovski et al., 2016; Trani et al., 2018) prefer human annotation using crowdsourcing. Although crowdsourcing may outperform automated methods, it is inevitably noisy and can suffer from *subjective bias* issues (Maddela et al., 2022) since people have different judgments on what they consider *salient*. In this study, we follow a more regimented approach similar to the NYT salience corpus (Dunietz and Gillick 2014), which considers entities that also appear in an abstract or summary as salient. Unlike Dunietz and Gillick, which uses automatic coreference resolution to detect mentions in newspaper summaries, we use fully manual annotation coupled with GUM's carefully written summaries which have consistent guidelines and style across 12 English genres (Liu and Zeldes, 2023), rather than found abstracts or teasers limited to news or academic data.

Entity-centric Summarization Research on entities in automatic summarization has seen a surge of interest in the NLP community recently. However, numerous studies (Cao et al., 2018; Kryscinski et al., 2019; Nan et al., 2021) have pointed out that abstractive summarization models suffer from entity hallucination, i.e. summaries contain entities that never appear in the source document. Previous attempts to solve such problems include training models to classify whether generated summaries are factually consistent with input documents (Kryscinski et al., 2019) and filtering out entities that have no match in the source document (Nan et al., 2021; Xiao and Carenini, 2022). In this study, we propose adapting methods from controllable summarization (Fan et al., 2018; Nan et al., 2021), which enable users to specify for example keywords to control information included in generated summaries, and help provide a better quality summary with fewer hallucinated entities.

Unlike our approach, previous controllable summarization methods (Fan et al., 2018; He et al., 2022) are often evaluated compared to generic summarization methods through qualitative analysis by human evaluators, which may suffer from biases. In this study, we combine qualitative and quantitative metrics of factual consistency at the entity level following Nan et al. (2021) and Xiao and Carenini (2022). We analyze the factual quality of summaries using controllable entity-centric methods compared with generic supervised methods and prompt-based methods.

3 Annotation process

Our dataset, GUMsley, is based on the open access GUM (Zeldes, 2017), a manually annotated multilayer corpus with Universal Dependencies (UD) parses (de Marneffe et al., 2021), entity information (entity types, Wikification links and more), coreference resolution and discourse parses, as

well as a human-written summary for each document. Data comes from 12 text types and covers over 200K tokens (see Table 1).

GUMsley adds a layer of entity salience labels to all named and non-named entities in GUM, annotated by three trained experts (PhDs/PhD students in Computational Linguistics). The main goal is to annotate a subset of entities as salient if they are mentioned in the summaries, regardless of subjective judgments about their importance. Annotators are asked to look at the source document as well as the human-written summary first, and then make a binary judgment on every mention as to whether it is in both. In this way, annotators mark a judgment for every mention in both documents. This approach, which goes back to Dunietz and Gillick 2014, assumes that if something is salient it should appear in the summary, and conversely, if it appears in the summary, it must be salient, since summaries are meant to be as short and informative as possible. This assumption is mirrored in GUM's summary guidelines[1]. We reason that while this approach could over-generate, it should have high recall, since it would be hard to summarize a document while omitting salient entities. Despite this, we note that our approach still flags only a fraction of entities as salient (around 7%, see Table 1). Using the gold standard manual coreference clusters, we ensure that all mentions of each cluster (=entity) are included as salient mentions, meaning annotations are consistent with the coreference layer in terms of entities.

We also double-annotated 21,770 tokens of the data corresponding to the 24 test documents of the UD release of GUM, containing 3,283 entities ($\approx$ 10% of the data in Table 1) with binary SEE labels (salient vs. non-salient). To measure inter-annotator agreement (IAA), we compute raw percentage agreement and Cohen's κ agreement at entity level for all 12 genres (if any mention of the entity is considered salient, the entity is considered salient). Since entity mention spans are given in GUM, our task only involves matching such spans to the summaries, and we achieve very high IAA across 12 genres (0.981 for raw agreement and 0.978 for Cohen's κ agreement), with most of the texts achieving an agreement score over 0.9 (see the genre-breakdown IAA scores in Table 2). While this indicates very reliable results, annotators did disagree on some difficult cases:

- **Canonical mentions**: Some entities are mentioned in the summary differently than in the text, which sometimes makes it hard for annotators to locate the right salient mention in the text. This usually happens when the entity is a singleton (only being mentioned once in the document). For example, one summary mentions "demographic information about the respondents", which does not appear in the text. In this case, annotators flagged the mention "Demographic variables" in the text as salient, since it was judged to refer to the same thing in context. By contrast, another summary mentioned "the history of the concept of atoms", but the nearest mention in the text, "early ideas in Atomic Theory" was deemed not equivalent in its denotation.

- **Lack of explicit speaker information**: This type of issue occurred frequently in conversations, where no explicit speaker information is given in the text supplied to annotators, who needed to track who is speaking. For example, if a summary mentions "Miles tells his friends about...", then all interlocutors (Miles and his friends) should be marked as salient. However, the pronouns (*I* and *you*) in the conversation do not unambiguously indicate 'who is talking to whom'. In this case, annotators were provided with gold speaker information from the dataset to help make the right decision.

- **Non-nominal mentions**: According to our guidelines[2], for an entity to be considered salient it must be (i) a markable mention in the source document, which include referential NPs and verbal markables (if they are coreferred to by an NP)[3] (ii) a verbal event in the summary coreferent with a nominal event in the source document. If the entity is mentioned in the source document as a non-nominal mention, then it is only considered salient if it is referred back to by a pronoun or noun. For example, in Figure 1 the summary mentions "dancing lewdly", which corefers with a non-nominal mention "the whole dance" in the source document. In this case, "dancing lewdly" will not be

[1] https://wiki.gucorpling.org/gum/summarization

[2] https://wiki.gucorpling.org/gum/salience

[3] See entity annotation guidelines here: https://wiki.gucorpling.org/gum/entities

	Documents	Mentions	Entities	% Salient Entities	Avg # of Entities	Tokens
academic (ac)	18	5,046	3,067	3.13	170	17,169
bio (bi)	20	5,768	3,326	5.11	166	18,213
conversation (cn)	14	4,094	1,352	9.62	97	16,416
fiction (fc)	19	4,974	2,344	7.04	123	17,510
interview (it)	19	5,211	2,604	5.18	137	18,190
news (nw)	23	4,720	2,544	11.08	111	16,145
reddit (rd)	18	4,543	2,302	4.13	128	16,364
speech (sp)	15	4,847	2,550	5.88	170	16,720
textbook (tx)	15	4,719	2,881	5.41	192	16,693
vlog (vl)	15	4,498	1,629	11.42	109	16,864
voyage (vy)	18	4,471	2,952	7.79	164	16,514
whow (wh)	19	4,468	2,348	11.33	124	17,081
Total	213	57,359	29,899	7.26	146	203,879

Table 1: Overview of GUMsley. % salient entities = number of salient entities / total number of entities; Avg entities per summary = # of entities / # of documents in genre.

	Percentage/ Cohen's κ agreement		
ac	0.9979/0.9983	**rd**	0.9928/0.9906
bi	0.9846/0.9840	**sp**	0.9913/0.9897
cn	0.9780/0.9666	**tx**	0.9942/0.9931
fc	0.9684/0.9621	**vl**	0.9993/0.9983
it	0.9860/0.9834	**vy**	0.9976/0.9967
nw	0.8955/0.8889	**wh**	0.9902/0.9869
Total	0.9813/0.9782		

Table 2: Genre-breakdown inter-annotator agreement on the GUMsley test set at entity level.

marked as salient because "the whole dance" is not coreferred to by a pronoun in the document.

- **Aggregate and specific mentions**: When documents enumerated the members of an aggregate set mentioned in the summary but not the document, we decided to include the members as salient. For example, "the remaining three shuttles" are mentioned in one summary, while the document contains the three specific shuttles ('Space shuttle Endeavour', 'Discovery', and 'Enterprise'). These are thus all marked as salient.

4 Experimental setup

In order to evaluate the usefulness of SEE annotations we apply our data to the task of automatic summarization and test 1) whether system summaries capture gold salient entities identified by humans, and 2) whether SEE information can improve summarization quality. We evaluate the following models:

BRIO BRIO (Liu et al., 2022b) is a recent SOTA abstractive summarization model, trained and fine-tuned on three newswire datasets: the CNN/Daily Mail dataset (CNN/DM, Hermann et al. 2015), XSum (Narayan et al., 2018), and the NYT dataset (Sandhaus, 2008). It uses a novel training paradigm that introduces a contrastive learning component to estimate the probability of the generated summaries more accurately.

We chose the pre-trained XSum BRIO model[4], which most closely resembles the style of GUM's single sentence summaries (cf. Figure 1). We test whether the summaries generated by the model are able to capture gold salient entities in GUMsley using the UD test partition (see Table 3). We also include summary level scores on the full dataset in Table 4 to see whether SEE information can enhance summarization quality.

CTRLSum CTRLSum (He et al., 2022) is a summarization model used for generating abstractive summaries. It is considered a controllable summarization method because it produces summaries based on user input, which can specify entities of interest (in the form of keywords), summary length, and questions that the summary should answer. The system is a fine-tuned version of the $\text{BART}_{\text{LARGE}}$ model (Lewis et al., 2020) based on three training datasets: CNN/DM, arXiv scientific papers (Cohan et al., 2018), and BIGPATENT (patent documents, Sharma et al. 2019).

CTRLSum is designed to separate test-time user control of summarization and the training process. During training, summaries are conditioned on the

[4] `Yale-LILY/brio-xsum-cased` on Huggingface

source document and automatically extracted keywords. At test time, a *control function* is applied to map control aspects to keywords, while model parameters from training remain unchanged. Thus CTRLSum differs from other controllable summarization methods in not requiring separate models for each control aspect, generalizing to new keywords at test time.

We use the pre-trained CTRLSum model[5] in three scenarios: GOLD, PRED and ZERO. For GOLD we use the 3 most frequently mentioned gold salient entities in each document[6] as "keywords" (all unique mentions of these entities are used, excluding pronouns); in PRED we generate predicted salient entities using GPT-4 (OpenAI 2023), a generative LLM that achieves human-level performance on a range of benchmarks, using the prompt '*Find the top 3 salient entities in the following document.*', and in ZERO we test without adding salient entities.

GPT-4 GPT-4 (OpenAI, 2023) is the latest version of Generative Pre-trained Transformers at the time of writing. Although training details for GPT-4 are not released (incl. model size, architecture, dataset, training method, etc.), we know from technical reports (OpenAI, 2023) that it was trained using masking and reinforcement learning from human feedback (RLHF).

For a more robust comparison between the fine-tuned models (BRIO and CTRLSum) and prompt-based models, we control the length of GPT-4 prompts using the following prompt: *Summarize the following article in* N *sentences.* Since BRIO's XSum model produces one-sentence outputs and CTRLSum summaries are mostly 2-3 sentences, we can compare GPT-4 summaries with both systems using the sentence-count length prompt *N*. In order to test whether adding gold or predicted entities to the model helps generate better summaries, we use the following prompt format: *Summarize the following article in N sentences. In your summary, make sure to include the following words: <gold or predicted entity 1,2,3>.*

[5] https://github.com/salesforce/ctrl-sum

[6] The choice of using the top 3 salient entities rather than all salient entities is because the minimum count of salient entities (i.e. several documents only have 3), and therefore it represents a reasonable prompt for the GPT model, which would otherwise potentially be asked to generate more salient entities than the document contains, leading to precision errors.

5 Evaluation

In this section, we evaluate model performance and the impact of SEE on two aspects of summarization: Section 5.1 shows a manual evaluation of entity-level performance (are all salient entities included in summaries?) on the test set[7] (24 documents/ 6k entity mentions, Table 3), and Section 5.2 shows the overall summary quality on the entire dataset (213 documents/ 30K entity mentions) based on automatic metrics.

5.1 Entity Level Evaluation

We use GUMsley to test the two systems above, as well as GPT-4 itself, and examine whether baseline results differ from settings where predicted or gold-standard salient entities are provided (for CTRLSum and GPT-4; BRIO does not provide summary control mechanisms). Apart from investigating whether the systems are able to capture entities that appear in the summary (see Table 3), we conducted a quantitative (see Appendix C for additional quantitative analysis of system output factuality using automated scores i.e. $SummaC_{Conv}$ (Laban et al., 2022)) and qualitative analysis on entity hallucination (see Figures 2,3,4), which examine entities that didn't appear in the summary or source document.

Following Nan et al. (2021), we evaluate summaries at the entity level by taking precision, recall and F1 score for unique predicted entities (rather than mentions), e.g. $P_t = N(h \cap t)/N(h)$, is the precision, where $N(h \cap t)$ is the number of distinct gold salient entities also mentioned in the summary and $N(h)$ is the number of entities mentioned in the generated summary. For all the system summaries, we performed a manual evaluation to ensure the quality of mention/entity detection in all 12 genres. That is, the mentions/entities in the generated summaries are identified manually by one of the authors rather than automatically by an entity resolver or coreference system, which are known to perform poorly on out-of-domain genres (Moosavi and Strube, 2017; Zhu et al., 2021).

Overall, we see that both dedicated summarization systems and prompt-based LLMs show poor performance in capturing all salient entities,

[7] The entity level evaluation was performed only on the test set because it needed to be carried out manually and separately for each of the 7 system outputs, making evaluation of the full dataset unfeasible.

	P_t	R_t	$F1_t$	P_t	R_t	$F1_t$	P_t	R_t	$F1_t$	P_t	R_t	$F1_t$
	CTRLGold			**CTRLPred**			**CTRL0**			**BRIO**		
ac	0.615	0.615	0.615	0.536	0.583	0.558	0.857	0.462	0.600	0.571	0.333	0.421
bi	0.813	0.765	0.788	0.607	0.403	0.434	0.600	0.176	0.273	0.600	0.462	0.522
cn	0.471	0.500	0.485	0.583	0.339	0.413	0.500	0.250	0.333	0.330	0.250	0.284
fc	0.583	0.500	0.538	0.750	0.417	0.533	0.333	0.214	0.261	0.166	0.154	0.160
it	0.769	0.526	0.625	0.900	0.436	0.564	0.875	0.368	0.519	0.647	0.478	0.550
nw	0.632	0.500	0.558	0.833	0.350	0.452	0.900	0.375	0.529	0.636	0.292	0.400
rd	0.500	0.643	0.563	0.875	0.354	0.500	0.600	0.214	0.316	0.455	0.385	0.417
sp	0.217	0.481	0.299	0.486	0.217	0.299	0.857	0.222	0.353	0.666	0.296	0.410
tx	0.632	0.750	0.686	0.500	0.198	0.283	0.333	0.125	0.182	0.222	0.133	0.166
vl	0.800	0.615	0.696	0.833	0.244	0.377	0.667	0.154	0.250	0.538	0.292	0.379
vy	0.577	0.455	0.508	0.479	0.292	0.360	0.500	0.094	0.158	0.200	0.066	0.099
wh	0.857	0.514	0.643	0.500	0.153	0.229	1.000	0.057	0.108	0.500	0.152	0.233
Total	0.555	0.555	0.555	0.657	0.332	0.417	0.658	0.206	0.313	0.512	0.255	0.340
	GPTGold			**GPTPred**			**GPT0**					
ac	0.409	0.692	0.514	0.400	0.615	0.485	0.304	0.538	0.389			
bi	0.619	0.765	0.684	0.375	0.529	0.439	0.455	0.588	0.513			
cn	0.524	0.688	0.595	0.435	0.625	0.513	0.333	0.438	0.378			
fc	0.650	0.929	0.765	0.391	0.643	0.486	0.409	0.643	0.500			
it	0.737	0.737	0.737	0.364	0.632	0.462	0.400	0.526	0.455			
nw	0.647	0.458	0.537	0.462	0.500	0.480	0.481	0.542	0.510			
rd	0.650	0.929	0.765	0.579	0.786	0.667	0.647	0.786	0.710			
sp	0.459	0.630	0.531	0.300	0.222	0.255	0.432	0.593	0.500			
tx	0.571	0.750	0.649	0.367	0.688	0.478	0.423	0.688	0.524			
vl	0.450	0.346	0.391	0.550	0.423	0.478	0.421	0.308	0.356			
vy	0.433	0.394	0.413	0.533	0.485	0.508	0.615	0.485	0.542			
wh	0.696	0.457	0.552	0.690	0.571	0.625	0.760	0.543	0.633			
Total	0.570	0.648	**0.594**	0.454	0.560	0.490	0.473	0.556	0.501			

Table 3: Entity level scores and the macro-averaged scores per model on the GUMsley test set for several systems. The blue text is the highest score across 12 genres and red text is the lowest. The top $F1_t$ score across all models is bolded. See Table 1 for genre codes.

with $F1$ scores ranging from 30s to 50s. Table 3 shows that BRIO trained on XSum (Narayan et al., 2018) performs poorly in all 12 genres, but especially in genres rich in conversations (*conversation* and *fiction*).[8] This is expected, as models trained solely on news may not generalize to out-of-domain (OOD) data like *conversation* and *fiction*. Interestingly, we also found that entity hallucination is most severe in these genres, see e.g. P score of 0.166 for *fiction*, mainly due to hallucinations. For example, the BRIO summary in Figure 2 for one of the *fiction* mentions 'the German writer and photographer Barbara Hepworth' even though it has not been appeared in the source document.

We also tested whether providing salient entities to the model would improve performance. Table 3 shows CTRLSum and GPT scores in three settings: adding gold salient entities that have the top 3 frequent mentions in the document (GOLD), adding predicted salient entities from the GPT-4 model (PRED), and no salient entities provided (ZERO). Unsurprisingly, GPT with gold salient entities (GPTGold) outperforms all models, with $F1 = 0.594$. The models in the GOLD setting also outperform those with the other two settings, as can be seen in the $F1$ scores of CTRLSum methods and GPT methods. Interestingly, despite having a lower $F1$ score, CTRL0 has a surprisingly high P score, while CTRLGold has the lowest P score. This is because CTRL0 often picks out the first sentence in the document as the generated summary, which usually contains a large number of salient entities (high precision) but not all of the

[8] Although *fiction* is considered a written genre, the data contains substantial dialogue between characters.

important ones (low recall), as shown in Figure 3. We do not see this pattern in GPT methods, suggesting that the position of entities is not the only factor to take into account for GPT-4 to generate summaries.

For genre comparison, most of the models perform relatively well in written genres (e.g. *academic, biography, interview*[9]) but not in spoken genres (e.g. *speech*). This is reasonable, as spoken genres are considered "unfamiliar" and out-of-domain for models trained on written data. However, this modality (written vs. spoken) effect does not seem to explain the performance of GPT methods, as can be seen in the lower scores of both spoken (*speech*, *vlog*) and written (*academic*) genres. The poor performance of *academic* and *speech* might be explained by the fact that they have a rather low % of salient entities and a rather large number of entities per document (see Table 1), which makes it hard for the model to capture the targeted salient entities in a document. Surprisingly, GPT methods perform well in *fiction* and *reddit*, which are usually hard for other models. This might be because the Pile dataset (Gao et al., 2021), which is known to be used as training data for GPT models, includes diverse data sources including books and web text.

We also saw that adding gold salient entities to CTRLSum methods is especially beneficial for genres like *voyage* and *wikihow*. However, we did not see the power of adding gold salient entities for *voyage* and *wikihow* in GPT methods. Without adding gold entities, CTRL0 and CTRLPred models often produce summaries that are too short and abstractive, leaving out important details. By contrast, GPT summaries in all three settings are rather similar in terms of the mentioned entities. For example, the CTRL0 summary for one of the *wikihow* documents is simply its title: "*How to Grow Beavertail Cactus.*", while the CTRLGold and GPT summaries mention methods and materials that can be used to grow Beavertail Cactus, which human summarization also captured.

Interestingly, we observe that adding gold salient entities to GPT models is specifically useful for highly conversational genres like *conversation* and *interview*, whereas predicted entities added to the model in these genres are not as useful as the gold ones. This suggests that predicting such entities is still a difficult task for GPT-4, especially in spoken genres. A closer look at these predicted entities shows that GPT-4 tends to pick out PERSON entities in the document as salient, which is not always correct. Figure 3 shows the predicted entities (in italics) including several PERSON entities, which were disregarded by humans.

In terms of entity hallucination, we can see from Figures 2,3,4 that BRIO summaries contain the most hallucinated entities, while we hardly see any hallucinations in CTRLSum and GPT-4 summaries. Our quantitative analysis in Appendix C also shows that adding salient entities to the model enhances the faithfulness of the summaries. However, it is worth noting that our analysis focuses on 'intrinsic' hallucinations (Ji et al., 2023), which are those that do not appear in and/or contradict the source document. We did notice other types of hallucinations (i.e. 'extrinsic' ones) in GPT-4 outputs. These include entities neither supported nor contradicted by the source document. For example, in Figure 3, GPTPred outputs 'the speaker's "long" friendship with. . . ,' although the document does not specify whether the friendship is "long" or not. We believe that further work could be done on central propositions or claims in text and their role in curbing this type of hallucination but a rigorous evaluation of this issue lies beyond the scope of the experiments we conducted.

5.2 Summary Level Evaluation

We evaluate the quality of the generated summaries from SOTA models and prompt-based models, including BRIO, CTRLSum and GPT-4, using the widely used ROUGE scores (Lin, 2004) and BERTScore (Zhang et al., 2020). ROUGE-1 and ROUGE-2 are used to measure the unigram and bigram overlap with the reference summary, respectively. ROUGE-L score (longest common subsequence) is used to measure the sentence level structural similarity between the generated and reference summaries. BERTScore measures the semantic similarity between the generated and reference summaries by computing the similarity score for each token in the generated and reference summary.

In general, we found that all models perform the best with the GOLD setting, and the ZERO setting has the lowest performance. This is unsurprising, as adding gold salient entities to the model enhances both lexical/content overlap and semantic

[9]Although *interview* is considered a spoken genre, the source of the data is Wikinews interviews with politicians, which makes the language similar to news articles.

Model	Metrics	ac	bi	cn	fc	it	nw	rd	sp	tx	vl	vy	wh	Avg
BRIO	ROUGE-1	31.49	30.44	14.81	11.77	31.53	30.66	19.54	27.25	14.69	15.56	18.89	15.60	21.85
	ROUGE-2	10.06	12.96	2.27	1.94	13.43	11.79	2.89	11.39	1.57	5.48	3.63	4.59	6.83
	ROUGE-L	23.94	25.11	11.49	8.97	26.59	23.70	17.27	23.42	11.69	13.32	15.51	13.70	17.89
	BERTScore	.65	.65	.55	.51	.66	.63	.56	.61	.56	.56	.58	.54	.59
CTRL0	ROUGE-1	25.11	32.92	7.18	8.66	35.02	29.22	12.50	15.41	15.29	10.13	17.39	16.89	18.81
	ROUGE-2	5.32	17.30	2.72	1.46	18.73	11.08	1.56	8.49	4.49	0.51	5.51	6.25	6.95
	ROUGE-L	20.85	31.98	6.16	7.94	29.38	23.70	11.70	13.66	13.81	7.47	15.54	15.80	16.50
	BERTScore	.59	.65	.45	.49	.64	.63	.51	.54	.52	.50	.56	.53	.55
CTRLPred	ROUGE-1	23.20	38.14	16.61	23.49	34.65	31.58	20.85	21.89	21.31	17.92	23.57	18.76	24.33
	ROUGE-2	8.44	20.31	4.36	4.59	18.98	13.90	5.43	9.34	6.30	7.38	6.58	4.87	9.21
	ROUGE-L	19.27	34.31	13.86	20.35	31.20	25.58	18.72	18.24	18.38	15.33	19.78	16.46	20.96
	BERTScore	.55	.67	.47	.51	.63	.61	.50	.55	.51	.51	.56	.51	.55
CTRLGold	ROUGE-1	29.48	42.44	20.05	19.42	39.76	41.07	23.60	27.51	23.58	24.35	28.26	26.12	28.80
	ROUGE-2	7.83	25.67	5.62	2.66	18.24	18.32	4.60	13.10	7.09	6.26	12.12	6.83	10.69
	ROUGE-L	22.72	39.91	18.14	17.40	31.67	31.06	18.10	23.96	19.32	17.75	23.31	21.86	23.77
	BERTScore	.59	.69	.50	.52	.65	.68	.54	.57	.57	.53	.60	.56	.58
GPT0 N=1	ROUGE-1	29.89	46.20	18.79	29.67	38.04	46.69	28.97	32.26	28.81	25.30	29.27	35.90	32.48
	ROUGE-2	9.72	25.19	4.23	5.37	15.21	20.99	6.05	15.35	9.13	8.99	7.64	10.35	11.52
	ROUGE-L	22.86	37.97	16.91	23.05	28.12	36.07	22.79	25.62	23.09	20.75	22.27	28.68	25.68
	BERTScore	.64	.72	.59	.64	.69	.73	.64	.66	.64	.64	.66	.67	.66
GPTPred N=1	ROUGE-1	29.49	41.35	21.27	30.46	38.56	46.18	31.04	29.78	24.86	40.84	31.83	33.12	33.23
	ROUGE-2	9.83	20.28	4.29	9.07	14.51	20.84	6.25	10.00	5.75	15.96	9.71	8.86	11.28
	ROUGE-L	24.29	33.77	18.30	24.97	28.42	32.81	24.48	22.76	17.94	31.29	24.52	27.58	25.93
	BERTScore	.64	.70	.62	.65	.69	.72	.63	.65	.61	.69	.66	.67	.66
GPTGold N=1	ROUGE-1	33.96	47.63	24.05	30.57	39.00	47.17	30.78	30.55	28.59	31.01	33.10	34.30	**34.23**
	ROUGE-2	10.96	27.00	5.84	5.62	16.04	20.46	7.92	12.96	9.21	11.78	12.47	10.30	**12.55**
	ROUGE-L	24.78	40.68	21.81	23.39	30.54	33.94	23.72	24.15	20.46	25.15	26.86	26.03	**26.79**
	BERTScore	.66	.73	.62	.65	.70	.72	.64	.69	.63	.64	.68	.67	**.67**

Table 4: Summary level scores on GUMsley with the SOTA abstractive summarization method (BRIO), controllable summarization method (CTRLSum), and zero-shot LLM GPT-4 with three different settings: with gold salient entity information (GOLD), with predicted salient entities from GPT-4 (PRED) and without salient entity information (ZERO). N represents the sentence-count length in GPT-4 methods. The blue text is the highest score across 12 genres and red text is the lowest. The highest average scores across all models are bolded.

similarity between the generated summary and the ground truth summary.

As can be seen in Table 4, GPTGold outperforms all the other models on all metrics. Within GPT methods, we found that models prompted to summarize in 1 sentence generally outperform those prompted in 2-3 sentences (compare the numbers in Table 4 with those in Table 5). This is because longer summaries are usually too specific, leading to lower quality summaries. Figure 3 shows a qualitative example of this.

In terms of the differences between models, we observe that CTRLSum summaries are more extractive than BRIO summaries, containing more document entities (predicted or gold) in the output. We also found that BRIO summaries suffer a lot from entity hallucination, which can be alleviated by CTRLSum methods. GPT-4 summaries are considered the closest to the gold summaries for the following reasons: First, they contain as many gold entities as the CTRLSum summaries but with more subjective coherence. Second, they have the least hallucinated entities compared to the other two models. See Figure 3 for a comparison between the ground truth summary and summaries generated by all the models from a *conversation* document in our dataset.

For genre comparison, all the models perform relatively well in written genres such as *biography*, *interview*, and *news* since they are trained on written genre data. Although the training details for GPT-4 are unknown, it is likely that the majority of data it has been trained on is relatively similar to news language. For example, we see that GPT0 and GPTPred methods perform extremely well on *news* data (see the blue text in news genre in Table 4). However, we see that most of the models perform poorly on *conversation* and *fiction* data, which shows that both the SOTA models and prompt-based models are less familiar with spoken genres and are thus incapable of generating high-quality summaries for these "outlier" genres. Specifically, we found that BRIO summaries for *fiction* data suffer from severe entity hallucination problems, which makes the generated summary not factually consistent with the gold one.

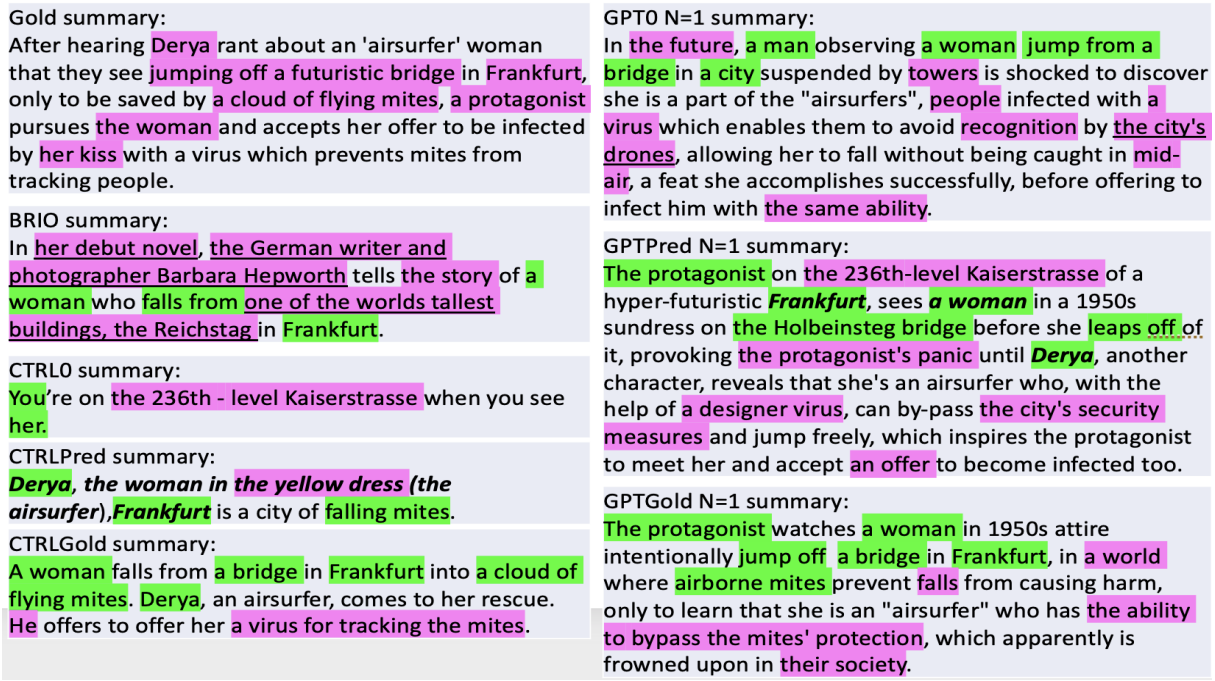

Figure 2: Generated summaries with several models from a *fiction* document in GUMsley. The first mention of the entity has been highlighted in pink. Those that have a match in the ground truth summary are highlighted in green. The hallucinated entities are underlined and predicted entities by GPT-4 are in italics.

GPT summaries, on the other hand, suffer less from entity hallucinations but provide overly specific details of document contents (e.g. *the protagonist's panic*, *the city's security measures* in Figure 2). Similar to GPT summaries, CTRLSum summaries have few hallucinated entities. However, they are usually too short to cover a more exhaustive overview of document content. Figure 2 shows the ground truth summary for one of the *fiction* documents with generated summaries from all the models. The hallucinated entities are underlined and predicted entities from GPT-4 are in italics.

6 Conclusion

This paper presented GUMsley, the first manually annotated entity salience dataset covering all named and non-named entities for 12 genres of English text. GUMsley achieves a high level of agreement in all 12 genres, creating a high-quality entity salience dataset that allows the evaluation of SEE annotations in diverse genres. Our evaluation shows that a significant amount of salient entities are not captured by SOTA abstractive summarization models and prompt-based LLMs and that adding salient entities to model inputs substantially enhances the coverage. We also show that adding such entities helps reduce hallucinations in less common genres (e.g. textbooks and travel guides) to a large extent, generating higher-quality summaries. We hope that GUMsley will enable further research on entity salience and can serve as a challenging dataset for testing text summarization methods in a wide range of genres focusing on entities.

Limitations

This paper has several limitations. First and most important is the restriction of the data to English, the highest resource language in NLP research – it is likely that our findings underestimate the contribution of providing salient entities for summarization in lower resource languages, while also overestimating the performance of pretraind models on the summarization and salient entity prediction tasks for the same languages. It is also possible that SEE annotation would not generalize well, or suffer from more disagreements in other languages, though we believe this is unlikely.

A further limitation in terms of the evaluation of pretrained LLMs is that we cannot rule out that models have seen some of the evaluation data in some form during pretraining. GUM data is part of the Universal Dependencies project, which is managed over GitHub, and is therefore susceptible to inclusion in the Pile dataset, known to be used

as training data for GPT models. If such effects are present in our evaluation, they should minimize, rather than maximize the contribution of providing SEE information. Our data is also relatively small in terms of summarization datasets, meaning that while it may not substantially affect LLM training, more data would lead to better results.

Additionally, we would like to point out that the summary level evaluation in Section 5.2 could benefit from a human evaluation study on the quality of the system and reference summaries. First, it has been pointed out in Goyal et al. (2022) that automatic metrics (e.g. ROUGE, BERTScore), though being commonly used, may not always correlate well with human evaluation. In this respect, we conducted the first manual, qualitative evaluation of the system summaries for the included entities (see Figures 2,3,4). Our analysis shows that adding predicted or gold salient entities to summarization models helps enhance summary quality by alleviating hallucinations in summaries. Despite this, we certainly believe that a more systematic human evaluation of system summaries would be beneficial, and it's worth exploring in future work. Second, previous research (Liu et al., 2022a; Pu et al., 2023) has found that the reference summaries in the existing datasets are not always of good quality, especially when compared with summaries generated by LLMs. The expert-written summaries in GUM (see Liu and Zeldes (2023) for more details), however, are considered high quality because the summaries follow stricter guidelines than other 'found' summarization datasets (Gamon et al., 2013). This is supported by the human evaluation study conducted in Liu and Zeldes (2023), where human evaluators strongly preferred GUM human-written summaries to summaries generated by LLMs such as GPT-3. Also, GUM summaries were found to be best at substituting reading the text, while summaries from LLMs and pre-trained supervised models were considered less substitutive.

Finally, we note that the reference-based summarization paradigm is fundamentally limited in scoring outputs based on gold standard comparisons, despite the fact that alternative summaries may be equally good. We counter this issue by performing manual, qualitative human evaluation in this paper, and argue that while different summaries may include other ancillary entities, ones that are truly salient are likely to appear in almost any valid summary of a document, suggesting that at least the SEE recall of our manual approach should be satisfactory. We feel that this is valuable new data that can contribute especially to existing, automatically constructed datasets using click data (Gamon et al., 2013), NER/coreference resolution tools (Dunietz and Gillick, 2014) or hyperlinks (Wu et al., 2020), which have also covered rather few domains in the past, and no spoken data. We leave the study of precision in SEE with multiple reference summaries for future papers.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomáš Vitvar, and Harald Sack. 2016. Crowdsourced corpus with entity salience annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3307–3311.

Jesse Dunietz and Dan Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013. Identifying salient entities in web pages. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2375–2380.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Janet Yang Liu and Amir Zeldes. 2023. GUMSum: Multi-genre data and evaluation for English abstractive summarization. In *Findings of ACL 2023*, Toronto.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022a. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. 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.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, pages 33–40.

Salvatore Trani, Claudio Lucchese, R. Perego, David E. Losada, Diego Ceccarelli, and Salvatore Orlando. 2018. SEL: A unified algorithm for salient entity linking. *Computational Intelligence*, 34:2 – 29.

Chuan Wu, Evangelos Kanoulas, Maarten de Rijke, and Wei Lu. 2020. Wn-salience: A corpus of news articles with entity salience annotations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2095–2102.

Wen Xiao and Giuseppe Carenini. 2022. Entity-based spancopy for abstractive summarization to improve the factual consistency. *arXiv preprint arXiv:2209.03479*.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. Ontogum: Evaluating contextualized sota coreference resolution on 12 more genres. 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 461–467.

A GPT-4 summary level scores with different length constraints

We found that GPT models prompted to summarize in 1 sentence (N=1) usually outperform those prompted in 2-3 sentences (N=2,3). Compare Table 5 with the GPT results in Table 4. Figure 3 shows qualitative differences between GPT-4 models with different length controls.

B API costs

At the time of running our experiments, GPT-4 API costs \$0.06 / 1K tokens.[10] We generated around 1,278 GPT-4 summaries for all evaluations in Section 5. The total cost of API requests was about \$88.

[10]See https://openai.com/pricing for more details.

C Quantitative evaluation of hallucination in summaries

Apart from the qualitative analysis on entity hallucination (Figures 2,3,4), we conducted a quantitative evaluation on the GUMsley test set to evaluate whether generated summaries are factually consistent with the source article. We used the $SummaC_{Conv}$ factuality metric (model_name = 'vitc', granularity=sentence-level) from the SummaC model (Laban et al., 2022), which is an NLI (Natural Language Inference, or more specifically textual entailment) model that is used to measure hallucination based on the assumption that a faithful summary will be entailed by the gold source document. Table 6 shows the $SummaC_{Conv}$ scores ranging from 0 to 1, with 0 indicating the generated summary logically follows from the source document (entailment) and 1 representing that the generated summary contradicts the information in the source document (contradiction).

Overall, we can see that adding predicted or gold salient entities to the model significantly improves the factuality of generated summaries and reduces hallucinations (lower scores are better), compare e.g. the total scores of the CTRL0 (0.737), CTRLPred (0.658) and CTRLGold models (0.537). Among all models, GPTPred produces summaries with the best entailment score (0.222). This demonstrates that adding predicted entities to GPT-4 contributes the most to improving faithfulness.

For genre comparison, we found that pretrained SOTA abstractive summarization models (BRIO and CTRLSum) have higher factuality scores in written genres (e.g. *fiction*, *textbook*) but not in spoken genres (e.g. *conversation*). This is unsurprising, as most of the summarization models were trained on written data, whereas spoken data like *conversation* are considered out-of-domain for these models. As such, it is easier for the models to generate summaries that are more factually consistent with the source document (or contain fewer hallucinations) in these "familiar" genres. However, similar to our findings in Section 5.1, this genre effect does not seem to appear in GPT methods, where spoken genres like *speech* surprisingly outperform written genres like *fiction* and *academic*. As we have indicated in Section 5.1, this might be because GPT-4 was trained on a wide variety of data sources, which include political speeches, etc.

Model	Metrics	ac	bi	cn	fc	it	nw	rd	sp	tx	vl	vy	wh	Avg
GPT0 N=2,3	ROUGE-1	28.58	43.39	22.51	27.88	34.23	43.61	26.67	31.27	29.34	23.73	29.30	37.66	31.51
	ROUGE-2	8.35	25.17	5.06	5.24	13.88	18.12	4.13	15.45	8.49	9.20	7.93	11.57	11.05
	ROUGE-L	22.85	36.13	20.26	21.89	26.20	33.00	20.59	23.99	22.43	19.91	23.29	29.27	24.98
	BERTScore	.62	.69	.60	.63	.64	.69	.60	.63	.59	.61	.65	.66	.63
GPTPred N=2,3	ROUGE-1	29.61	43.21	22.47	26.73	37.44	44.15	27.92	28.33	23.46	24.36	27.54	34.03	30.77
	ROUGE-2	9.98	21.76	3.75	5.32	15.76	18.17	5.01	12.35	5.09	8.97	7.76	10.37	10.36
	ROUGE-L	22.95	35.41	19.07	20.42	27.74	32.39	22.04	23.16	17.13	18.77	22.27	28.27	24.14
	BERTScore	.61	.68	.59	.61	.65	.68	.60	.63	.57	.62	.64	.65	.63
GPTGold N=2,3	ROUGE-1	32.01	43.29	23.87	28.58	37.72	46.31	29.00	29.67	28.22	26.44	29.92	36.19	32.60
	ROUGE-2	11.25	23.80	4.64	5.04	12.75	19.95	4.24	13.91	7.13	10.38	11.64	11.19	11.33
	ROUGE-L	23.69	34.90	20.79	20.01	27.65	32.39	21.81	24.27	19.53	21.09	24.99	28.38	24.96
	BERTScore	.62	.69	.57	.62	.65	.70	.61	.63	.59	.63	.65	.66	.64

Table 5: Summary level scores on GUMsley with GPT-4 N=2,3. The blue text is the highest score across 12 genres and red text is the lowest.

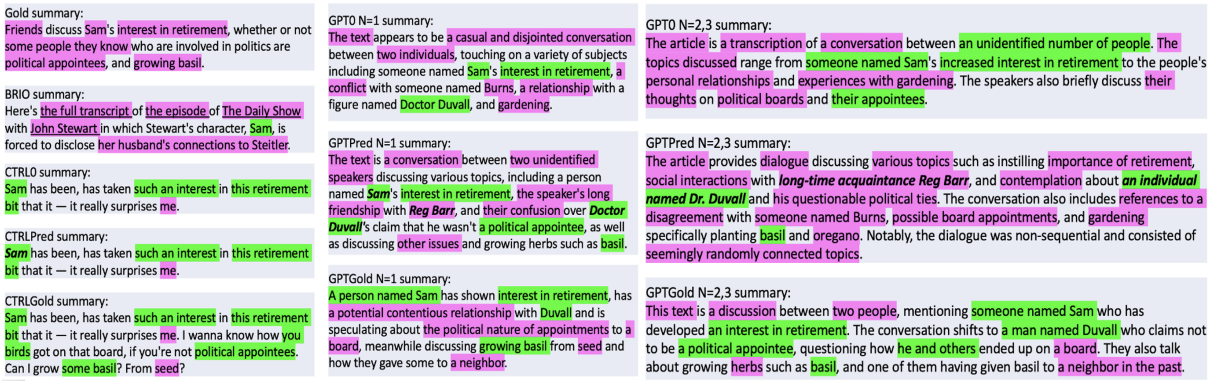

Figure 3: An example of generated summaries with all the models from a *conversation* document in GUMsley. The first mention of the entity has been highlighted in pink. Those that have a match in the ground truth summary are highlighted in green. The hallucinated entities are underlined and predicted entities by GPT-4 are in italics.

Interestingly, we found that $SummaC_{Conv}$ scores in *textbook, voyage, and interview* improve the most after adding predicted or gold salient entities to the model (see the scores with † in Table 6). This indicates that the addition of salient entities is most effective in enhancing faithfulness in these 'unusual' genres.

D Detailed summary examples

Although most of the models perform generally well in written genres, we found that BRIO and GPT models perform unsatisfactorily in the *textbook* genre in summary level evaluation. The possible reasons for this include: (1) The BRIO summaries for *textbook* are too short to include important details of the article and they often contain hallucinated entities. (2) GPTPred summaries contain incorrect salient entities predicted by GPT-4, leading to low ROUGE scores. We found that the headings of textbooks are sometimes misleading. GPT-4 tends to select entities based on their position in the textbook i.e. entities that are in the beginning or the headings of the textbook article are more likely to be selected as "salient", which is not always correct (e.g. *Abraham Lincoln* is mentioned but never discussed in the underlying document, and the human summary omits his name). See Figure 4 for an example of this.

	BRIO	CTRL0	CTRLPred	CTRLGold	GPT0	GPTPred	GPTGold
ac	0.228	0.845	0.805	0.832	0.230	0.235	0.349
bi	0.495	0.708	0.689	0.537	0.245	0.234	0.212
cn	0.223	0.885	0.890	0.439†	0.273	0.241	0.237
fc	0.233	0.300	0.355	0.367	0.228	0.251	0.217
it	0.543	0.862	0.714†	0.625	0.430	0.215†	0.231†
nw	0.425	0.879	0.830	0.850	0.289	0.203	0.205
rd	0.255	0.840	0.851	0.739	0.268	0.242	0.217
sp	0.203	0.735	0.764	0.488	0.218	0.200	0.203
tx	0.235	0.820	0.372†	0.337†	0.338	0.208†	0.206†
vl	0.248	0.439	0.484	0.412	0.245	0.232	0.231
vy	0.371	0.820	0.709	0.412†	0.545	0.201†	0.203†
wh	0.553	0.719	0.435†	0.412	0.241	0.207	0.207
Total	0.334	0.737	0.658	0.537	0.296	**0.222**	0.226

Table 6: $SummaC_{Conv}$ scores on the GUMsley test set for all systems. The blue text shows the highest entailment score across 12 genres and red text is the highest contradiction score across 12 genres. The best entailment score across all models is **bolded**. Scores with † are the ones that have top 3 Δ scores in each model compared to the corresponding ZERO setting. See Table 1 for genre codes.

Gold Summary:
This excerpt from a history book tells about the American Civil War, describing economic hardships for the Confederacy, its failure to obtain support from Great Britain and France, and the role of African American Union soldiers in the war.

BRIO Summary:
The Civil War was the largest conflict in the history of the United States, and it was the conflict that led to the creation of the United States of America.

GPT0 N=1 Summary:
The Union utilized its resources, including the mobilization of African American soldiers and the creation of new railroad infrastructure, to overpower the Confederacy and secure victory in the Civil War, despite initial doubts about President Lincoln's reelection and the potential involvement of foreign powers such as France and Great Britain.

GPTPred N=1 Summary:
The Union overpowered ***the Confederacy*** in the Civil War due to greater resources, industrialization, and manpower, despite initial doubts about ***Abraham Lincoln***'s re-election and the Confederacy's hopes for European intervention; ***African American soldiers*** played a significant role in the Union's victory, with over 190,000 enlisting despite initial discriminatory practices and harsh treatment from Confederates, and their gallantry despite their predicament helped further cement public sentiment against the Confederacy.

GPTGold N=1 Summary:
The Union, fueled by its strong economy, successfully overpowered the Confederacy during the American Civil War due to superior resources and the contributions of African Americans, who enlisted in large numbers and served in various capacities, despite the existence of racism within the ranks.

Figure 4: An example of generated summaries with BRIO and GPT-4 from a *textbook* document in GUMsley. The first mention of the entity has been highlighted in pink. Those that have a match in the ground truth summary are highlighted in green. The hallucinated entities are underlined and predicted entities by GPT-4 are in italics.