NewsEdits 2.0: Learning the Intentions Behind Updating News

Anonymous ACL submission

Abstract

As events progress, news articles often update with new information: if we are not cautious, we risk propagating outdated facts. In this work, we hypothesize that linguistic features indicate factual fluidity, and that we can *predict which facts in a news article will update* using solely the text of a news article (i.e. not external resources like search engines). We test this hypothesis, first, by isolating fact-updates in large news revisions corpora (Spangher et al., 2022). News articles may update for many reasons (e.g. factual, stylistic, narrative). We introduce the *NewsEdits 2.0* taxonomy, an edit-intentions schema that separates fact updates from stylistic and narrative updates in news writing. We annotate over 9,200 pairs of sentence revisions and train high-scoring ensemble models to apply this schema. Then, taking a large dataset of silver-labeled pairs, we show that we can predict when facts will update in older article drafts with high precision. Finally, to demonstrate the usefulness of these findings, we construct a language model question asking (LLM-QA) abstention task. Inspired by Kasai et al. (2022), we wish the LLM to abstain from answering questions when information is likely to become outdated. Using our predictions, we show, LLM absention reaches *near oracle levels of accuracy*.

1 Introduction

News is the “first rough draft of history” (Croly, 1943). Its information is both valuable and fluid, prone to changes, updates, and corrections. As shown in Figure 1, the first sentence on the left has a factual update, while the second does not. Intuitively, we might be able to predict this: an “*advisory*” is not likely to indefinitely stay in effect, while details about the “*quake*” are less likely to change. Indeed, if someone asks “*Q: Is an advisory still in place?*”, we might want to abstain from answering definitively. However, “*Q: How large was the quake?*” can be answered directly.

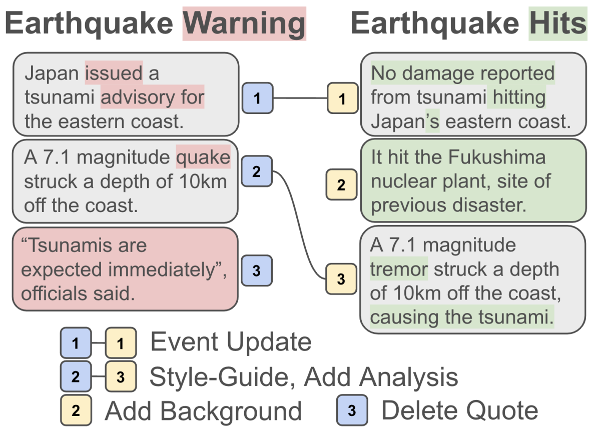

Figure 1: Updates can occur for many different reasons. Shown here, we identify factual updates (e.g. “Event Update” between 1-1), stylistic updates (e.g. “Style-Guide” between 2-3) and narrative updates (e.g. “Add Background” for sentence addition 2).

Recent work has recognized the importance of testing LLM-QA in dynamic settings (Jia et al., 2018; Liska et al., 2022). Kasai et al. (2022)’s RealTimeQA benchmark specifically measures LLM-QA performance for updating news documents. However, current approaches rely on search engines retrieving updated information[1]. This neglects potentially salient linguistic and commonsense information. As the example shown in Figure 1 demonstrates, cues exist that we, as humans, intuitively understand to signal fluidity. *Can we learn these cues, and predict which facts in a news article will update? Can this help LLMs better abstain from answering questions they may not have updated information for?*

We answer these questions in three steps, shown in Figure 2. In **Part 1**, we start by studying update patterns in *NewsEdits*, a large corpus of article revision histories (Spangher et al., 2022). Articles update for many different reasons (e.g. factual,

[1]The latest entry of RealTimeQA was *RAG + Google Custom Search*. https://realtimeqa.github.io/.

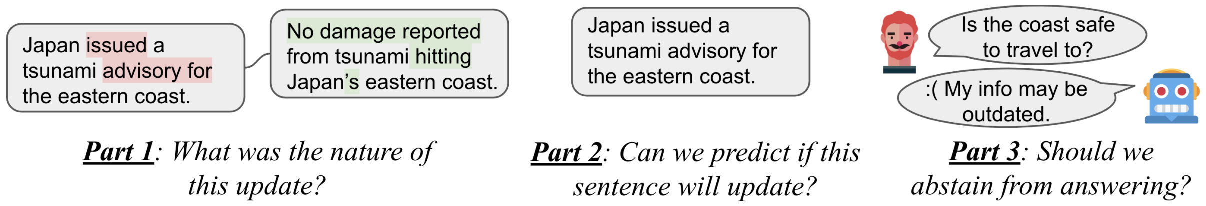

Figure 2: Overall paper flow. In **Part 1** of our paper, we develop an edits-intention scheme to describe news edits and train models to apply this schema to existing news revision corpora (Spangher et al., 2022). In **Part 2**, we use these models to silver-label a large corpus and ask how well we can *predict* whether a sentence will factually update. In **Part 3**, we show these predictions can be beneficial for increasing abstention rates during LLM-QA.

stylistic, etc.), and it is difficult to identify these reasons. So, we introduce *NewsEdits 2.0*, a taxonomy of edit-intentions for journalistic edits (Figure 3), to help us do this. We hire professional journalists to annotate 9,200 pairs of sentence revisions across 507 article revision pairs with the *NewsEdits 2.0* schema. We then train an ensemble model to tag pairs of revisions, with 75.1 Micro F1 and create a large silver-label corpus of revision pairs.

Next, in **Part 2**, we use this silver-labeled corpus to predict which facts in articles might update. We find that models achieve a moderate macro-F1 of .58, overall, on a gold-labeled test set. *Although these scores are noisy, we notice that our models are learning reasonable linguistic cues*. We observe key linguistic patterns: the use of future-tense verbs, statistics and commonly updating events. We validate these cues with human measurement. Further, by focusing on the sentences our models predict are *highly likely* to update, we notice a much higher precision of .74. Finally, in **Part 3**, we simulate a RealTimeQA-style case where an LLM using Retrieval Augmented Generation (RAG) retrieves an outdated document. Without our predictions, the LLM abstains wrongly more than it should. With them, the LLM achieves near-oracle level performance. In sum, our contributions are:

- We introduce the *NewsEdits 2.0* schema, with 4 coarse and 20 fine-grained categories, developed with professional journalists; train models to label these with 75.1 micro-F1; and release a large corpus of 4 million revision histories silver-labeled with edit intentions.
- We show that pretrained LLMs perform poorly at *predicting which facts in the old versions articles will update*, indicating that this important capability is not emergent during pretraining. While fine-tuning helps performance, LLMs still lag humans.
- Finally, we show via a use-case, Question Answering with Outdated Documents, that a failure to address these shortcomings can result in decreased performance for leading LLMs.

Finally, two subtle yet significant contributions of this work are (1) preprocessing improvements we introduce to improve the *NewsEdits* corpus (e.g. improving sentence boundary detection); and (2) visualization tools to make revision histories more accessible to users. Because these advances are not relevant to the main ideas of our paper, we save a deeper discussion these for Appendix A.2. Taken together, we hope that our work can increase utilization and understanding of news dynamics.

2 Related Work

Although most LLM Q&A benchmarks assume that information is static, recent work has increasingly explored LLM performance in the presence of dynamic, updating information (Jia et al., 2018; Liska et al., 2022). This growing direction is concisely captured by Kasai et al. (2022)'s statement: "*GPT-3 tends to return outdated answers when retrieved documents [are outdated]. Can [we] identify such unanswerable cases?*"

To our knowledge, the use of revision-histories to address this question, which we discuss in Section 5, is novel. News updates are an especially crucial domain to study: (1) news is socially important (Cohen et al., 2011); (2) LLMs are increasingly using news to better serve users (Hadero and Bauder, 2023); (3) news is more likely to deal with updating events than other domains (Spangher et al., 2022). Indeed, Kasai et al. (2022)'s RealTimeQA benchmark is built entirely on news data.

Edit-intention schemas have been developed for other types of revision histories, like Wikipedia (Yang et al., 2017), and Student Learner Essays (Zhang and Litman, 2015). In these works, researchers categorize the intention of each edit us-

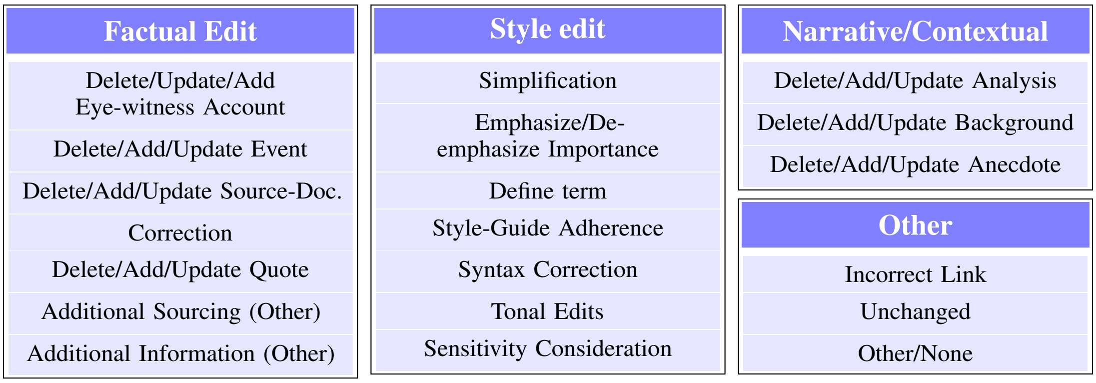

Figure 3: *NewsEdits 2.0*: Edit-Intentions Schema categories and their subcategories. In this work, we focus mainly on the *Factual Edit* category. See Appendix C.1 for definitions for all categories.

ing similar schemas to what we have developed. While building *NewsEdits 2.0*, we were inspired by the schemas developed by prior work and they provided a starting point for our taxonomy. We added edit-categories that were more journalism specific, like "Add Eye-witness Account", and removed categories that were more specific to the aforementioned domains (Section 3.1,). The use-cases of these schemas has mainly focused on stylistic prediction tasks (e.g. text simplification (Woodsend and Lapata, 2011) and grammatical error correction (Faruqui et al., 2018)) or tasks specific to these corpora (e.g. building models to assess the validity of a student's draft (Zhang and Litman, 2015), or counter vandalism on Wikipedia (Yang et al., 2017)). We are the first, to our knowledge, to develop tasks centered on news articles (Section 4) and to apply predictive analyses to fact-based edits.

3 Part 1: Learning Edit Intentions in Revision Histories

News articles update for different reasons, especially during breaking news cycles where facts and events update quickly (Saltzis, 2012). In this section, we introduce the edit-intentions schema we use for *NewsEdits 2.0*, our annotation, and our models to label edit-pairs. This lays groundwork for Section 4, where we will predict when facts change.

We wish to identify categories of edits, in order to enable different investigations into these different update patterns. In other words, we describe the following update model:

$$p(l|s_i, s_j^{'}, D, D^{'}) \tag{1}$$

where l is an *intention* (e.g. a "Correction" needs to be made), D and $D^{'}$ represent the older and newer versions of a news article, respectively, and s_i and $s_j^{'}$ are individual sentences where the update occurred. i, j are sentence indices, ranging from $i \in \{1, ...n\}$, $j \in \{1, ...m\}$ (where n, m are the number of sentences in D, $D^{'}$).

3.1 Edit Intentions Schema

We work with two professional journalists and one copy editor[2] to develop an intentions schema. Building off work by Zhang and Litman (2015) and Yang et al. (2017), we start by examining 50 revision-pairs sampled from *NewsEdits*. We developed our schema through 4 rounds of conferencing: tagging examples finding edge-cases and discussing whether to add or collapse schema categories. Figure 3 shows our schema, which we organize into coarse and fine-grained labels. We incorporate existing theories of news semantics into our schema. For instance, "Event Updates" incorporates definitions of "events" (Doddington et al., 2004), while "Add Background" incorporates theories of news discourse (Van Dijk, 1998). "Add Quote" incorporates definitions from informational source detection (Spangher et al., 2023) and "Add Anecdote" incorporates definitions from editorial analysis (Al-Khatib et al., 2016). See Appendix B.2 for a deeper discussion of the theoretical schemas that inform the NewsEdits 2.0 schema. Finally, "Incorrect Link" is an attempt to correct sentence pairs that were erroneously (un)linked in *NewsEdits*.

3.2 Schema Annotation

We build an interface for annotators to provide intention labels for news article sentence pairs (see Appendix C.2). Annotators are shown definitions for each fine-grained intention and the articles to

[2]Collectively, these collaborators have over 50 years of experience in major newsrooms.

	All		Fact		Style		Narrative	
Features	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
Baseline, *fine-grained*	45.8	73.6	32.0	47.2	58.6	39.9	52.0	39.9
+ NLI	48.6	74.1	45.7	50.4	55.2	38.7	43.6	38.7
+ Event	46.7	74.1	39.0	49.0	59.3	41.4	41.7	41.4
+ Quote	46.3	72.8	49.8	54.7	31.9	28.0	42.4	28.0
+ Collapsed Quote	**51.2**	73.9	38.7	47.6	58.3	39.4	51.4	39.4
+ Discourse	45.8	75.1	37.7	49.6	63.8	44.6	43.2	44.6
+ Argumentation	48.9	73.6	37.1	47.9	57.1	37.7	53.5	37.7
+ Discourse & Event	46.3	74.3	38.9	49.9	62.1	42.2	42.4	42.2
+ Discourse & Argumentation	47.8	74.1	56.8	50.5	31.4	32.2	41.1	32.2
+ Argumentation & Event	50.0	75.1	38.0	48.6	46.4	44.9	58.5	44.9
+ Quote & Discourse	**51.2**	72.2	40.5	45.3	62.8	43.0	48.7	43.0
+ Collapsed Quote & Discourse	49.6	73.9	45.6	49.4	58.9	39.1	47.9	39.1
+ Collapsed Quote & NLI	45.4	72.8	41.9	50.4	46.7	31.2	39.3	31.2
+ Collapsed Quote & NLI & Event	49.0	73.8	44.9	48.9	57.4	37.0	44.0	37.0
+ All	47.2	73.6	40.0	49.7	58.6	36.0	43.5	36.0
Baseline, *coarse-grained*	49.4	56.7	46.6		65.1		10.4	
+ Discourse & Arg. (Best model, Fact)	65.4	70.7	59.4		66.2		49.2	

Table 1: Various F1 scores (%) on our test set of the fine-tuned LED model with different combinations of features. Fact/Style/Narrative F1 scores are computed on instances that contain the corresponding labels, whereas All F1 scores are derived from all instances.

tag; they are instructed to tag each sentence. To recruit annotators, we posted on two list-serves for journalism industry professionals[3]. We train our annotators until they are all tagging with $\kappa > .6$ agreement, compared with a gold-set of 50 article revision-pairs that we annotated, described previously (Section 3.1). See Appendix for more details.

3.3 Edit Intentions Modeling

Now, we are ready to classify edit intentions between sentences in article revisions. Edit intentions are labeled on the sentence-level, and each sentence addition, deletion or update has potentially multiple intention-labels. Document-level context is important: as shown in Figure 1, understanding that Sentence 2, right, adds background ("*It hit the Fukushima plant, site of previous disaster.*") is aided by the surrounding sentences contextualizing that a major event had just occurred. So, we wish to construct models that can produce flexible outputs and reason about potentially lengthy inputs.

Generative models have recently been shown to outperform classification-based models in document understanding tasks (Li et al., 2021; Huang et al., 2021). Inspired by this, we develop a sequence-to-sequence framework using Long-Former [4] (Beltagy et al., 2020) to predict the intent behind each edit. Specifically, our model processes the input $x = [s_i||s'_j||D||D']$. s_i or s'_j can also be $\varnothing$, which corresponds to the other sentence being a addition/deletion. The decoding target $y_{i,j} = [l_1||\ldots||l_k]$ is a concatenation of ≥ 0 intention labels $1_1, ..., 1_k$ annotated for the pair s_i, s'_j.

Experimental Variants As discussed in Section 3.1, we developed our schema to bring together different theories of news semantics. So, we hypothesize that incorporating insights from these theories into our modeling – specifically, by utilizing labels from trained models in these domains – might improve our performance. We run models from the following papers over our dataset: *Discourse* (Spangher et al., 2021), *Quote-Type Labeling* (Spangher et al., 2023), *Event Detection* (Hsu et al., 2021), *Textual Entailment* (Nie et al., 2020) and *Argumentation* (Al-Khatib et al., 2016). Labels generated from these models, denoted as f_{s_i} and $f_{s'_j}$, are appended to the model input $x = [s_i||s'_j||D||D'||f_{s_i}||f_{s'_j}]$.

Edit-Intention Taggin Model Performance As shown in Table 1, our baseline tagging models that solely use article features score 45.8 Macro F1 and 73.6 Micro F1, respectively. These scores

[3] The Association of Copy Editors (ACES) `https://aceseditors.org/` and National Institute for Computer-Assisted Reporting (NICAR) `https://www.ire.org/hire-ire/data-analysis/`.

[4] `https://huggingface.co/allenai/led-base-16384`

are moderate-to-low. The category we are most interested in, Factual updates, scores at 32 Macro-F1 (derived from macro-averaging the fine-grained categories). However, incorporating additional features increases overall Macro and Micro F1 by 5.5 and 1.5 points, respectively, in the *Quotes & Discourse* trial. And for Factual updates, additional features increase Macro and Micro F1 accuracy by 17.8 and 7.5 points, respectively. While low-to-moderate scores are not ideal, this likely reflects the noisy nature of our problem. *We hope in future work to assess an upperbound on these scores.* For details and schema definitions, see Appendix B.

3.4 Exploratory Insights

Different edit-intentions distribute differently across different edit types (Add, Deletion, Update). We run the models trained in the last section over the entire *NewsEdits* corpus to generate silver-labels on all edit pairs. We present an exploratory analysis of these silver labels, with more material shown in the appendix. Table 2 shows the correlation between syntactic edit categories (defined by (Spangher et al., 2022)) and our semantic categories. As can be seen, categories like Addition have far more Narrative and Factual updates than Stylistic updates; Stylistic updates, on the other hand, are far more likely to occur between sentences. This is logical; Stylistic updates are likely smaller, local updates, while Narrative and Factual updates might include more rewriting.

Different edit-intentions distribute differently across different kinds of news (e.g. Business, Politics). Next, we explore if certain *kinds of articles* are more likely to have certain *kinds of edits*. We start by looking at broad news categories, shown in Table 3, obtained from classifier we train on CNN News Groups dataset[5]. "Politics" and "Sports" coverage are observed to have the highest level of Factual updates, relative to other categories, while Stylistic updates are prevalent in "Health" and "Entertainment" pieces. Although we focus on Factual updates for the rest of the paper, we believe that there are many fruitful directions of future work examining other categories of updates. For instance, stylistic edits made in "Health" news might reach more readers – understanding these patterns might be crucial during times of crisis. We include additional exploration in Appendix A.

[5] https://www.kaggle.com/code/faressayah/20-news-groups-classification-prediction-cnns

	Narrative	**Fact**	**Style**
Addition	840329	358900	104
Deletion	330039	21671	6088
edit	411292	102499	644243

Table 2: Counts of coarse-grained semantic edit types, broken out by syntactic categories (for fine-grained counts, see Appendix).

	Fact	Style	Narrative
Business	1.6	62.0	36.4
Entertainment	3.3	65.5	31.1
Health	2.1	61.0	36.9
News	2.8	57.0	40.2
Politics	5.9	57.8	36.3
Sport	3.5	59.3	37.2

Table 3: Distribution over update-types, across CNN section classifications.

4 Part 2: Predicting Factual Updates

In Section 3, we learned high-scoring models to categorize edit pairs (Equation 1). Now, we wish to leverage these to learn a predictive function:

$$p(l = \text{Factual-Update} | s_i, D) \quad (2)$$

Where s_i and D are the *older* half of a revision pair. Eq 2 seeks to predict how D *might* change.

The problem statement builds off of a line of inquiry introduced in Spangher et al. (2022). Authors introduced tasks aimed at predicting news article developments across time. They tried to predict whether a "sentence will be *Added* to, *Deleted* from, or *Updated* in" an older draft, to induce reasoning about article changes. However, authors stopped at this "syntactic" analysis. Here, we build off of this mode of inquiry: with the semantic understanding of edits introduced in the prior section, we try to predict *how* information will change.

4.1 Factual Edit Prediction Dataset

To construct our task dataset, we sample revision pairs with a non-negligible amount of updates. We sample a set of 500,000 articles from *NewsEdits* that have > 10% sentences added and > 5% deleted. *We acnkowledge that this introduces bias into our dataset*, as we focus solely on a subsection of data we *know* will update. However we build off Spangher et al. (2022)'s broader analysis of syntactic edits patterns, where they found that these kinds of articles could be predicted with reasonable accuracy. We reason that our construction makes it more likely that we are focusing on factual updates

Model	Features	Fact F1	Not Fact F1	Macro F1	Micro F1
GPT-3.5	Sentence-Only	11.3	79.1	30.4	74.2
	Direct Context	3.4	91.8	32.2	85.2
	Full Article	7.9	91.1	49.8	85.4
GPT-4	Sentence-Only	11.1	66.3	38.9	62.4
	Direct Context	14.8	88.8	52.7	84.1
	Full Article	15.4	90.6	53.2	84.9
FT Longformer	Sentence-Only	21.2	92.3	57.4	87.0
	Direct Context	22.3	93.0	87.8	87.4
	Full Article	25.4	91.4	58.0	86.4
Human Performance	Sentence-Only	41.2	75.3	58.6	69.2

Table 4: How well can models predict if a sentence will have a fact update, or not? We test GPT3.5 and GPT4. Individual, macro and micro F1 scores (%) on the golden test set for various evaluated models.

that have more significant impact on the article (as they require more substantial rewrites.)

Then, we use the best-performing edit-intentions model, in Section 3.3, to produce silver labels. We assign labels l using both versions of a revision pair (Equation 1); then we discard D', s'_j and try to predict l using *just* D, s_i (Equation 2).

4.2 Predicting Factual Edits

For training and development, we chronologically split our dataset into train/development sets with 80/20 ratios. The earliest 80% is our training set, the next 20% for development, etc. To keep cost reasonable, we sample 16,000 sentences for the training set and 2,000 for the development set. *We test all approaches on the same gold-labeled documents D^{gold}_{test}, which were part of our gold-annotated test set* (Section 3.2). In early experiments, we noticed that many fine-grained labels were too infrequent to model well, so we switched to predicting coarse-grained labels. We balance the training dataset to have an equal number of classes.

Factual Edit Prediction Experiments We test different variants of Equation 2 to provide different degrees of article context to the model. This helps us understand how much local vs. global article features predict Factual Updates.

(1) Sentence-Only, $p(l|s_i)$;
(2) Direct Context, $p(l|s_{i-1}, s_i, s_{i+1})$
(3) Full Article, $p(l|s_i, D)$.

For each variant we test zero-shot (i.e. prompted `gpt-3.5-turbo` and `gpt-4`); and fine-tuning approaches (i.e. longformer models)[6].

Sent. Contains:	Fact U.	$\overline{\text{Fact U.}}$	Δ
Recent Event	50%	8%	42%
Developing Event	30%	0%	30%
Statistic	28%	8%	19%
Info. request	12%	0%	12%
Historical Event	0%	17%	-17%
Opinion/Analysis	2%	39%	-36%
Description	10%	50%	-40%

Table 5: **Linguistic Cues characterizing Factual Updates**: Manual annotations of characteristics in D^{gold}_{test} sentences that either Factually Update, or not. We show the % of sentences containing these characteristics, ordered by those most salient for Factual Updates.

Results are shown in Table 4. Performance is moderate-to-low for detecting factual updates. However, we do observe performance increases from fine-tuning the longformer model, so to some degree this task is learnable. We recruit a former journalist, with 4 years of experience in major newsrooms, to predict labels for this task, in order to provide a human upper bound to Equation 2. The journalist observes the training data, and then scores the test set. At 41.2 F1-score, the journalist sets a moderately higher upper bound.

Discussion: Linguistic Cues Characterize Factual Edits. LLMs are bad at detecting these. Interestingly, sentence-level characteristics seem to contain much of the signal for this task: as shown in Table 4, the performance barely increases by including the Full Article as context (a finding we did not observe in our tagging task, in Section 3.1). To gain a deeper intuition about these sentence-level cues, we sample 100 sentences from D^{gold}_{test}

[6]The longformer is trained with the same approach as the silver-label prediction step from Section 3.3In early trials, we try different variations on these experiments, like restricting the dataset to different subsets based on topic, like "Disaster" or "Safety". These topic categories, as shown in Section 3.4, are more fact-heavy. However, we find negligible impact on F1-score.

Sentences with ↑ $p(l|s_i, D)$
There are no immediate reports of casualties.
His trial has not yet started.
Officials said attackers fired as many as 30 rockets in Friday's assault.
The rebel group did not immediately comment.

Table 6: A small sample of sentences in the high-likelihood region of $p(l|s_i, D)$. More examples shown in Table 12.

Old sentence: The White House ~~is~~ on lockdown after a vehicle struck a security barrier.
New sentence: The White House was on lockdown for about an hour after a vehicle struck ...
Question: *"Can I visit the White House right now?"*

Table 7: **LLM Abstention Demonstration**: In this example, the LLM only has access to the old, outdated article. We wish to probe whether LLMs can reason about the information's likelihood of being outdated and be cautious about answering this question.

that have been labeled as either having a Factual Update or not (i.e. another kind of update, or no update at all). We show results in Table 5. We identify cues like the temporality of an event described in the sentence as important, and whether the sentence contains statistics, analysis or other kinds of news discourse (Van Dijk, 1998). Interestingly, sentences that Factual Update are more likely to contain Recent Events and Developing Events, compared with Opinion, Historical Events and Description. (See Appendix B.2 for definitions of these discourse patterns).

This would explain in part why language models underperform human reasoning in predicting updates. We find that GPT4 generally has low agreement with human annotators on these tasks, at κ = .2. Researchers have generally found that LLMs struggle with this kind of reasoning (Han et al., 2020; Tan et al., 2023). Recent modeling advancements might help us perform these tasks better (Xiong et al., 2024).

This prediction task is noisy: many sentences may look similar, but may or may not have had Factual Updates, due to chance. Indeed, even expert human annotators have low prediction scores. However, we hypothesize that data that the model is most confident about (or the high-precision region), are more uniformly predictable. We show samples of these sentences in Table 6. These sentences contain many of the linguistic cues identified in 5. See Table 12 for more examples of high-probability sentences (and Table 13 for examples of low-probability sentences). We focus on these high-precision sentences in the next section.

5 Part 3: Question Answering with Outdated Documents

We are ready to test whether the prediction models learned in the last section, to predict whether a sentence will have a Factual update, can help us in dynamic LLM Q&A tasks. We set up a RealTimeQA-style task (Kasai et al., 2022), where an LLM is supplied by a retrieval system with potentially *out-of-date* information. We would like the LLM to *abstain* from answering a question if it suspects it's information might be outdated.

Consider the scenario in Table 7. As humans, we could infer that the ongoing events in the old sentence would be of relatively short time-scale. Thus, if a retriever retrieves the old sentence for the LLM, without knowledge of the new sentence, we would like the LLM to answer the question with something like: "*I do not have the most updated information and this might change quickly*". Confidently answering without any caution as to the updating nature of events is *wrong*.

5.1 LLM-QA Experiments

Experimental Design We take pairs of sentences in the gold test set of our annotated data where an update occurred, and we ask GPT4 to ask questions based on the older sentence.

(1) No-Conflict: 5 questions based on information in the older sentence that does *NOT* update in the newer one.

(2) Maybe-Conflict: 5 questions based on information in the older sentence that *might* update in the newer one.

(3) Likely-Conflict: 5 questions based on information from the older sentence *likely* updates with a newer one. (For all prompts, see Appendix D).

Experimental Variants We devise the following experimental variants. Each variant take in the *old sentence* and a *question*, generated previously.

(1) No Warning (Baseline #1): We formulate a basic prompt to GPT4, without alerting it to any possibly outdated material.

(2) Uniform Warning (Baseline #2) We warn GPT4 that some information might be outdated.

	No-Conflict			Maybe-Conflict			Likely-Conflict		
	Micro F1	Macro F1	Avg.	Micro F1	Macro F1	Avg	Micro F1	Macro F1	Avg.
No Warning	55.9	35.8	55.9	8.8	8.1	8.8	38.8	28.0	38.8
Uniform Warning	52.9	**49.6**	52.9	90.0	47.4	90.0	64.7	54.0	64.7
w. Update Pred.	**59.4**	48.9	**59.4**	**90.6**	61.1	**90.6**	**67.1**	**62.4**	**67.1**
w. Oracle Update	57.6	47.7	57.6	90.0	**63.3**	90.0	66.5	61.1	66.5

Table 8: **LLM-QA Abstention Accuracy**: we measure how often GPT4 correctly abstains from answering user-questions, based on the ground truth of whether the facts in an article updated or not. Each variant shows different information that GPT4 is given. We generate questions in three categories: No-Conflict, Maybe-Conflict, Likely-Conflict, representing how likely the answer to the question will be outdated after a factual update.

	No	Maybe	Likely
No Warning	0.0	0.0	0.0
Uniform Warning	30.0	87.1	98.8
w. Update Pred.	10.6	74.1	95.9
w. Oracle Update	12.4	75.9	94.1

Table 9: **Likelihood of abstaining** in the three test cases: **No** factual conflict, **Maybe** factual conflict, **Likely** factual conflict. In general, we wish to refrain only when we need to. Over-refraining is bad.

The warning is the same for all questions, so GPT has to rely on its own reasoning to detect information that could be potentially outdated.

(3) w/ Our Update Likelihood: We give GPT4 predictions from our Factual Update model, binned into "low", "medium", "high" update likelihood. (We use the highest-scoring LED variation).

(4) w/ Oracle Update: We give GPT4 gold labels that a fact-update *did* or *did NOT* occur. This is designed to give us an upper bound on abstention.

Abstention Rate Evaluations We evaluate performance of each prompting strategy using a GPT4-based evaluation. We ask GPT4: (1) Is this question answerable given the information in the old sentence? (2) Is the answer consistent with the information presented in the revised sentence?

We manually label a small set of 100 questions, to verify that GPT4 can perform this task, and find high agreement $\kappa > .74$ for both questions. If the answer to both questions is yes, the LLM should attempt to provide an answer. If either of the answers is "no", then we want the LLM to ABSTAIN from answering. Abstaining when it *should* is a success; any other answer is a failure. We show F1 scores in Table 8. Interestingly, and perhaps unexpectedly, the variant with Update Predictions does as well if not better than the variant with Oracle Updates. Perhaps the categories of the prediction score helps GPT4 better understand the task compared with the simple yes/no gold labels.

The Uniform Warning (Baseline #2) variation has surprisingly strong performance as well, perhaps an indication that GPT4 does have some emergent abilities to detect the linguistics of outdated information. However, when we examine overall abstention rates, shown in Table 9, we find that this baseline has a far abstention rate. Meanwhile, the variant with Update Predictions abstains at nearly the same rates as that with Oracle Updates.

6 Discussion and Conclusion

The ability of our prediction tags to recover near-oracle performance signals that factual edit prediction can serve a useful role in LLM Q&A. Although we have mainly tested our results in a high-likelihood region of the problem domain as a proof of concept, we suspect that if future work improves the models trained in Section 4.1, then we will see an increase in the ability to drive such abstentions.

We do suspect there to be an inherent upper bound in our ability to model such revision patterns. Randomness undoubtedly exists in the editing and revision process; for many factual updates where, perhaps, the ethical stakes of outdated information are lower, journalists may choose not to go back and revise. We still see such work as promising. Indeed, it is surprising that, despite low scores on the modeling components for Part 1 (Edit-Intention Tagging) and Part 2 (Factual Edit Prediction), we still observe useful downstream applications in Part 3. The linguistic insights we are observe concord with human intuition, and identify known shortcomings of current language models.

Thus, we hope more broadly that the taxonomy introduced in *NewsEdits 2.0* has many rich directions for yielding linguistic insights and better benchmarks. We hope in future work to revise directions around stylistic and narrative edits, both of which we believe can lead to better tools for computational journalists.

7 Ethical Considerations

7.1 Dataset

NewsEdits is a publicly and licensed dataset under an `AGPL-3.0 License`[7], which is a strong "Copy-Left" license.

Our use is within the bounds of intended use given in writing by the original dataset creators, and is within the scope of their licensing.

7.2 Privacy

We believe that there are no adverse privacy implications in this dataset. The dataset comprises news articles that were already published in the public domain with the expectation of widespread distribution. We did not engage in any concerted effort to assess whether information within the dataset was libelious, slanderous or otherwise unprotected speech. We instructed annotators to be aware that this was a possibility and to report to us if they saw anything, but we did not receive any reports. We discuss this more below.

7.3 Limitations and Risks

The primary theoretical limitation in our work is that we did not include a robust non-Western language source. As our work builds off of NewsEdits as a primary corpora, it contains only English and French.

This work should be viewed with that important caveat. We cannot assume *a priori* that all cultures necessarily follow this approach to breaking news and indeed all of the theoretical works that we cite in justifying our directions also focus on English-language newspapers. One possible risk is that some of the information contained in earlier versions of news articles was updated or removed for the express purpose that it was potentially unprotected speech: libel, slander, etc. Instances of First Amendment lawsuits where the plaintiff was successful in challenging content are rare in the U.S. We are not as familiar with the guidelines of protected speech in other countries.

We echo the risk of the original *NewsEdits* authors: another risk we see is the misuse of this work on edits for the purpose of disparaging and denigrating media outlets. Many news tracker websites have been used for good purposes (e.g. holding newspapers accountable for when they make stylistic edits or try to update without giving notice). But we live in a political environment that is often hostile to the core democracy-preserving role of the media. We focus on fact-based updates and hope that this resource is not used to unnecessarily find fault with media outlets.

[7] https://opensource.org/licenses/AGPL-3.0

7.4 Computational Resources

The experiments in our paper require computational resources. Our models run on a single 30GB NVIDIA V100 GPU or on one A40 GPU, along with storage and CPU capabilities provided by our campus. While our experiments do not need to leverage model or data parallelism, we still recognize that not all researchers have access to this resource level.

We use Huggingface models for our predictive tasks, and we will release the code of all the custom architectures that we construct. Our models do not exceed 300 million parameters.

7.5 Annotators

We recruited annotators from professional journalism networks like the NICAR listserve, which we mention in the main body of the paper. All the annotators consented to annotate as part of the experiment, and were paid $1 per task, above the highest minimum wage in the U.S. Of our 11 annotators, all were based in large U.S. cities. 8 identify as white, 1 as Asian, 1 as Latinx and 1 as black. 8 annotators identify as male and 3 as female. This data collection process is covered under a university IRB. We do not publish personal details about the annotations, and their interviews were given with consent and full awareness that they would be published in full.

7.6 References

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News Editorial Corpus for Mining Argumentation Strategies. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

H.D. Croly. 1943. *The New Republic*. v. 108. Republic Publishing Company.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. Wikiatomicedits: A multilingual corpus of wikipedia edits for modeling language and discourse. *arXiv preprint arXiv:1808.09422*.

Haleluya Hadero and David Bauder. 2023. New york times sues microsoft, open ai over use of content. *Globe & Mail (Toronto, Canada)*, pages B1–B1.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020. Econet: Effective continual pretraining of language models for event temporal reasoning. *arXiv preprint arXiv:2012.15283*.

I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, et al. 2021. Degree: A data-efficient generation-based event extraction model. *arXiv preprint arXiv:2108.12724*.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Adam Liska, Tomás Kociskỳ, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, and Devang Agrawal. 2022. Cyprien de masson d'autume, tim scholtes, manzil zaheer, susannah young, ellen gilsenan-mcmahon, sophia austin, phil blunsom, and angeliki lazaridou. 2022. streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999.

Kostas Saltzis. 2012. Breaking news online: How news stories are updated and maintained around-the-clock. *Journalism practice*, 6(5-6):702–710.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying informational sources in news articles. *arXiv preprint arXiv:2305.14904*.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Alexander Spangher, James Youn, Matthew Debutts, Nanyun Peng, and Jonathan May. 2024. Explaining mixtures of sources in news articles.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Teun A Van Dijk. 1998. *News as discourse*. Lawrence Erlbaum Associates.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 927–932.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

W Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 133–143.

	Fact	Style	Narrative
Disaster	6.4	43.4	50.0
Elections	5.1	47.9	46.9
Environment	1.9	56.8	41.2
Labor	2.0	49.6	48.2
Other	3.7	50.7	45.5
Safety	4.7	46.6	48.6

Table 10: Distribution over update-types, across social-interest categories (Spangher et al., 2023).

A Additional EDA

We show the following different analyses to support the findings in the main body.

Table 10 shows the kinds of edits in 6 different categories of news determined "socially beneficial", by (Spangher et al., 2023)[8]. As can be seen, even though Factual updates are rarer overall in sentence-level updates, they are more represented in Disaster and Safety categories.

In Figure 7, we perform an error analysis on our best-performing ensemble model, which includes tags from Argumentation and Discourse. We inspect the categories we are most likely to get wrong. As can be seen, our fine-grained accuracy is actually quite low, indicating the value of future work, perhaps collecting more training data or employing LLMs to label more silver-standard data. Many categories on the diagonal have 0 labels, both because many categories are low-count categories (e.g. "Define Term", which does not have *any* gold-truth labels in the test set), as well as that more dominant categories capture many of the predictions (e.g. "Tonal Edits").

However, the problem is slightly less sever on the coarse-grained level, shown in Figure 6. By comparing these two categories, we can see that many of the errors we observed are on the fine-grained level are within the same coarse-grained category. We suspect that to raise accuracy for fine-grained labels further, we need further experimentation is needed. Perhaps we can experiment with approaches involving more specific fine-grained models or with data augmentation.

A.1 Further details about high-precision sentences

Figure 4 shows more details of our exploration into the predictability of higher-precision fact-update sentences: as we restrict the pool of documents, we increase the performance.

[8] To group news articles in these categories, we use a classifier released by the authors

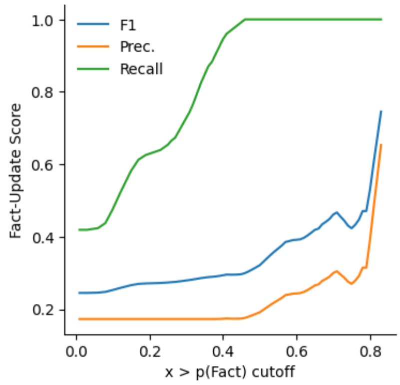

Figure 4: Performance of Fact-update model increases as we increasingly focus on a pool of documents that are categorized as high-likelihood under the top-performing LED model (in Table 1). In other words, the model truly shines in the high-precision, high-probability realm.

A.2 Technical Improvements over *NewsEdits* Edit-Action Algorithm

Spangher et al. (2022) identified "edit-actions", or "syntactic" edits in article revision histories (i.e. sentence additions, deletions and updates), which requires them to match sentences across article versions. They report a 89.5 F1 efficacy at matching sentences, a significantly higher rate than we might expect for lexical matching. We examined *NewsEdits*'s sentence matches and found that a large source of errors stem from poor sentence boundary detection (SBD). Poor SBD creates an abundance of sentence stubs, which often over-match across revisions. We reprocessed the dataset from scratch using spaCy[9] instead of SparkNLP for SBD[10], which we qualitatively observe to be better. For word-matching, we use `albert-xxlarge-v2`[11]'s embeddings (Lan et al., 2019) instead of `TinyBert` (Jiao et al., 2019). These steps, we find, increase our linking accuracy to 95 F1-score. We reprocess and re-release *NewsEdits*. In addition, we release a suite of visualization tools, based on D3[12] to enable further exploration of the corpus. See Appendix C.2 for an example.

[9] https://spacy.io/, specifically, the en_core_web_lg model.

[10] https://sparknlp.org/api/com/johnsnowlabs/nlp/annotators/sbd/pragmatic/SentenceDetector.html

[11] https://huggingface.co/albert/albert-xxlarge-v2

[12] https://d3js.org/

	Addition	Deletion	Edit
Add/Delete/Update Background	806909	329652	411025
Add/Delete/Update Quote	303451	17995	46300
Incorrect Link	191022	125362	237437
Other (Please Specify)	84646	66929	65077
Add/Delete/Update Event Reference	37409	3645	56098
Add/Delete/Update Analysis	33426	390	268
Add/Delete/Update Eye-witness account	9772	0	3
Add/Delete/Update Source-Document	6639	2	28
Add/Delete/Update Information (Other)	1058	13	3
Additional Sourcing	573	15	29
Tonal Edits	102	6000	616514
Emphasize/De-emphasize Importance	1	32	1076
Syntax Correction	1	2	21729
Emphasize/De-emphasize a Point	0	53	1668
Simplification	0	0	3
Style-Guide Edits	0	1	3253
Correction	0	1	47

Table 11: Counts of fine-grained semantic edit types, broken out by syntactic categories

B Details of the LED Model

In this section, we describe the specifications of the LED model described in Section 3.3.

B.1 Input Template

The input to the LED model is shown below:

```
Predict the edit intention from
version 1 to version 2.
Version 1: SOURCE_SENTENCE
Version 2: TARGET_SENTENCE
Version 1 Document: SOURCE_DOCUMENT
Version 2 Document: TARGET_DOCUMENT
```

Here, **SOURCE_DOCUMENT** (D) and **TARGET_DOCUMENT** (D') refer to the newer and older articles, while **SOURCE_SENTENCE** (s_i) and **TARGET_SENTENCE** (s'_j) represent a sentence with these articles.

B.2 Additional Schema

NLI We use textual entailment from (Dagan et al., 2005), which consists of *Entail*, *Contradict* and *Neutral*. These categories indicate whether two pieces of information refute each other, complement each other, or are neutral. We use a trained model by (Nie et al., 2020), which is an adversarially-trained Albert-xxlarge model, to label pairs of sentences (one from the old version, one from the new version).

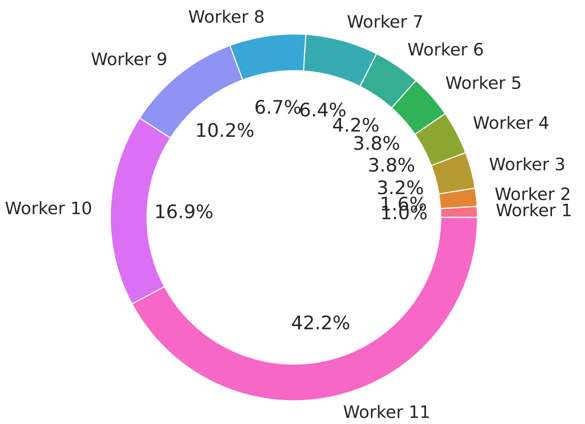

Figure 5: The portion of annotation tasks assigned to each worker.

Event Detection As described by Doddington et al. (2004) in the coding guidelines for the ACE-2005 dataset, "*An Event is a specific occurrence involving participants. An Event is something that happens. An Event can frequently be described as a change of state.*" Several datasets exist which label events in text, like ACE-2005, and a wide body of research has since emerged to model and detect events in text. Such models detect *triggers* (i.e. mostly verb-forms that signal the presence of an event); *types* (i.e. broad taxonomies that events fall into) and *arguments* (i.e. people, places or other lexical units associated with the occurrence of the event which further define it).

We use a model by (Hsu et al., 2021), designed to detect events in a wide variety of settings. We only consider whether an event trigger exists in a sentence, as a binary variable (0=no trigger exists, 1=trigger exists). Our theory is that this can help with tags like "Delete/Add/Update Event".

Argumentation Defined in Al-Khatib et al. (2016), *Argumentation* is a type of discourse schema that defines what kinds of evidence the writer marshalls to make their point. Authors define the following categories: *Anecdote*, *Assumption*, *Common Ground*, *Statistics*, *Testimony*, *Other*. They primarily study news editorials (i.e. opinion pieces), where they assume they have the most different kinds of argumentation categories. Spangher et al. (2021) and Spangher et al. (2024) show that these models can generally be applied helpfully across a broader news domain. We include them in the present study to capture aspects like "Anecdote" that capture framing aspects of journalistic writing.

Quote Quote-detection is a long-standing task, usually involving detecting the presence of direct or indirect quotes (Pareti et al., 2013). We use the broad definition of a "quote" as "information derived from any source external to the news article and the journalist's own thoughts", as defined in Spangher et al. (2023). Authors developed and released models for detecting when sentences had information that could be attributable to a named or unnamed source in the news article. We use these models to apply a simple binary indicator for whether or not the sentence contained a quote (1=sentence contains a quote, 0=it does not). We include this under the hypothesis that it can help us improve our detection in categories like "Delete/Add/Update Quote".

News Discourse The News Discourse schema, as defined by Van Dijk (1998) views news stories as a sequence of structural elements, each serving a different narrative role. As implemented separately by (Choubey et al., 2020), (Yarlott et al., 2018) and (Spangher et al., 2021), the news discourse schema has undergone some modifications since Van Dijk (1998)'s original formulation, most notably to include current theories on event detection. It includes the following elements: *Main Event*, *Consequence*, *Previous Event*, *Current Context*, *Evaluation*, *Expectation*, *Historical Events*,

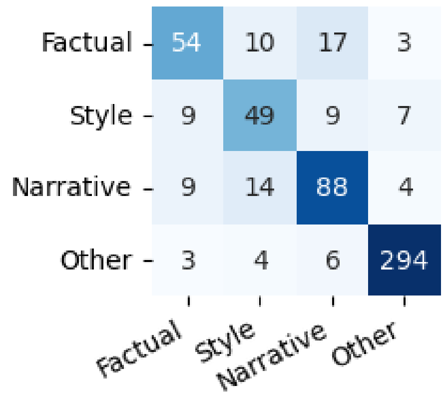

Figure 6: Coarse-grained confusion matrix for the LED model trained with Discourse and Argumentation features.

Anecdotal Event. We believed that, since much of our edit schema was inspired by notions of narration, like "Delete/Add/Update Background", we could get signal from this schema.

C Annotation Details

In this section, we provide details of the annotation process, such as annotation guidelines and task allocation.

C.1 Annotation Guidelines

To complete the task, look at each sentence: if it's been added, updated, or deleted between drafts, try to determine based on your knowledge of the journalistic editing process why this was done.

You can specify multiple intentions for each add/delete/edit operation. Please also pay attention to when sentences are moved around in a document (i.e. if that was done to emphasize or de-emphasize that sentence), and when there might be errors to how we are linking sentences.

We devised these in consultation with professional journalists. However, if you are consistently annotating edits with "Other" (i.e. we are missing something in our schema), please let us know!

Fact Edits:

- **Delete/Add/Update Eye-witness Account**: The writer deletes/adds/updates the contents for the events being described. This can either take the form of a quote (in which case this edit should be paired with a Quote Update), or a first-person account by the journalist.
- **Delete/Add/Update Event**: There is a change to some event in the world that the article

covers and the article needs to be updated to reflect this. Usually, there are changes to the verbs in the article, but this can also include increased death counts, stock-market changes, etc.

- **Delete/Add/Update Source-Doc**: Additional written documents have been released by a government or company that warrant deletion/inclusion/update of the content of the article. For example, additional information included in an SEC filing, quarterly earnings report, IPCC report, etc.

- **Correction**: There are factual errors in the original version. The new version corrects the error.

- **Delete/Add/Update Quote**: There is an addition, editing or deletion of quotes in the article. Or, a quote from one person is swapped for a quote from another. Sometimes these updates are made with other intentions (e.g. to include a punchier quote, in which case it would also be a Preferential Edit. In these cases, please use the "+" button to add another intention dropdown.)

- **Additional Sourcing (Other)**: The new version includes evidence of new sources for additional information, usually added for confirmation purposes. Note that this is different from Quote Update or Document Update since Additional Sourcing doesn't have to result in a new quote or document reference. Can simply be an indication that the journalist obtained new evidence.

- **Additional Information (Other)**: This edit intention is applied when the new version of the article includes details or context not present in the original version, which doesn't necessarily fall under specific updates like eyewitness accounts, event changes, document updates, or sourcing alterations.

Style Edits:

- **Simplification**: educes the complexity or breadth of discussion. This edit might also remove information from the article.

- **Emphasize/De-emphasize Importance**: The sentence is moved up or down in the document in order to make the sentence MORE/LESS prominent, or to emphasize/de-emphasize it's connection to the events being described in another sentence.

- **Define term**: The author provides meaning or differentiation to a term or concept that might be unknown to the reader. Note that this intention is DIFFERENT from the Background intention, which is more about providing context, e.g. historical or geographic context for a person, company, or place.

- **Style-Guide Adherence**: Edits that are made specifically to address a formal style guide (when in doubt, defer to the Associated Press style-guide). The first version violates the style guide and the revised version fixes it.

- **Syntax Correction**: Improve grammar, spelling, or punctuation. These are strictly to correct errors in syntax, not **Preferential Edits**. And, they need not be adhering to a formal style-guide (when a **Syntax Correction** is also adhering to a **Style Guide**, please use the "+" button to add another intention dropdown and annotate both).

- **Tonal Edits**: The journalist or copy-editor made the edits due to a specific personal or artistic preference. Use your intuition here: these are usually edits that introduce punch, elegance or scenery. These edits often also have the effect of some other edit intention, see the example, but cannot be fully ascribed to other aims.

- **Sensitivity Consideration**: The journalist rewrote the sentence because the original version is inappropriate/ may be considered insensitive.

Narrative Edits:

- **Delete/Add/Update Analysis**: The writer deletes/adds/updates inferences from the presented information. These can be in the form of analyses, expectations, or deeper understandings. These are usually forward-looking rather than Background information, which is usually past-looking.

- **Delete/Add/Update Background**: Delete/add/update contextualizing information to the article to help readers understand the history, geography or significance of a term, personal, place or company. Note that contextualizing information is not analysis, expectations, or projections, which would fall into the Analysis intention category.

- **Delete/Add/Update Anecdote**: The writer deletes, adds, or updates a brief, revealing account of a person or event. This can be a personal story, a particular incident, or a narrative snippet that exemplifies a point or adds a humanizing or illustrative dimension to the news piece. These anecdotes may serve to engage the reader's interest, illuminate a fact, or provide a real-world example of abstract concepts.

Others:

- **Incorrect Link**: This refers to an error in our original linking of sentences. We have linked two sentences that should NOT be linked. This only pertains to 'Edit'ed or 'Unchanged' sentences. Sentences should not be linked if they are entirely unrelated — they have substantially different syntax, intent, and purpose — and, by error, our algorithm said they were. If you identify an **Incorrect Link** AND there are more than one links, please specify (A) the index of the sentence in the other version that it should NOT be linked to via the dropdown (B) any other intention ascribed to this pair (i.e. Fact Deletion).

C.2 Annotation Interface

Figure 8 shows the annotation interface for our task. Users are shown pairs of sentences, as identified in NewsEdits (Spangher et al., 2022) and have the option to annotate edits, additions and deletions with different edit intentions. Additionally, users can annotate when the links are incorrect.

C.3 Annotation Task Distribution

We asked prospective applicants to describe their journalism experience, and selected this pool based on those having one or more year of professional editing experience. Then, we asked them to label revised sentences in five news articles, which we checked. We recruited 11 annotators who scored above 90% on these tests.

In Figure 5, we show the portion of annotation tasks assigned to each worker. As can be seen, we have a broad mix of users. Worker 11 is a professional journalist we worked most often with, and annotated a plurality of the tasks.

D Prompts for Use-Case

D.1 Question-Asking Prompts

D.1.1 No-Conflict

Prompt Outline

```
I will give you a sentence and you
will give me 5 different questions.
It should be directly answerable by
the sentence.
Here are some examples:
Example 1: EXAMPLE
Example 2: EXAMPLE
Example 3: EXAMPLE
Ok, now it's your turn.
Here is a sentence: SENTENCE Ask 5
different questions, output in a list.
Don't say anything else.
```

Examples *sentence*: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." *question*: "What did the vehicle strike?"

sentence: "The death count from the 42nd street bombing is 49 injured, 2 killed so far." *question*: "Where did the bombing take place?"

sentence: "The construction work left the bridge badly damaged and unsafe for passengers and is expected to remain so for days." *question*: "What kind of work was being done?"

D.1.2 Maybe-Conflict

Prompt Outline

```
I will give you a sentence and you
will give me an answer. It should
be timely and related to the facts in
the sentence. It should be a question
that could go stale, especially for
ongoing events, or facts like death
counts that might update.
Here are some examples:
Example 1: EXAMPLE
Example 2: EXAMPLE
Example 3: EXAMPLE
Ok, now it's your turn.
Here is a sentence: SENTENCE Ask 5
different questions, output in a list.
Don't say anything else.
```

Examples *sentence*: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." *question*: "Is the White House currently in lockdown – if I visit, will I get turned away?"

sentence: "The death count from the street bombing is 49 injured, 2 killed so far." question: "How many people have died so far?"

sentence: "The construction work left the bridge badly damaged and unsafe for passengers and is expected to remain so for days." *question*: "What route should I take? The bridge is the quickest way to work."

D.1.3 Likely Conflict

Prompt Outline

```
I will give you two sentences from an updating news article and you will give me 5 different questions. They should ideally focus on information that changes in between the sentences. So, if someone were to just look at the old sentence and you asked them your question, they would get it wrong.
Ok, now it's your turn. Here is the old sentence: OLD_SENTENCE Here is the new sentence: NEW_SENTENCE Ask 5 different questions, output in a list. Don't say anything else.
```

Examples *old sentence*: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." *new sentence*: 'WASHINGTON (AP) – The White House was on lockdown for about an hour Friday after a passenger vehicle struck a security barrier.' *question*: "Is the White House currently in lockdown – if I visit, will I get turned away?"

old sentence: "ISTANBUL (AP) – An earthquake with a preliminary magnitude of 6.2 shook western Turkey and the Greek island of Lesbos Monday, scaring residents and damaging buildings." *new sentence*: "ISTANBUL (AP) – An earthquake with a preliminary magnitude of 6.2 shook western Turkey and the Greek island of Lesbos on Monday, injuring at least 10 people and damaging buildings, authorities said." *question*: "Was anyone injured?"

old sentence: "Turkey's emergency management agency said there were no reports of casualties in the country." *new sentence*: "Turkey's emergency management agency said there were no reports of casualties and has dispatched emergency and health teams, and 240 family tents to the area as a precaution." *question*: "Is the Turkish emergency management doing anything as a precaution?"

D.2 Question Answering Prompts

D.2.1 Experimental Prompt

```
You are a helpful assistant who answers questions based on this news information:
NEWS_ARTICLE_SENTENCE

We give this a HIGH/MEDIUM/LOW chance of there being a fact update in this sentence. That might mean some new information could make some of the information in this sentence outdated. The user will ask a question. Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it. Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else.

Keep our estimate in mind.
```

D.2.2 Baseline 1

```
You are a helpful assistant who answers questions based on this news information:
NEWS_ARTICLE_SENTENCE

Try to directly answer the users question and say nothing else.
```

D.2.3 Baseline 2

```
You are a helpful assistant who answers questions based on this news information:
NEWS_ARTICLE_SENTENCE
```

This sentence might go out of date. Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it.

Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else.

D.2.4 Oracle

You are a helpful assistant who answers questions based on this news information:
NEWS_ARTICLE_SENTENCE

This sentence **DOES / DOES NOT** have a major fact update. That might mean some new information, updating information. Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it.

Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else.

D.3 Evaluation Prompts

You are a helpful assistant. You will be shown an old sentence, a revised sentence, and a user-question. you will answer the following 2 questions:
1. Is this question answerable given JUST the old sentence?
Answer with "yes" or "no". Do not answer anything else. If the answer to 1 was yes, then proceed to the second question, otherwise respond to question 2 with n/a
2. Does the question ask about something that is factually consistent with the information presented in the revised sentence?
Answer with "yes", "no" or "n/a." Do not answer with anything else.

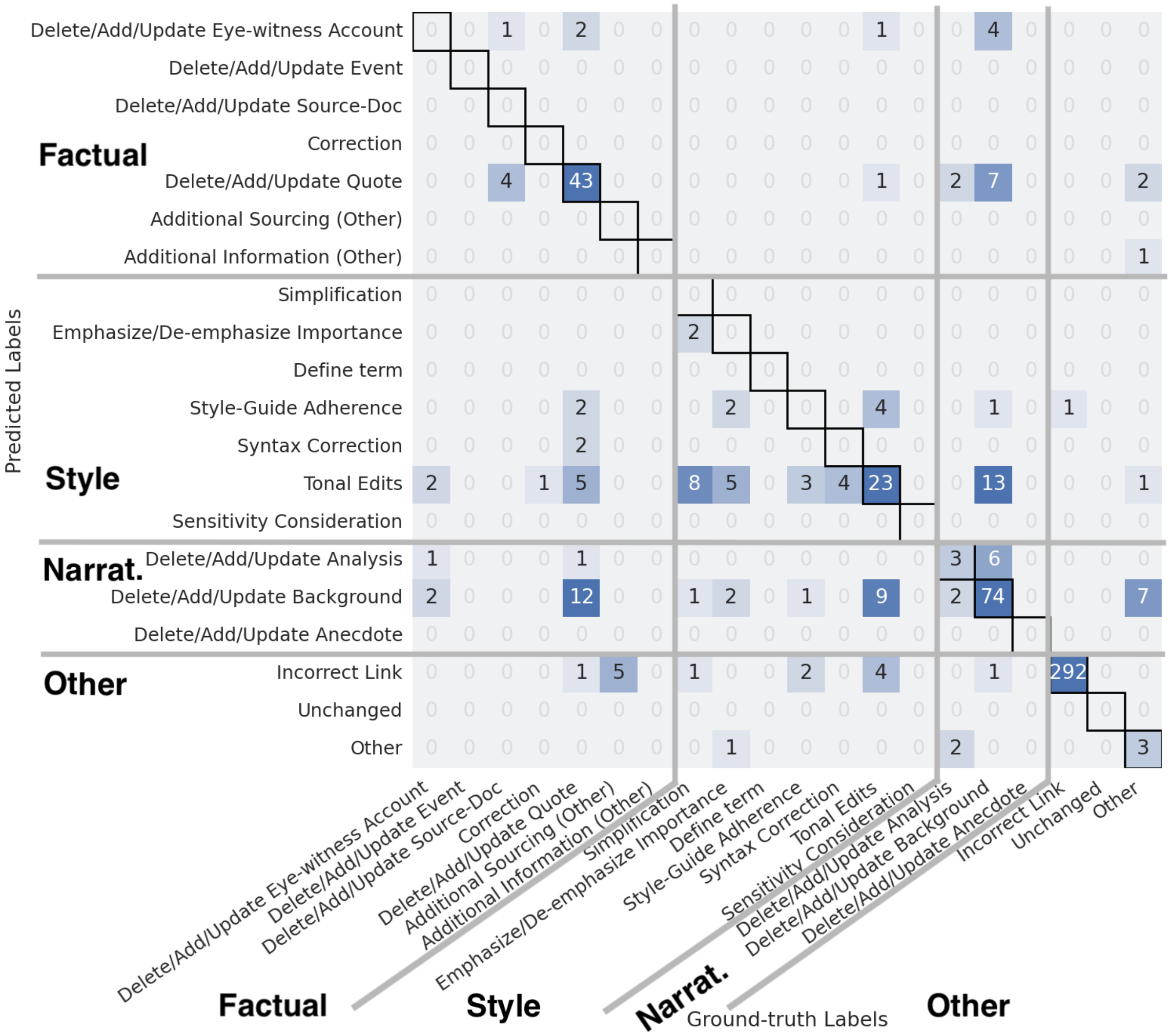

Figure 7: Fine-grained confusion matrix for the LED model trained with Discourse and Argumentation features.

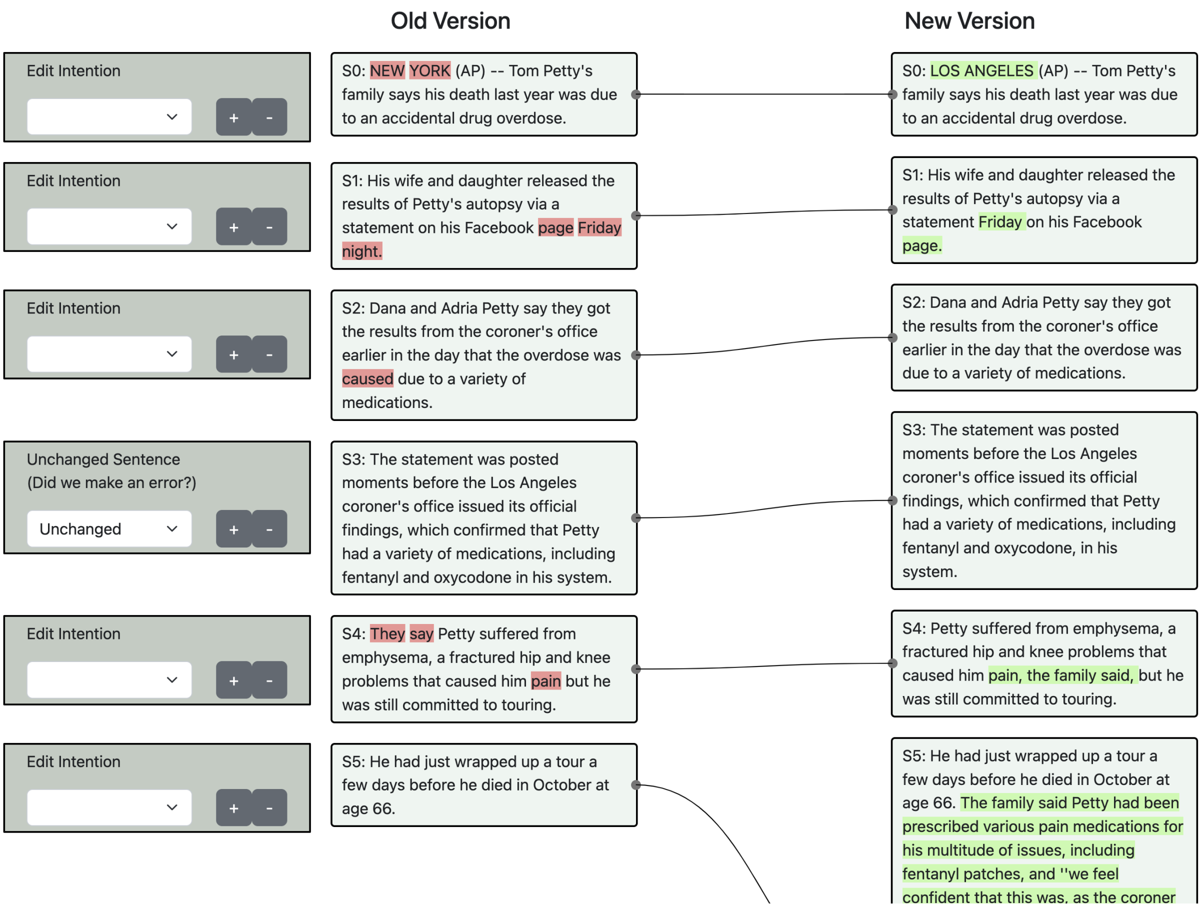

Figure 8: The interface for annotating edit intentions.

Top Predictions for Content Evolution Prediction, $p(l = \text{Fact Update}|s_i, D)$
The company takes this recommendation extremely seriously," it said in a statement.
KABUL, Afghanistan — An Afghan official says a powerful suicide bombing has targeted a U.S. military convoy near the main American Bagram Air Base north of the capital Kabul.
WASHINGTON — The U.S. carried out military strikes in Iraq and Syria targeting a militia blamed for an attack that killed an American contractor, a Defense Department spokesman said Sunday.
Mr. Causey, who reported his concern to authorities, was not charged in the indictment, which a grand jury returned last month, and did not immediately comment.
His trial has not yet started.
MEXICO CITY — A fiery freeway accident involving a bus and a tractor-trailer killed 21 people in the Mexican state of Veracruz on Wednesday, according to the authorities and local news outlets.
The indictment accuses Mr. Hayes, a former congressman, of helping to route $250,000 in bribes to the re-election campaign of Mike Causey, the insurance commissioner.
No Kenyans died in the attack, Kenya's military spokesman Paul Njuguna said Monday.
Mr. Manafort, 70, will most likely be arraigned on the new charges in State Supreme Court in Manhattan later this month and held at Rikers, though his lawyers could seek to have him held at a federal jail in New York, the people with knowledge said.
Officials said attackers fired as many as 30 rockets in Friday's assault.
KABUL, Afghanistan — Gunmen attacked a remembrance ceremony for a minority Shiite leader in Afghanistan's capital on Friday, wounding at least 18 people, officials said.
BEIRUT — A senior Turkish official says Turkey has captured the older sister of the slain leader of the Islamic State group in northwestern Syria, calling the arrest an intelligence "gold mine. "
Paul J. Manafort, President Trump's former campaign chairman who is serving a federal prison sentence, is expected to be transferred as early as this week to the Rikers Island jail complex in New York City, where he will most likely be held in solitary confinement while facing state fraud charges, people with knowledge of the matter said.
The watchdog, the Securities and Exchange Surveillance Commission, said Tuesday it made the recommendation to the government's Financial Services Agency on the disclosure documents from 2014 through 2017.
There are no immediate reports of casualties.
It said the U.S. hit three of the militia's sites in Iraq and two in Syria, including weapon caches and the militia's command and control bases.
The rebel group did not immediately comment.
Kep provincial authorities later announced a total of five dead and 18 injured.
QUETTA, Pakistan — Attackers used a remotely-controlled bomb and assault rifles to ambush a convoy of Pakistani troops assigned to protect an oil and gas facility in the country's restive southwest, killing six soldiers and wounding four, officials said Tuesday.
WASHINGTON — Senator Bernie Sanders of Vermont raised $18.2 million over the first six weeks of his presidential bid, his campaign announced Tuesday, a display of financial strength that cements his status as one of the top fund-raisers in the sprawling Democratic field.

Table 12: Sample of the most likely fact-update sentences, as judged by our top-performing model. Top predictions reflect a combination of statistics, recent or upcoming events, and waiting for quotes.

Lowest Predictions for Content Evolution Prediction, $p(l$ = Fact Update$|s_i, D)$
Sir Anthony Seldon, vice-chancellor of the University of Buckingham, said: "Cheating should be tackled and the problem should not be allowed to fester any longer. "
He added: "This shows the extent to which a party which had such a proud record of fighting racism has been poisoned under Jeremy Corbyn. "
But he said his dream of making it in the game had turned into a nightmare. “
Adam Price, Plaid Cymru leader, said: "There is now no doubt that Wales should be able to hold an independence referendum. "
Others told how excited they had been when they were scouted by Higgins. “
The former Conservative deputy prime minister said it was “complete nonsense” to suggest Brexit could be done by Christmas. “
He said the QAA identified 17,000 academic offences in 2016 - but it was impossible to know how many cases had gone undetected. "
Nationalism leads a "false trail" in ""exactly the opposite direction", he argued, "one that pits working people against each other, based on the accident of geography".
He also suggested that universities should adopt "honour codes", in which students formally commit to not cheating, and also recognise the consequences facing students who are subsequently caught.
He added: "But my experience is, if you make that threat, you don’t actually need to follow through with the dreaded milkshake tax. "
He said: “There’s an anger inside of me, a feeling of disgust that turns my stomach. ”
Damian Hinds says it is "unethical for these companies to profit from this dishonest business".
She added: “His plan to hold another two referendums next year – and all the chaos that will bring – will mean that his government will not have time to focus on the people’s priorities. “
We would be happy to talk to the Department of Education about their concerns." ’
I am determined to beat the cheats who threaten the integrity of our system and am calling on online giants, such as PayPal, to block payments or end the advertisement of these services - it is their moral duty to do so," said Mr Hinds.
The chief executive of Action on Smoking and Health, Deborah Arnott, also warned it would be a "grave error" to move away from taxing cigarettes. "
Rather than just taxing people more, we should look at how effective the so-called ’sin taxes’ really are, and if they actually change behaviour. "
He added: "How many more red lines will be laid down by sensible Labour MPs, only for the leadership to trample right over them?
This shows that the complaints process is a complete sham," she tweeted. "
Mr Hinds added that such firms are "exploiting young people and it is time to stamp them out". "
One said he was abused by Higgins in a gym.

Table 13: Sample of the least likely fact-update sentences, as judged by our best-performing model. Predictions represent a combination of opinion quotes or anecdotes, projects and longer-term plans.