Factual or Contextual?
Disentangling Error Types in Entity Description Generation

Navita Goyal
University of Maryland
navita@cs.umd.edu

Ani Nenkova
Adobe Research
nenkova@adobe.com

Hal Daumé III
University of Maryland
Microsoft Research
me@hal3.name

Abstract

In the task of entity description generation, given a context and a specified entity, a model must describe that entity correctly and in a contextually-relevant way. In this task, as well as broader language generation tasks, the generation of a nonfactual description (factual error) versus an incongruous description (contextual error) is fundamentally different, yet often conflated. We develop an evaluation paradigm that enables us to disentangle these two types of errors in naturally occurring textual contexts. We find that factuality and congruity are often at odds, and that models specifically struggle with accurate descriptions of entities that are less familiar to people. This shortcoming of language models raises concerns around the trustworthiness of such models, since factual errors on less well-known entities are exactly those that a human reader will not recognize.[1]

1 Introduction

Gricean maxims of effective communication (Grice, 1975) as they pertain to referring expressions (Dale and Reiter, 1995) posit that referring expressions should not convey false information and that they should be relevant to context. *Factuality* and *congruity* are thus the two main properties of pragmatically appropriate referring expressions.

Following these maxims, human-written referring expressions strongly adhere to the principles of factuality and congruity (Dale and Reiter, 1995; Kheirabadi and Aghagolzadeh, 2012). Standard evaluation practices for referring expression generation (Belz et al., 2009; Kang et al., 2019; Cao and Cheung, 2019), however, only distinguish between model generated referring expressions being *accurate* (ground-truth) versus *inaccurate* (not ground-truth), without considering factuality and congruity of the model outputs. However, this distinction between factual and contextual errors is important, as contextually-relevant factual errors are likely to be harder for people to identify; this concern is supported by evidence that human annotators trust translations that are fluent/coherent but inadequate over translations that are adequate but disfluent (Martindale and Carpuat, 2018; Popović, 2020).

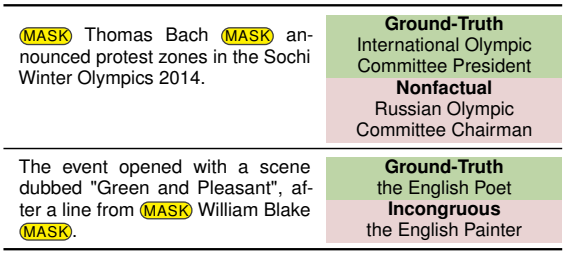

Figure 1: Two example contexts with MASKs representing potential location of a target referring expression. In the top case, the inaccurate generation is nonfactual (not true) but contextually plausible (given "Sochi"); in the bottom, the inaccurate generation is factual (Blake was both a poet and a painter) but incongruous.

In this work, we design an evaluation framework to study the distribution of factual and contextual errors in referring expression selection and generation, for descriptions of people mentioned in English news articles. In the top example in Figure 1, the model generates the description *Russian Olympic Committee Chairman*, plausibly based on the contextual cues (*Sochi*), leading to factual error that is contextually relevant. We call such contextual but not factual descriptions **nonfactual**. In the bottom example in Figure 1, the generated description *the English Painter* is factually correct (Wiliam Blake was both a poet and painter), however it is contextually not relevant. We call such factual but not contextual descriptions **incongruous**.

To tease apart these two failure modes, we automatically construct potentially nonfactual and incongruous reference texts using article context and other factual sources and evaluate generated descriptions against these reference texts (§ 2.1).

[1]The code and data used in the paper is available at https://github.com/navitagoyal/Factual-or-Contextual-Errors-in-LM-Desc-Gen.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8322–8340
July 9-14, 2023

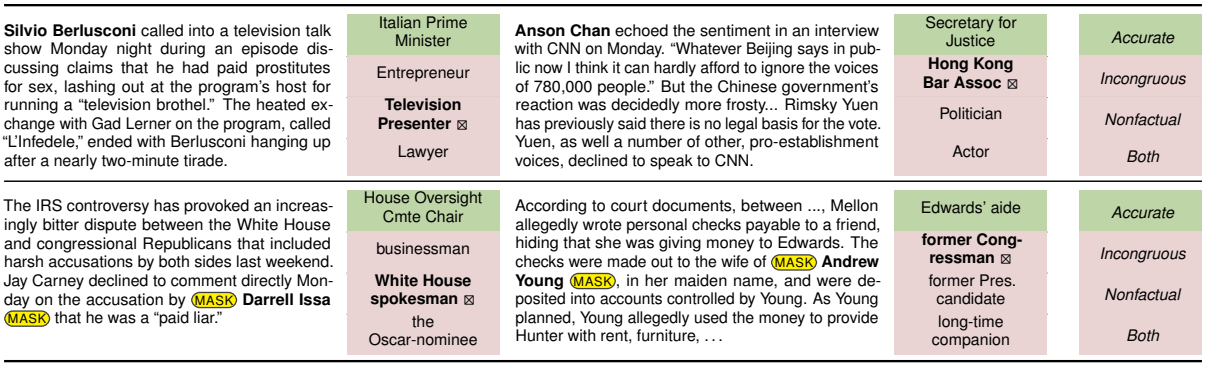

Figure 2: Example for claim identification (top) and description identification (bottom) tasks with context and four distractors. The bold distractor (⊠) indicates the model's predicted class: nonfactual (left) and incongruous (right).

Additionally, we consider a multiple-choice identification experiment, with distractors that are nonfactual, incongruous, or both (§ 2.2, § 2.3). Our goal is to evaluate the ability of language models to select between factually and contextually plausible alternatives and to determine which aspects are more problematic for language models. We find that models commonly make both factual and contextual errors.

Because people reading the generated descriptions might fail to recognize factual errors for entities they are unfamiliar with, we augment our evaluation to distinguish between more and less familiar entities. We find that models disproportionately predict nonfactual descriptions for unfamiliar entities, raising concerns about the trustworthiness of text where description selection is guided by language models (§5). We validate our nonfactuality and incongruity assumptions on the proposed distractors (§6) and show that our findings continue to hold for a completely validated test set. Finally, we discuss the validity of our proposed evaluations in measuring what we purport them to measure from the measurement modeling perspective (§7).

2 Referring Expression Generation

Open-book referring expression generation involves two steps: claim selection and description generation (Kang et al., 2019). Claim selection identifies facts from the target entity's WikiData (referred as *claims*) relevant to the context. Description generation produces the text, conditioned on the context and relevant claims, for instance using an auto-regressive encoder (Kang et al., 2019). Traditionally, such generation tasks are evaluated using text similarity metrics, such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002), of the predicted description with respect to the ground-truth description. However, the failure modes beyond the ground-truth are usually left unexplored.

Here we develop a framework for evaluating generated descriptions against various reference texts, controlled to be accurate, nonfactual, or incongruous, to sharpen our understanding of model errors. We also design multiple choice evaluations for claim and description identification, with distractors that are nonfactual, incongruous, or both. This multiple-choice setup allows more control over the chosen alternatives to highlight the factual and contextual preferences in language models. For instance, in the top-left example in Figure 2, the claim identification model identifies Silvio Berlusconi as a television presenter rather than as the Italian prime minister, possibly because the context text mentions "television talk show."

2.1 Description Generation

The description generation task calls for generating referring expressions describing people in a text. For training a description generation model, we mask the description text in an article and feed the masked context t, along with the target entity e, to a language model as input to generate the description left-to-right. To investigate factual and contextual errors, we evaluate the similarity of generated description (y^{gen}) not only with the ground-truth ($y^{\star}$) but also with alternative factual, or contextual, reference text that are *incongruous* (y^{inc}), or *nonfactual* (y^{nf}), respectively.

- ▷ y^{inc}: we take that entity's entire Wikipedia text as an incongruous reference; this presumably contains true facts about that entity, but which are not relevant in the article context.
- ▷ y^{nf}: we take the descriptions of all *other* entities mentioned in the source article (excluding target

entity's descriptions) as a nonfactual reference; these are very unlikely to be facts about the candidate entity, but are definitionally contextually relevant.

In both cases, we exclude ground-truth descriptions from incongruous and nonfactual reference texts.

Given these reference texts, we measure precision of the generated description (y^{gen}) with the $y^{\text{ref}} \in \{y^\star, y^{\text{inc}}, y^{\text{nf}}\}$ (after excluding stop words) using BLEU (unigram) (Papineni et al., 2002), ROUGE-L (Lin, 2004) and Bertscore (Zhang et al., 2019). We use precision, instead of recall, as the generated description and ground-truth tend to be short (1.43 ± 0.89 words and 1.64 ± 0.85, respectively), whereas the distractor reference texts are much longer. The recall of distractor reference is not possible, or even expected. The distractor reference texts, instead, act as noisy proxy of the space of incongruous and nonfactual tokens that the model might generate. We would expect a "good" system to have high precision with the ground-truth and low precision with either distractor. A high precision with respect to distractor reference texts indicates model tendency to generate factually incorrect or contextually incongruous descriptions.

2.2 Claim Identification

The claim identification task is to find a relevant Wiki property (*claim*) for an entity (e) mentioned in the article given the masked article context (t). We frame claim identification as a multiple-choice problem where the ground-truth ($y^\star$) is taken from the original text and we automatically construct *incongruous* (y^{inc}), *nonfactual* (y^{nf}), and *both incongruous and nonfactual* (y^{both}) distractors (see Figure 2 *(top)*). For a given text t and a target entity e, we construct alternatives $\mathcal{A}_{\text{CI}} = \{y^\star, y^{\text{inc}}, y^{\text{nf}}, y^{\text{both}}\}$ as:

- ▷ $y^\star$: claims in the target entity's Wiki tagged as relevant to the ground-truth description. If multiple, we choose the claim with the highest unigram overlap with the description.
- ▷ y^{inc}: claims in the target entity's Wiki tagged as irrelevant to the ground-truth description.
- ▷ y^{nf}: claims pertaining to the *other* entities in the article, drawn from the article directly.
- ▷ y^{both}: claims pertaining to entities not mentioned in the article, drawn from a random other article.

For $y^\star$ and y^{inc}, we use the contextual relevance mark up from Kang et al. (2019). To minimize the chance that a distractor is accidentally accurate, we sample all distractors from respective candidate sets ensuring zero unigram overlap with the ground-truth description. We ascertain nonfactuality of y^{nf} and y^{both} distractors by enforcing zero overlap with all claim associated with the target entity.

2.3 Description Identification

Description identification aims to identify the ground-truth description ($y^\star$) from distractors. In contrast with claim identification, which uses structured Wiki claims as alternatives, description identification uses free-text descriptions in the article. We construct alternatives $\mathcal{A}_{\text{DI}} = \{y^\star, y^{\text{inc}}, y^{\text{nf}}, y^{\text{both}}\}$ as:

- ▷ y^{inc}: descriptions of e from another article.
- ▷ y^{nf}: description of another entity in the article.
- ▷ y^{both}: description of an entity not mentioned in the article.

We ensure that distractor descriptions ($y^{\text{inc}}, y^{\text{nf}}, y^{\text{both}}$) have zero unigram overlap with $y^\star$. Also, y^{nf} and y^{both} do not have any overlap with any description of the target entity.

For both claim and description identification, the model takes (t, e, y) as input and outputs a probability $p(y)$, for each $y \in \mathcal{A} = \{y^\star, y^{\text{inc}}, y^{\text{nf}}, y^{\text{both}}\}$. We report the relative frequency that each type of distractor is highest ranked: $\textsc{Top} = \arg\max_{y \in \mathcal{A}} p(y)$. We also compute a mean reciprocal rank (MRR) for the ground-truth and distractor classes on the held-out test data. MRR averages the reciprocals of the rank of each class: large MRR values correspond to higher ranking alternatives.

For uniform comparison of the generation experiment with the identification experiments, in addition to reporting the BLEU, ROUGE-L, and Bertscore of the generated description with respect to the ground-truth and distractor reference texts, we also rank the different classes (*accurate*, *incongruous*, and *nonfactual*) based on their BLEU, ROUGE-L, and Bertscores on each example and report the TOP and MRR of each class. Note, we do not have a *both incongruous and nonfactual* class in the description generation task as the generation task is free-form, making the scope of description that is *both incongruous and nonfactual* too open-ended to be expressed by some reference text.

3 Entity Familiarity

Our proposed evaluation attempts to disentangle factual and contextual preferences in language mod-

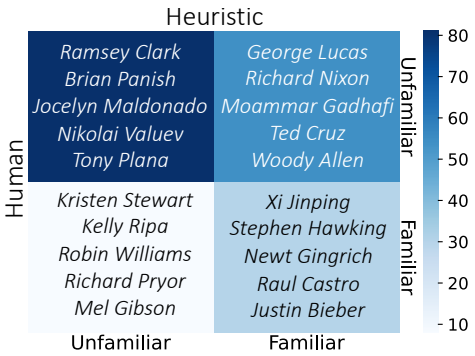

Figure 3: Familiarity: heuristic vs human annotation.

		Train	Valid	Test
Claim Ident.	Familiar	2050	27	66
	Unfamiliar	3608	112	93
Desc. Ident.	Familiar	10305	57	59
	Unfamiliar	5888	46	22
Desc. Gen.	Familiar	13373	353	365
	Unfamiliar	19721	726	631

Table 1: Number of examples with *familiar* and *unfamiliar* entities across tasks.

els in generating referring expression for entities mentions in news articles. However, these preferences need not be uniform across entities. Language models may possess disparate information about different entities, leading to disparate reliance on factual and contextual cues. We aim to understand how factual and contextual errors relate to the familiarity with the entity being described.

We develop a heuristic for approximating entity familiarity, adapted from Siddharthan et al. (2011). We tag entities that always appear in article summaries without descriptions (referring expressions) as *familiar* and the rest as *unfamiliar*.

4 Experiment Details

We conduct our experiments on the PoMo dataset (Kang et al., 2019), which contains English news articles from CNN, Daily Mail and New York Times, along with their summaries. The dataset also identifies post-modifier description for an entity mentioned in the article and a set of *claims* from target entity's Wikidata. We extract all *person* entities mentioned in the article using NER tagging (Finkel et al., 2005). We use claims associated with the target entity and other entities mentioned in the article to construct alternatives for claim identification.

For description identification and generation experiments, we extract referring expressions—pre-modifier, relative clause, appositive, participle clause, adjective/adverb clause and prepositional phrase—using the regular expressions of Staliūnaitė et al. (2018). We use the same set of referring expressions for familiarity heuristics. We use a dependency parser (Manning et al., 2014) to extract the descriptions associated with entities' first mention.[2] We identify alternative descriptions of an entity in other articles based on full name matches. Within an article (or summary), we match entities based on first or last name (Siddharthan et al., 2011).

We validate the familiarity heuristic by collecting human annotations on a subset of 200 person entities in the validation set, with 3-4 annotations per entity. We ask crowdworkers on AMT (paid ~US$15/hour on AMT) if they are familiar with the given entity. For the entities marked as familiar, we ask annotators to add a short description of the entity to ensure that annotators perform the task carefully. We specifically ask the annotators to not use external search for this task and clarify that there is no penalty for tagging entities as familiar or unfamiliar. We use two attention check questions: one corresponding a well-known entity (e.g., Joe Biden) and another corresponding to a made up entity, expecting the response as familiar and unfamiliar, respectively. We further discard the annotations where 90% of entities are marked as familiar or unfamiliar.

Inter-rater agreement according to the generalized Kappa measure for multiple raters (Gwet, 2014) is 0.768, indicating "substantial agreement" (Viera et al., 2005). We consider entities tagged unfamiliar by the majority annotators as *unfamiliar* and the rest as *familiar*, and measure precision/recall of our heuristic with respect to this ground-truth. The proposed heuristic has precision of 79%, recall of 64% and F-score of 67%. This is close to precision, recall, and F-score of baseline human annotations comparing one-vs-rest (74%, 75%, and 74%, respectively). See Figure 3 for examples of (dis)agreement between human annotations and our heuristics.

Table 1 shows the distribution of examples with familiar and unfamiliar entities across different tasks. We apply our evaluation procedure on mod-

[2]First mentions of entities are generally longer and descriptive, and serve to introduce relevant information about the entity. Later references tend to be mostly referential (Siddharthan et al., 2011) (see Appendix A for details).

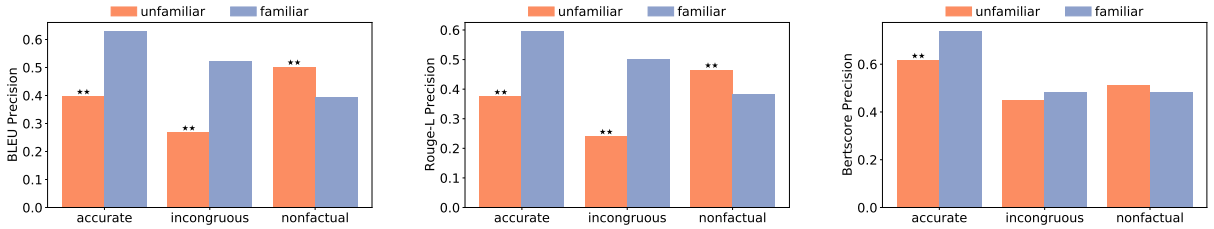

Figure 4: Description Generation: BLEU, ROUGE-L, and Bertscore Precision across familiar and unfamiliar entities for T5-base fine-tuned model. The model exhibit higher overlap with nonfactual reference texts for unfamiliar entities and incongruous reference texts for familiar entities.[3]

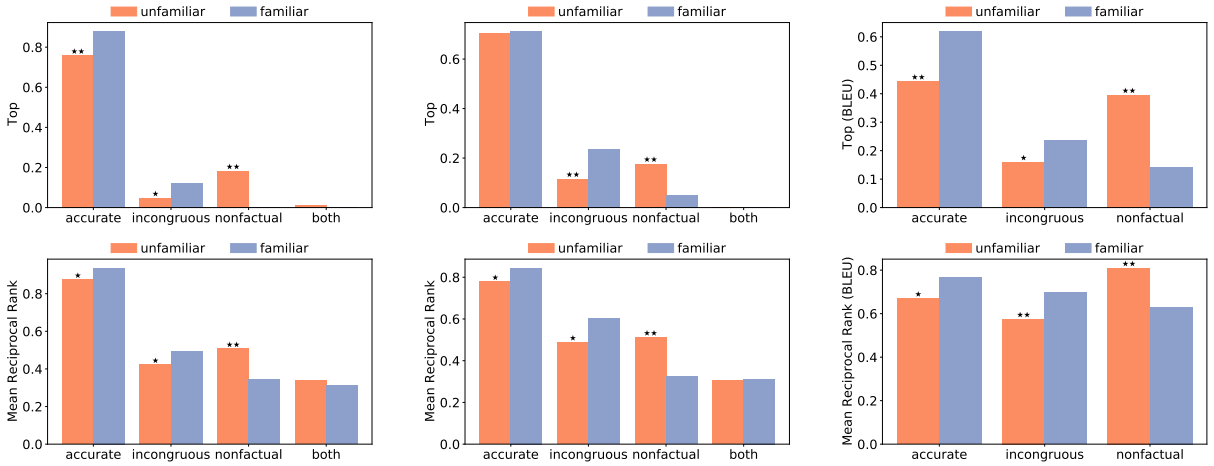

Figure 5: Claim Identification (left), Description Identification (middle), Description Generation (right): Relative Frequency of Highest Rank and Mean Reciprocal Rank for each class across unfamiliar and familiar entities. Models predominately favor nonfactual alternatives for unfamiliar entities and incongruous alternatives for familiar entities.[3]

els that we fine-tune on each task. We consider 3 sequence-to-sequence models for the generation task: T5-small, -base (Raffel et al., 2020) and BART models (Lewis et al., 2020) with $60M$, $220M$, and $140M$ parameters, respectively. We consider 3 multiple-choice models for the identification tasks: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020) base models with $110M$, $130M$, and $110M$ parameters, respectively. See Appendix D for details.

5 Do error types differ for familiar vs unfamiliar entities?

Figure 4 shows the generative evaluation with respect to ground-truth and distractor reference texts across familiarity for descriptions generated by the fine-tuned T5-base model. We observe a higher overlap with nonfactual reference texts for unfamiliar entities and a higher overlap with incongruous reference texts for familiar entities across all evaluation metrics. The difference is significant for BLEU and ROUGE metrics with p<0.005, but non-significant for Bertscore. TOP and MRR of classes (accurate, incongruous, or nonfactual) in description generation task, obtained by ranking the BLEU overlap between the generated description and reference texts, yield similar results (Figure 5). We observe a higher TOP and MRR for incongruous reference texts for familiar entities and a higher TOP and MRR for nonfactual reference texts for unfamiliar entities. We include results for other models and metrics in Appendix E.

Controlled evaluation of RoBERTa model fine-tuned on Claim and Description Identification tasks (Figure 5) echo the same trends. Nonfactual distractors are ranked higher than incongruous distractors for unfamiliar entities, indicating more factual errors. In contrast, incongruous distractors are ranked higher for familiar entities, indicating more contextual errors. Additionally, comparing between familiar and unfamiliar examples, we find that rate of incongruous errors is significantly higher (p<$\{0.005, 0.05\}$) for familiar examples than that for unfamiliar examples. The converse is true for

[3] $\star$ and $\star\star$ represents significance level of 0.05 and 0.005, respectively, between the unfamiliar and familiar sets.

factual errors, with a significantly higher (p<0.005) rate of factual errors for unfamiliar examples than familiar examples. This reflects that the distribution of different types of models errors is different for different entity types, with a disproportionately higher rate of nonfactual predictions for unfamiliar entities. Unsurprisingly, we also observe a higher performance (accurate class) for familiar entities in most of the cases. Our findings are consistent across models (Appendix E).

6 Are nonfactual and incongruous alternatives really so?

Our results highlight model tendency to make factual or contextual errors insofar as the automatically extracted distractors and reference texts are faithful to their associated classes. Our automatic distractor extraction makes two assumptions: nonfactuality and incongruity of distractors. For nonfactual distractors, we assume that a claim or description associated with other people mentioned in the article is contextual, but not factual, for the given entity. This assumption is easy to verify: we consider the claims and descriptions associated with the target entity in the PoMo corpus ensuring no overlap with the facts associated with the target entity.

For the incongruous distractors, on the other hand, we have assumed that a random alternative factual claim/description associated with the target entity that is not present in the current context is incongruous. However, it is entirely plausible that alternative factual descriptions are actually sometimes congruous, representing a threat to the validity of this measurement. To ascertain how reasonable this assumption is, we conduct a human study. Because assessing the contextuality of a description is difficult in isolation, we ask annotators to compare automatically extracted incongruous (but factual) descriptions with the congruous (ground-truth) descriptions. Annotators are shown the article context and the two descriptions and asked to give their preference on a 5-point scale ranging from *strongly prefer description 1* to *strongly prefer description 2*. We randomize description 1 and description 2 as incongruous or ground-truth descriptions. We collect 3 annotations for 50 samples from both claim and description identification tasks, compensating AMT crowdworkers at US$15/hour.

For the description generation task, we construct nonfactual reference text as the context excluding factual description of the entity, to enforce the non-factuality assumption. To assess the validity of incongruous reference texts in the description generation task, we collect human annotations. We consider the generated descriptions that overlap with the target entity's Wikipedia, aka, the incongruous reference text as incongruous description. We ask annotators to compare this incongruous description with the ground-truth congruous description, as we do not have a generated congruous description for the same input. We ask the annotators to rate the contextual appropriateness of the two on a scale of 1–5 for 50 examples (3 annotations each). More details on human annotations and interface are included in Appendix F.

	Claim Ident.	Desc. Ident.	Desc. Gen.
Rating congruous	3.74 ± 0.24	3.68 ± 0.25	3.71 ± 0.21
Rating incongruous	2.26 ± 0.24	2.31 ± 0.25	2.29 ± 0.21
Effect size	0.87^{**}	0.78^{*}	0.96^{**}
Inter-rater agg.	0.61	0.54	0.73
Agreement	0.82	0.76	0.79

Table 2: Human verification of incongruity of extracted incongruous distractors for identification tasks and generated incongruous description for the generation task. Annotators rate incongruous descriptions as less contextual than the ground-truth (congruous) descriptions, as expected. The effect size (Cohen's d) is significant ($p < \{0.05^{*}, 0.005^{**}\}$) for all tasks. We observe fair agreement with our automatically extracted incongruous class.

We observe a moderate inter-rater agreement of 0.63 (Fleiss' kappa). The annotator rating for congruous and incongruous description is 3.71 ± 0.23 and 2.28 ± 0.23 (mean$\pm$ standard error), with statistically significant effect size of 0.87 (Cohen's d) at p<0.05. The inter-rater agreements and effect sizes for different tasks are shown in Table 2. We take the description with annotator rating <3 as incongruous. We observe a majority agreement of 0.79 with our automated annotations.

For the claim and description identification task, we construct a gold test set with the subset of examples where our extracted distractors agree with the human annotations. We obtain 41 and 38 gold examples in the claim and description identification, respectively. We also internally validate the factuality assumption for the gold set. We manually check that the incongruous distractors in the gold test set are factual and the nonfactual distractors are not. The evaluation on the gold set is limited to identification tasks. For description generation, human annotators verify only the generated description that overlap with incongruous reference text, not

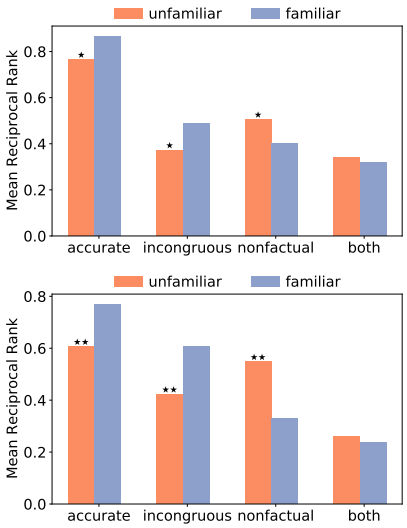

Figure 6: MRR on the gold test set for claim identification (top) and description identification (bottom). The gold set evaluation supports our previous findings: models favor nonfactual alternatives for unfamiliar and incongruous alternatives for familiar entities.[3]

the full reference text, which would require more involved and time-consuming annotations.

Figure 6 shows similar trends on familiar and unfamiliar entities in the human annotated set (gold test set) as seen previously on automatically annotated test sets (Figure 5). Models make more factual errors for unfamiliar entities and more contextual errors for familiar entities (p<0.05), both for claim and description identification tasks.

7 Measurement Validity

Above, we analyzed factual and contextual preferences in language models on three tasks pertaining to referring expression generation. Throughout, we have shown that the language models we measure tend to make both factual errors and congruity errors, and also have shown that these measures are—in various ways—actually measuring what we purport them to measure. Formally, we can conceptualize this from a measurement modeling perspective (Messick, 1995; Jacobs and Wallach, 2021), wherein we consider factual and contextual errors to be *unobservable constructs*.

In this view, we have proposed three measurement devices as proxies for the unobservable construct. This forms a measurement model, which we can example from the perspective of *measurement validity*, considering: *face validity* (the extent to which the measures look plausible), *content validity* (the extent to which the measures capture the substantive nature of the construct), *convergent validity* (the extent to which the measure matches related measures), *concurrent validity* (the extent to which the measures distinguish between groups that it should meaningfully be able to distinguish between), and *consequential validity* (the implication of using the measure).

Face validity is inherently subjective in nature. We pose that the incongruous (but factual) and nonfactual (but contextual) alternatives respectively encode the factual and contextual cues that the models refer to. Without any external criteria, readers are often the only judge of the face validity.

Content validity has two key sub-aspects: substantive validity and structural validity.

▷ *Substantive validity* asserts that the measure, fully and only, incorporates the properties related to the construct. We argue for substantive validity of the measures based on our design and human validation (§6). By design, nonfactual alternatives are extracted from the context and do not have any overlap with the factual information about the entity captured in Wikipedia or WikiData. So, to the extent that Wikipedia/WikiData are correct, these have strong substantive validity. The incongruous alternatives are extracted from factual information about the entity and verified to be incongruous using human validation. They subsequently point to factual preferences in the models. This confirms that the measures capture only the properties related to the construct for both identification and generation tasks.

In the case of the description generation task, however, the nonfactual and incongruous measures do not fully capture the respective constructs of contextual and factual preferences in language models. The unigram precision metric only accounts for factual and contextual indicators in the generated description dictated by the reference texts. Being open-ended, the reference text for incongruous and nonfactual generations is really broad. We design alternative reference texts to cover the range of factual and contextual cues that the model might parrot in generating entity descriptions, but the reference text is not necessarily exhaustive. This points to weaker substantive validity of the measures in the description generation task.

▷ *Structural validity* is a component of content validity that asserts that the measure captures the structure of relationship between the constructs.

Dimension	Description Generation	Claim Identification	Description Identification
Task	✓ Natural	✗ Semi-artificial	✗ Semi-artificial
Evaluation	✗ Difficult	✓ Easy	✓ Easy
Text	✓ Free-text	✗ Structured	✓ Free-text
Face validity	✓ Strong	✓ Strong	✓ Strong
Substantive validity	✗ Weak	✓ Strong	✓ Strong
Structural validity	✓ Strong	✗ Weak	✗ Weak
Convergent validity	✓ Agree	✓ Agree	✓ Agree
Concurrent validity	✓ Strong	✓ Strong	✓ Strong

Table 3: ✓ Pros & ✗ Cons of different tasks. Task, Evaluation, and Text consider the design differences for description generation and claim & description identification tasks. The remaining consider the different dimensions of construct validity for the incongruous and nonfactual measures in each task. The measures have *Strong* face validity and concurrent validity across all tasks, but *Weak* substantive and structure validity for some tasks (§7). For convergent validity, we note that measures *Agree* with each other across tasks.

We note that the identification task measures have relatively weak structural validity: the probability assigned to each alternative is relative and a higher probability to incongruous alternative might stem from being paired with a low probability nonfactual alternative. On the other hand, the description generation task measures have a strong structural validity as the generation task is free-form, so the model directly outputs the highest probability tokens. This represents a trade-off in content validity of the measures: identification has stronger substantive and generation has stronger structural validity.

Convergent validity considers the degree to which multiple measures of the same unobserved construct point in the same direction. Our three proposed measures have a clear convergent validity. As seen in the results (Figure 4-6), the measures point to the same effects across tasks: models default to contextual cues, leading to factual errors, for unfamiliar entities, whereas models tend to be more factual, even while compromising on the congruity of the description, for familiar entities. This distinctive trend across familiar and unfamiliar entities also points to the concurrent validity of the proposed measures.

Concurrent validity asserts that the measure should be able to distinguish between the meaningful groups. Previous works highlight that language models acquire information about entities seen during pre-training (Petroni et al., 2019; Kandpal et al., 2022). For such entities, we can expect the models to reference the factual information seen during pre-training. For unseen entities, on the other hand, we can expect the model to yield best guess based on contextual plausibility. Unsurprisingly, the models' exposure to certain entities would correlate with the human familiarity with these entities due to proliferation of online content on respective entities being a common cause for the two. The proposed measures (nonfactual and incongruous alternatives) are thus able to distinguish models' contextual and factual preferences between potentially seen and unseen groups, confirming the concurrent validity of the measures. Table 3 summarizes the validity of our proposed measures across different tasks.

Consequential validity is an assessment of what happens if a measure is adopted. The unanimous trend across various tasks points to a deeper concern for the variability in the priors that models use for familiar vs unfamiliar entities. Standard evaluations of tasks miss both what kind of errors models make and how these errors differ for different populations. This work brings attention to these problems that usually get hidden under the rug of accuracy measures. Based on our findings, we recommend that downstream tasks requiring a balance of factual and contextual information should probe into the model error distributions and their variances across familiar and unfamiliar entities. Human evaluation for such tasks should also account for human biases and limitations.

8 Related Work

Errors in Text Generation. Language models often generate erroneous information, not supported by source and/or background documents (Ji et al., 2023), often termed as hallucinations. Previous works in text generation—abstractive summarization (Maynez et al., 2020; Pagnoni et al., 2021), dialog generation (Dziri et al., 2021; Santhanam et al., 2021), and translation (Lee et al., 2019)—highlight these nonfactual or incongruous generations, without disentangling the factual and

contextual errors. Cao et al. (2022) study erroneous generations in summarization, which are factual but unverifiable from the source text. Although their work examines contextual errors, they do not contrast these with factual errors due to the context.

Familiarity Prediction. Previous works have studied familiarity of the reader with a person mentioned in news, using linguistic signals in articles and summaries (Siddharthan et al., 2011; Staliūnaitė et al., 2018). Siddharthan et al. (2011) distinguishes between familiar ("hearer-old") and unfamiliar ("hearer-new") based on how people are referred to in summaries— hearer-old entities are referred to with name only or title + name, while hearer-new are referred to with an additional description. Staliūnaitė et al. (2018) further studies the change in description length over time as entities evolve from hearer-old to hearer-new.

Knowledge Probing. There have been many discussions previously around "knowledge" encompassed in language models. Petroni et al. (2019) created the LAMA benchmark with (subject, relation, object) fact triplets, along with human-written templates to elicit these facts. Language models are designed to inherently focus on the context for generation. LAMA benchmark (Petroni et al., 2019) is specifically designed for factual probing. As a result, the context, i.e. template or prompt, is directly linked to the fact in question. More generally, language models are required to be both correct and contextually-relevant. Petroni et al. (2020) improves factual recall by augmenting templates with relevant contextual information retrieved from external sources, such as Wikipedia. The task is designed such that context aids factual probing. Our work deals with naturally occurring contexts.

Factuality and Congruity. Our paper focuses on the two main properties of referring expressions: factuality and congruity. These properties are borrowed from Gricean maxims of effective communication (Grice, 1975): quantity (give as much information as needed, and no more), quality (not to give information that is false), relation (stay pertinent to the discussion) and manner (be clear/brief). We adapt the maxims of quality and relation into factuality and congruity. We do not focus on the maxims of quantity and manner, because these maxims mainly apply to the use and frequency of referring expressions. For example, familiar entities are described without referring expressions in summaries for conciseness (Siddharthan et al., 2011).

In a different directions, research in event factuality deals with the interaction between lexical and syntactic information meant to convey varying levels of veracity or *factuality* of events mentioned in text (Nairn et al., 2006; de Marneffe et al., 2012). For instance, "XYZ, the supposed best artist" implies that the veracity of XYZ being the best artist is questionable. In contrast to this line of research, our work focuses on factuality in terms of the maxim-of-quality sense alone. We conduct a preliminary sanity check against lexicosyntactic triggers for event factuality (White et al., 2018) to confirm that referring expressions considered in our work do not contain any. Extending to broader pragmatic conditions related to event pragmatics would be fascinating future work.

9 Conclusion

In this work, we integrate indicators for factual inconsistencies and contextual incongruities in automated evaluation in referring expression generation. Our paper aims at conducting post-hoc analyses of language models in referring generation task to assess the differences in error types across familiar vs unfamiliar entities that are not reflected in the aggregate accuracy numbers alone. Comparisons with alternative factual and contextual reference text, in addition to those with ground-truth description, suggest that description generation models heavily mis-generate incongruous and nonfactual description. We also show this effect using controlled multiple-choice experiments for claim and description identification. Further, we show that language models disproportionately rely on context when describing less familiar people, resulting in factually incorrect descriptions.

The tasks discussed in our work are meant to exemplify the disparities in language model errors to advocate further research investment in expanding language model evaluation beyond accuracy. Our work opens avenues for future research on the effects of augmenting language models with retrieved information on the factuality and congruity of the generations: do retrieval-augmented language models still make factual errors and are these models able to maintain the contextuality of the generated descriptions. Further, more work is needed to study to what extent language models abide by the conversational pragmatics in the wild.

Limitations

Our work aims to uncover how the distribution of factual and contextual errors in referring expression generation varies based on the familiarity of entities. Our proposed experiments uncover this effect using pragmatics-driven heuristics. We need a more general deep-dive into what models "know" to estimate how language models handle known and unknown information differently, in a way that might even escape human scrutiny. Further, our discussion is limited to specific semi-artificial tasks. Although our work reveals an hitherto understudied shortcoming in language models, extending this diagnosis to general tasks might be non-trivial.

Ethics Statement

In this work, we adapt a pragmatic-driven familiarity heuristic to study factual and contextual errors for familiar and unfamiliar entities. Our work reveals that description generation disproportionately rely on context for unfamiliar entities leading to incorrect predictions. Our heuristic for tagging familiarity is aimed at contrasting model errors. We do not propose the use of this heuristic to filter *unfamiliar* entities in end services.

Acknowledgements

We sincerely thank the reviewers for their helpful comments and insights. We are very grateful to Ruiyi Zhang and the members of the CLIP lab at UMD for their constructive feedback and useful pointers for this work.

References

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. Generating referring expressions in context: The grec task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. 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 3163–3172, Hong Kong, China. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, 38(2):301–333.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, USA. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

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 Comput. Surv.*, 55(12).

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Jun Seok Kang, Robert Logan, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian. 2019. PoMo: Generating entity-specific post-modifiers in context. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 826–838, Minneapolis, Minnesota. Association for Computational Linguistics.

Reza Kheirabadi and Ferdows Aghagolzadeh. 2012. Grice's cooperative maxims as linguistic criteria for news selectivity. *Theory and Practice in Language Studies*, 2(3):547.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. In *International Conference on Learning Representations*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Samuel Messick. 1995. Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, 14(4):5–8.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim,

Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Ieva Staliūnaitė, Hannah Rohde, Bonnie Webber, and Annie Louis. 2018. Getting to "hearer-old": Charting referring expressions across time. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4359, Brussels, Belgium. Association for Computational Linguistics.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Why First Mention Description?

The choice of considering entity description associated with entities' first mention is inspired by previous studies in referring expressions (Siddharthan et al., 2011). Siddharthan et al. (2011) find that the first mentions of entities are generally longer and descriptive, and serve to introduce relevant information about the entity. Later references tend to be mostly referential. We confirm this property in our data: we find that only 14% instances of later mentions of entities, if any, have descriptions. We ignore these later descriptions to avoid any effect of description style varying across first vs later descriptions.

B Data

The PoMo dataset is built on top of CNN and Daily Mail data (See et al., 2017), available under MIT License with conditions only requiring preservation of copyright and license notices, and Wikidata[4] available under the Creative Commons Attribution/Share-Alike License. No additional copyright is listed for the PoMo dataset (Kang et al., 2019).

C Evaluation Metrics

For claim and description identification tasks, we use highest ranking prediction TOP and Mean Reciprocal Rank (MRR) evaluation metrics. TOP considers the highest ranking prediction as 1 and the rest of the predictions as 0:

$$\text{TOP}(c) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[c = \arg\max_{y^c \in \mathcal{A}}\{ p_i(y^c)\}] \tag{1}$$

where $c \in \{\star, \text{inc}, \text{nf}, \text{both}\}$ is accurate, nonfactual, incongruous, or both class, $p_i(y^c)$ is the prediction probability of the class c in the i^{th} sample, and $\mathbb{I}$ is the indicator function. Mean Reciprocal Rank, on the other hand, takes into account the rank order of different classes in each test example:

$$\text{MRR}(c) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\text{rank}_c} \tag{2}$$

where rank_c is the rank of alternative from class c in the i^{th} test sample.

For description generation, we consider three generative metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and Bertscore (Zhang et al., 2019) precision. For each example, we calculate the generative metrics for the generated description with respect to the accurate reference text (ground-truth description) and the incongruous and nonfactual reference texts as described in §2.1. We can consider the set of reference texts as $\mathcal{A} = \{r^\star, r^{\text{inc}}, r^{\text{nf}}\}$. Let $s_i(y, r^c)$ be generative score of the generated description y with respect to the reference text r^c for the i^{th} sample, where $c \in \{\star, \text{inc}, \text{nf}\}$. We further rank the three classes—accurate, incongruous, nonfactual—based on their generative scores on each example and calculate TOP and MRR as:

$$\text{TOP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[c = \arg\max_{r^c \in \mathcal{A}}\{ s_i(y, r^c)\}] \tag{3}$$

and

$$\text{MRR}(c) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\text{rank}_c}, \tag{4}$$

where rank_c is the rank of the reference text from class c in the i^{th} test sample.

D Experiment Details

We fine-tune BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) base models with $110M$, $130M$, and $110M$ parameters ($A = 12$, $H = 768$ and $L = 12$), respectively, for claim and description identification in multiple-choice setting to predict the accurate alternative. For description

[4] https://www.wikidata.org/wiki/Wikidata:Database_download

		Precision			TOP			MRR		
	Error class	BLEU	ROUGE (L)	BERT score	BLEU	ROUGE (L)	BERT score	BLEU	ROUGE (L)	BERT score
T5-base	Accurate	**0.483**	**0.456**	**0.662**	**0.508**	**0.517**	**0.658**	**0.706**	**0.714**	**0.795**
	Incongruous	0.362	0.337	0.462	0.189	0.185	0.091	0.62	0.635	0.463
	Nonfactual	0.463	0.436	0.503	0.303	0.298	0.251	0.745	0.752	0.576
T5-small	Accurate	0.35	0.324	0.584	0.349	0.357	0.508	0.609	0.616	0.699
	Incongruous	0.398	0.367	0.46	0.203	0.213	0.105	0.626	0.645	0.477
	Nonfactual	0.58	0.543	0.523	0.447	0.429	0.388	0.83	0.839	0.657
BART	Accurate	0.058	0.064	0.362	0.006	0.009	0.024	0.356	0.362	0.366
	Incongruous	0.374	0.27	0.533	0.148	0.195	0.163	0.561	0.586	0.563
	Nonfactual	0.588	0.431	0.623	0.846	0.796	0.813	0.937	0.929	0.904

Table 4: Description Generation: BLEU, ROUGE-L, and Bertscore precision on the test set. Precision represents the actual generation metrics; TOP and MRR represent the top and mean reciprocal rank of classes obtained by ranking the generative metrics (BLEU, ROUGE, Bertscore) between the generated description and reference texts from respective classes. Numbers in bold represent the best performing model (highest score on the accurate class).

	Error class	Claim Identification		Description Identification	
		TOP	MRR	TOP	MRR
BERT	Accurate	0.783	0.883	0.667	0.809
	Incongruous	0.089	0.462	0.247	0.576
	Nonfactual	0.115	0.404	0.049	0.373
	Both	0.013	0.335	0.037	0.325
RoBERTa	Accurate	0.809	0.901	**0.711**	**0.839**
	Incongruous	0.102	0.457	0.224	0.579
	Nonfactual	0.083	0.397	0.066	0.355
	Both	0.006	0.329	0.0	0.31
Electra	Accurate	**0.841**	**0.904**	0.368	0.599
	Incongruous	0.076	0.434	0.276	0.538
	Nonfactual	0.076	0.419	0.224	0.511
	Both	0.006	0.326	0.132	0.435

Table 5: Claim and Description Identification Evaluation. TOP and MRR on the test set. Numbers in bold represent the best performing model (highest score on the accurate class).

generation, we fine-tune a T5-small and T-5 base model (Raffel et al., 2020) with $60M$ parameters ($A = 8$, $H = 512$, $L = 12$) and $220M$ parameters ($A = 12$, $H = 768$, $L = 12$), respectively, and a BART base model with $140M$ parameters ($A = 12$, $H = 768$, $L = 12$). We use a learning rate of 2×10^{-5} with huggingface[5] implementation of Adam optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.01. The models are trained for 3 epochs and take about 2-3 hours each to train.

E Results

Table 4-5 show the evaluation of description generation and claim and description identification tasks across different models. We observe an accuracy of 78%, 81%, and 84% on claim identification task and accuracy of 67%, 71%, and 36% on description identification task for BERT, RoBERTa, and ELECTRA models, respectively. Presumably ELECTRA model performs distinctly worse on description identification due to lack of hyper-parameter tuning, which we keep same across all models for ease of experimentation. We find that models rank incongruous or nonfactual distractors at top in between $4 - 27\%$ test examples. We also note that the accuracy in description identification task is on average lower than the accuracy on the claim identification task. The lower accuracy in description identification task reflects the difficulty of identifying appropriate free-text descriptions, as opposed to structured claims.

For description generation, we observe best performance for T5-base model with a BLEU, ROUGE-L and Bertscore of 0.46, 0.52, and 0.71 with respect to the ground-truth description. Ranking the ground-

[5] https://huggingface.co/

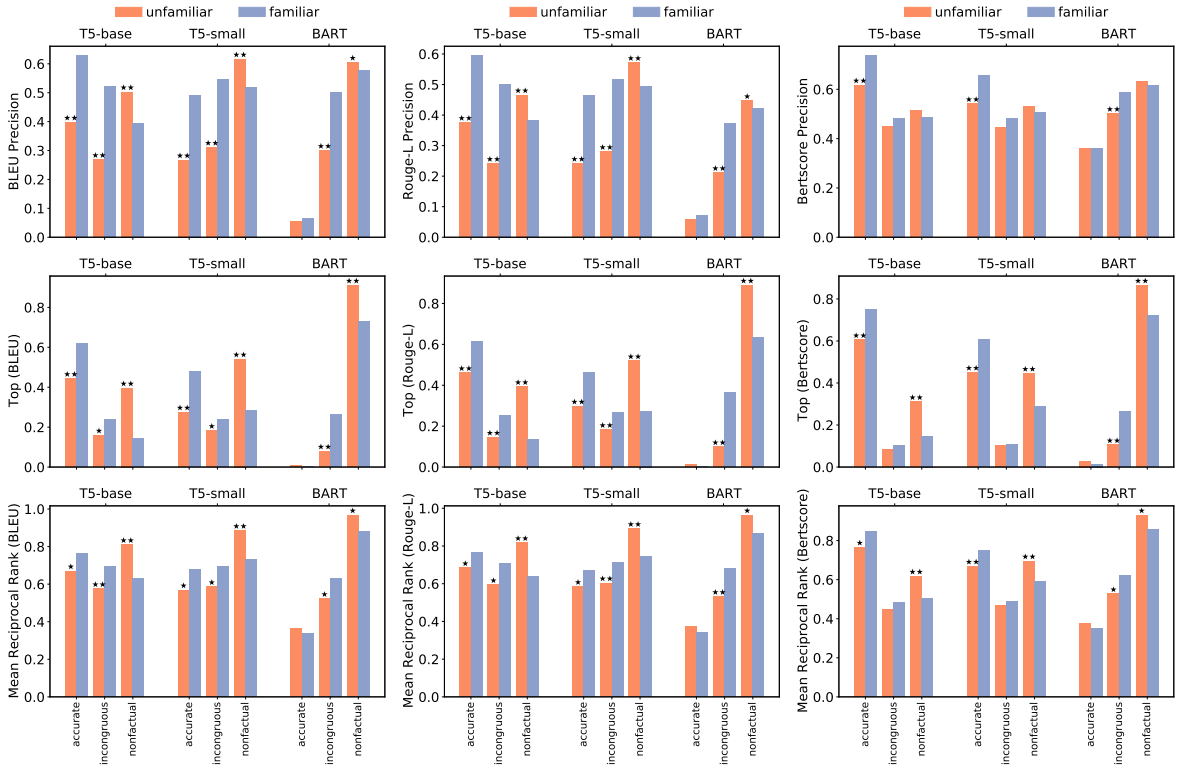

Figure 7: Description Generation: BLEU, ROUGE-L, and Bertscore Precision across familiar and unfamiliar entities. Models exhibit higher overlap with nonfactual reference texts for unfamiliar entities and incongruous reference texts for familiar entities. * and ** represents significance level of 0.05 and 0.005, respectively, between the unfamiliar and familiar sets.

truth and distractor classes based on the generative metrics (BLEU, ROUGE-L, Bertscore), we observe an accuracy of 56% on average for the T5-base model (that is, precision is highest with respect to the ground-truth description in 56% of test examples), with 15% and 28% incongruous and nonfactual errors on average, respectively.

Figure 7-8 show the comparison between factual and contextual errors across familiar vs unfamiliar entities for different models. The results are consistent with §5. Even though, the difference for Bertscore is non-significant, we observe significant and consistent difference between familiar and unfamiliar sets based on TOP and MRR of classes ranked by Bertscore.

F Human Annotations

We conduct two human annotation studies: verifying familiarity heuristics and task verification. For familiarity annotations, we consider 200 person entities in our validation set and have $3-4$ annotations per entity where user mark the entities as being familiar or unfamiliar. For the entities marked as familiar, we ask users to add a short description of who the person is to ensure that users are engaging with the task. We specifically ask the users to not use external search for this task and clarify that there is no penalty for tagging entities as familiar or unfamiliar (9). We set a small time window for the task, about 20 second per entity) to dissuade users from taking aid of any resources. We internally discard the annotations where 90% of entities are marked as familiar or unfamiliar. We further use two attention check questions: one corresponding a well-known entity (e.g., Joe Biden) and another corresponding to a made up entity, expecting the response as familiar and unfamiliar respectively. We collect annotations from English speaking participants in North America region as the news source (CNN) employed in our study is U.S. centric. We do not collect any other demographic information from the participants.

The second set of human annotations aim to verify the congruity assumption of incongruous distractor and reference text in claim & description identification and description generation tasks respectively. We

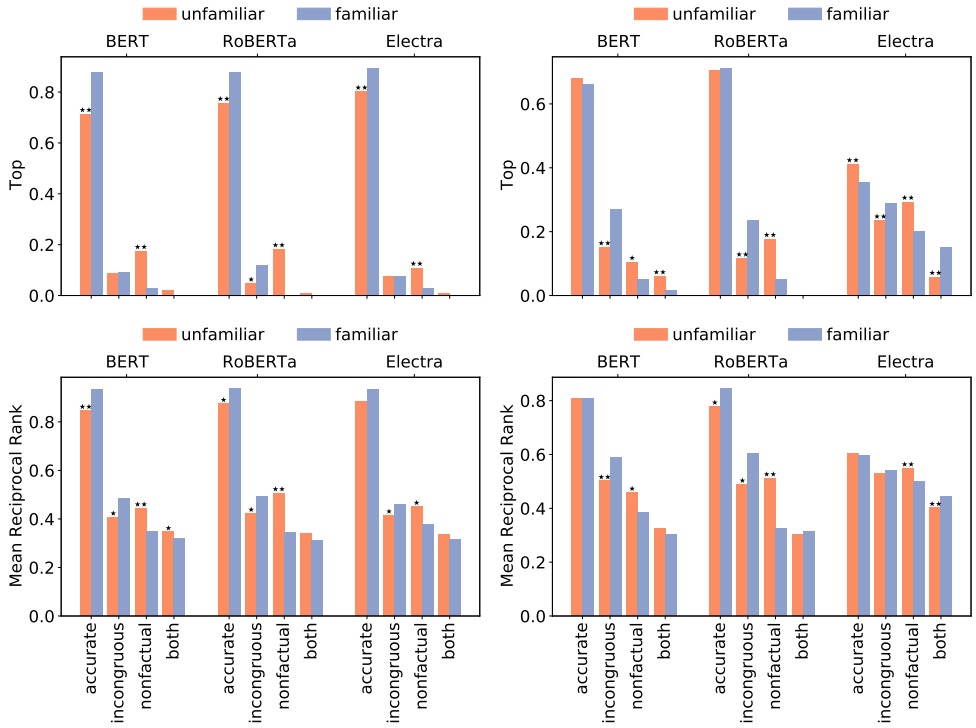

Figure 8: Claim Identification (left), Description Identification (right): Highest Rank and Mean Reciprocal Rank across familiar and unfamiliar entities. Models predominately favor nonfactual alternatives for unfamiliar entities and incongruous alternatives for familiar entities. * and ** represents significance level of 0.05 and 0.005, respectively, between the unfamiliar and familiar sets.

sample incongruous and congruous descriptions for 50 test examples in each task (details of incongruous and congruous description in §6). We ask 3 annotators to compare the two descriptions and rate them on a scale of $1-5$ from *Strongly Prefer description 1* to *Strongly prefer description 2* (Figure 10-11). We ascertain attention by two repeat questions where we asking them to restate their preference for an entity in previous example without context. We exclude responses that fail the attention checks, however, all the annotators are compensated at $15/hour.

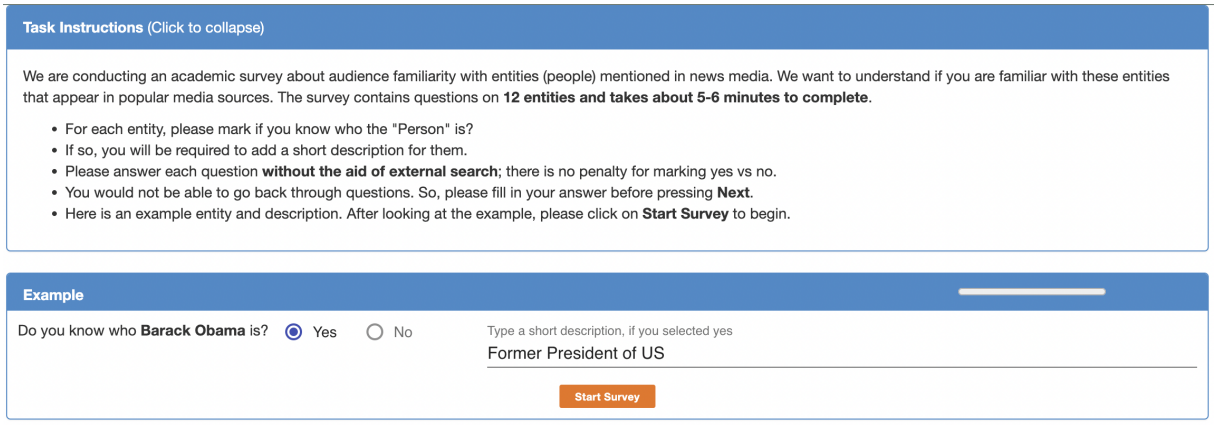

Figure 9: Instructions and example task for human annotations verifying familiarity heuristics.

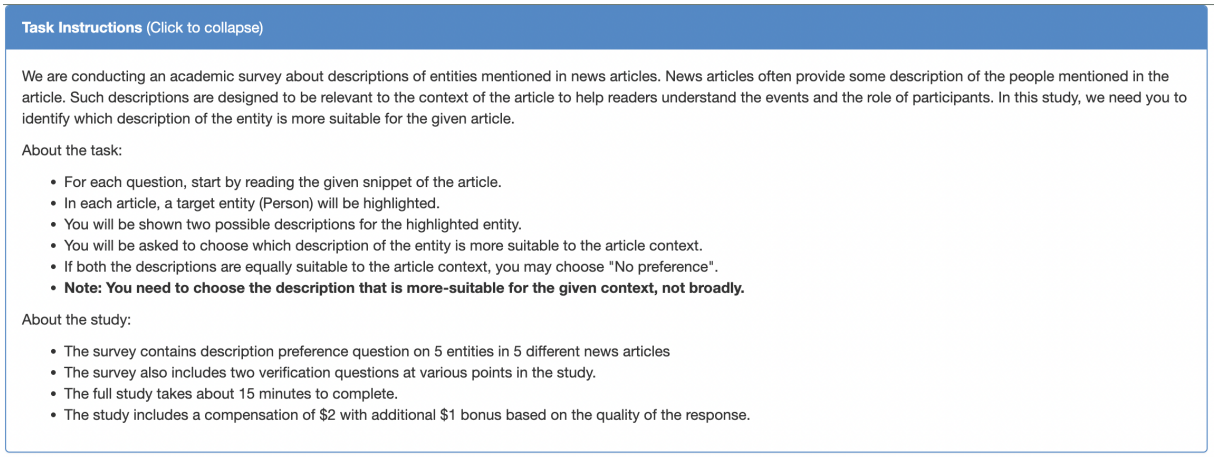

Figure 10: Instruction to participants for human annotations verifying the congruity assumption

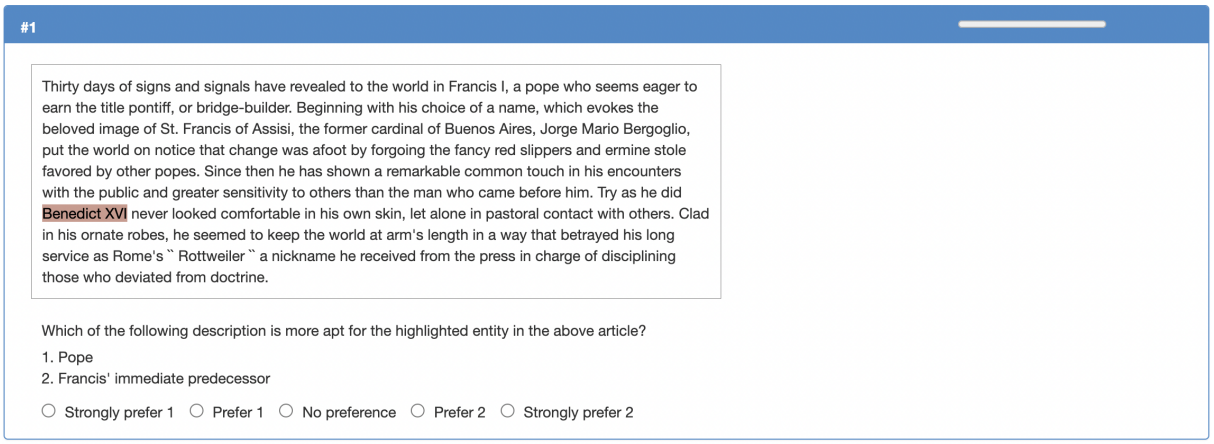

Figure 11: Example task for human annotations verifying the congruity assumption

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☑ A2. Did you discuss any potential risks of your work?
Limitations Section and Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction Section (Section 1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The artifacts used in the paper were available under creative common license and our use does not violate the license conditions.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data is public access news articles and wiki data and is fully non-anonymized. We expect the dataset to not contain any harmful or offensive content as is ensured by the publication standards of CNN, DailyMail and content moderation of Wikidata org.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and Appendix D

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We do not run hyper parameter tuning and use default same hyper parameters across all models and experiments

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 6 and Appendix F

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix F

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4, 6 and Appendix F

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We include instructions to participants regarding the purpose of the data annotation in Appendix F. However, we do not use the collected data for training our models. We only collect annotations to validate assumptions of our automatic data creation and evaluation.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The task and domain used in our setting does not include any potentially harmful content. The text in our article is sourced from open license news articles from CNN and Daily Mail. We only show participants content from the news articles, wherever applicable. We do not require participants to look at any external sources to complete the annotations.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix F