Not all Hallucinations are Good to Throw Away When it Comes to Legal Abstractive Summarization

Nihed Bendahman [♢♠], **Karen Pinel-Sauvagnat** [♢],
Gilles Hubert [♢], **Mokhtar Boumedyen Billami** [♠]
[♢]Université de Toulouse, IRIT, UMR 5505 CNRS, Toulouse, France
[♠]Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France
{nihed.bendahman, karen.sauvagnat, gilles.hubert}@irit.fr
{nihed.bendahman, mb.billami}@berger-levrault.com

Abstract

Automatic summarization of legal documents requires a thorough understanding of their specificities, mainly with respect to the vocabulary used by legal experts. Indeed, the latter rely heavily on their external knowledge when writing summaries, in order to contextualize the main entities of the source document. This leads to reference summaries containing many abstractions, that sota models struggle to generate. In this paper, we propose an entity-driven approach aiming at learning the model to generate factual hallucinations, as close as possible to the abstractions of the reference summaries. We evaluated our approach on two different datasets, with legal documents in English and French. Results show that our approach allows to reduce non-factual hallucinations and maximize both summary coverage and factual hallucinations at entity-level. Moreover, the overall quality of summaries is also improved, showing that guiding summarization with entities is a valuable solution for legal documents summarization.

1 Introduction

The extensive use of Large Language Models (LLMs) for summarization tasks brings new challenges. One of the most important is related to the LLMs' tendency to generate *hallucinations*, i.e., texts that give the impression of being fluid and natural, despite their lack of fidelity or nonsensical nature (Ji et al., 2023). Hallucinations can be non-factual or factual. In this last case, a source of knowledge, such as an expert, an ontology or a knowledge base, can indeed certify their veracity (Maynez et al., 2020; Feng et al., 2023).

Most of the current works on hallucinations aim at deleting, reducing or correcting non-factual hallucinations through dedicated models (Narayan et al., 2021; Nan et al., 2021a; Zhang et al., 2022) or post-processing approaches (Cao et al., 2020; Zhao et al., 2020; Chen et al., 2021). Some other works try to control factual hallucinations through the use of external knowledge, while keeping low the number of non-factual hallucinations (Cao et al., 2022; Dong et al., 2022; Chern et al., 2023).

This last line of research is particularly of interest while considering specific domains such as the legal field. Indeed, summaries of legal documents written by experts often contain domain-specific knowledge and references to previous laws that help the readers to fully grasp the content. This is illustrated in Figure 1 through an example of legal document with its associated reference summary, extracted from the EUR-Lex-Sum dataset[1]. It can be observed that both the document and the reference summary include a large number of entities of interest (in bold). Moreover, the reference summary mentions laws *Regulation 2019/979, Regulation 2020/1273*, and *Regulation 2020/1272* (colored in dark and light blue on Figure 1) that do not appear in the document. Such entities have been added to the summary by experts in order to track all the amendments made to the *Regulation 2019/980* presented in the document. They are named *abstractive* entities, in opposition to faithful entities, i.e., entities present in the source document (Dong et al., 2022) (in green in the reference summary). Abstractive entities, as they are written by experts, are de facto factual.

In this paper, we focus on abstractive summarization of legal documents, with the main objective to produce enriched, useful, and factual summaries which will lighten the task of information monitoring by legal experts. To do so, we aim at controlling factual hallucinations by knowledge arisen from the source dataset. In other words, we propose an approach which aims at fulfilling a two-objectives task: (1) retrieve factual hallucinations from the dataset; and (2) keep very low the number of non-factual hallucinations.

[1] https://eur-lex.europa.eu/eli/reg_del/2019/980/oj

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5331–5344

Source Document : Official Journal of the **European Union COMMISSION** DELEGATED **REGULATION (EU) 2019/980** of 14 March 2019 supplementing **Regulation (EU) 2017/1129** of the **European Parliament** and of the **Council** as regards the format, content, scrutiny and approval of the prospectus to be published when securities are offered to the public or admitted to trading on a regulated market, and repealing Commission **Regulation (EC) No 809/2004** (Text with **EEA** relevance) THE **EUROPEAN COMMISSION**, Having regard to the Treaty on the Functioning of the **European Union**, Having regard to **Regulation (EU) 2017/1129** of the **European Parliament** and of the **Council** of 14 June 2017 on the prospectus to be published when securities are offered to the public or admitted to trading on a regulated market, and repealing **Directive 2003/71/EC** (1), and in particular Article 9(14), paragraphs (1) and (2) of Article 13, Articles 14(3), 15(2) and 20(11) thereof, Whereas: (1) **Regulation (EU) 2017/1129** [...]

Reference Summary : [...] The **European Commission** has adopted the following acts : Delegated **Regulation (EU) 2019/980** laying down details of the precise content and format of the prospectus, as well as concerning the scrutiny and approval of the prospectus. Delegated **Regulation (EU) 2019/979** laying down details of the key financial information in the summary of the prospectus, the publication and classification of the prospectus, the advertisement for the securities and the supplement to a prospectus. Delegated **Regulation (EU) 2020/1273** amending and correcting Delegated **Regulation (EU) 2019/980**. Delegated **Regulation (EU) 2020/1272** amending and correcting Delegated **Regulation (EU) 2019/979** [...].

Generated Summary : [...] The Commission Delegated **Regulation (EU) 2019/980**, dated 14 March 2019, supplements **Regulation (EU) 2017/1129** of the **European Parliament** and **Council**. It addresses the format, content, scrutiny, and approval of the prospectus to be published when securities are offered to the public or admitted to trading on a regulated market, effectively repealing Commission **Regulation (EC) No 809/1273**. It includes different key points: - Different types of prospectuses are subject to specific information requirements depending on the issuer, type of security, and market admission, also mentioned in **Regulation (EU) 2020/1273**. - As in **Regulation 2019/977** Universal registration documents must clearly state whether they have been approved or simply filed with the competent authority. Secondary issuances, non-equity securities, collective investment undertakings [...]

Figure 1: Example of a EUR-Lex-Sum legal document with its reference summary (excerpts) produced by experts, as well as an example of generated summary. Entities of interest are marked in bold. In the reference summary, faithful entities are colored in green. Abstractive entities of the reference summary are colored in dark blue when present in the dataset (in-dataset abstractions) whereas the light blue one is an out-dataset abstraction. In the generated summary, faithful entities are also colored in green. The entities colored in red are non-factual hallucinations whereas the yellow one is a factual hallucination.

Our work focuses on hallucinations and abstractions at entity level for two main reasons: (i) entities are more striking to human readers (Ji et al., 2023), and (ii) the legal domain is overflowing with entities of interest (law references, institutions, etc.). First, we retrieve entities from the source dataset that closely match the set of abstractive entities of the reference summary. Then, we introduce various approaches for injecting the retrieved entities during the training of the model, in order to control the generated entities (faithful and hallucinated ones) at inference. To the best of our knowledge, this is the first approach aiming at controlling factual hallucinations using only the original dataset as a source of knowledge.

To evaluate our approach, we conducted various experiments on two different legal corpora in English and French. We aimed at answering the following questions:

(Q1) What are the best strategies to identify abstractive entities in the source dataset ?

(Q2) What are the best training scenarios to take advantage of the entities identified in Q1?

(Q3) Does guiding the model by entities improve the overall quality of the generated summary?

2 Context and Problem Formulation

2.1 Hallucinations and Abstractions

As there is no formal consensus on the vocabulary used by state-of-the-art approaches about the concepts of hallucinations and abstractions in a summarization context, we detail in this section the definitions considered in this article.

In literature, hallucinations and abstractions are usually considered at two different levels: the entity one and the relationship/information one (Lyu et al., 2022; Maynez et al., 2020; Ji et al., 2023). Since the approach proposed in this work is driven by entities of interest, we consider definitions at entity level in the following.

Figure 2 proposes a taxonomy of these definitions, while the colors highlighting entities on Figure 1 illustrate the taxonomy. *Faithful entities* (in green) are entities in the reference or generated summary that can be found in the source document. *Abstractions* may occur in the reference summary while *hallucinations* are specific to the generated summary. Abstractions are produced by experts, who own an omniscient knowledge, when summarizing a document, incorporating factual entities in the (reference) summary they write that are not present in the document. We classify abstractions into two categories: those that can be found in the dataset the source document belongs to (*in-dataset* abstractions, in dark blue) and those that come from external world knowledge (*out-dataset* abstractions, in light blue), which we cannot know. Hallucinations can be of two types: non ascertainable from the source document but factual with respect to the domain knowledge (in yellow) and non factual (in red). Domain knowledge can originate from an ontology, a knowledge base or the expert's own prior knowledge.

Our approach aims at controlling factual halluci-

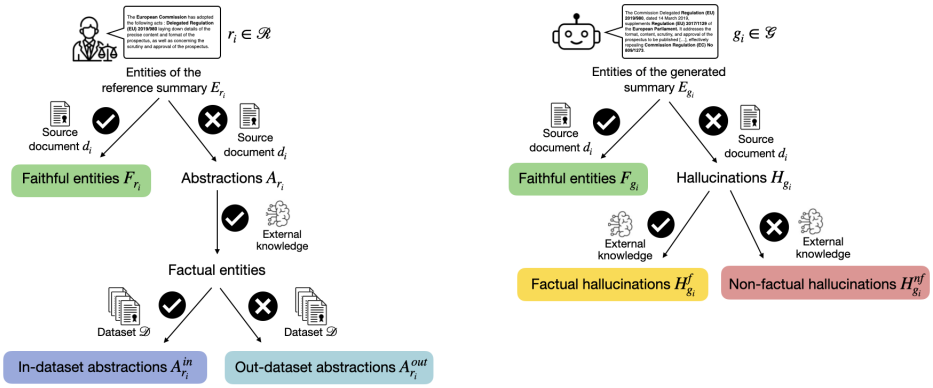

Figure 2: Taxonomy of entities in the reference summary (left) and generated summary (right). Examples of entities from each set are presented on Figure 1, using the same colors (green for faithful entities, blue for abstractions, yellow and red respectively for factual and non factual hallucinations).

nations to make them as close as possible to the abstractions contained in reference summaries. To ascertain the factuality of hallucinations we propose to exploit the dataset of the source documents as knowledge base through the construction of an entity graph. As *out-dataset* abstractions cannot be found from this graph, we simplify the problem as follows: we aim at controlling factual hallucinations to make them as close as possible to the *in-dataset* abstractions contained in reference summaries.

Other notions are introduced in the literature at the relationship/information level, such as extrinsic or extrinsic hallucinations (Maynez et al., 2020; Ji et al., 2023), but they are out of the scope of this paper.

2.2 Problem Formulation

Let us consider a dataset of the legal domain composed of $\mathcal{D}$ a set of source documents d_i and $\mathcal{R}$ a set of reference summaries r_i such as r_i is the reference summary associated to d_i. Let E_{d_i} and E_{r_i} be respectively the set of entities in d_i and r_i. $E_{r_i} = F_{r_i} \cup A_{r_i}$ where F_{r_i} is the set of faithful entities in r_i ($F_{r_i} = E_{r_i} \cap E_{d_i}$) and A_{r_i} is the set of abstractive entities in r_i. $A_{r_i} = A_{r_i}^{in} \cup A_{r_i}^{out}$ where $A_{r_i}^{in}$ is the set of abstractive entities that can be found in $\mathcal{D}$ and $A_{r_i}^{out}$ the set of abstractive entities that cannot be found in $\mathcal{D}$.

Our aim is to produce for each d_i in $\mathcal{D}$ a generated summary g_i such as E_{g_i} (the set of entities in g_i) is as close as possible to E_{r_i}. In a realistic world, we simplify the problem as: our aim is to produce for each d_i in $\mathcal{D}$ a generated summary g_i such as E_{g_i} is as close as possible to $E_{r_i} \setminus A_{r_i}^{out}$. In other words, we aim at having: $F_{g_i} = F_{r_i}$ and $H_{g_i}^{f} = A_{r_i}^{in}$ and $H_{g_i}^{nf} = \emptyset$, where $H_{g_i}^{f}$ and $H_{g_i}^{nf}$ are respectively the factual and non-factual hallucinations of g_i.

Notations are summarized in Appendix B.

3 Related Work

Hallucination Detection Several works have shown that abstractive summarization models suffer from generation of hallucinated entities (Cao et al., 2022; Nan et al., 2021b; Kryscinski et al., 2019). Cao et al. (2018) showed that 30% of the named entities generated by the BART model are hallucinated in the XSum dataset (Narayan et al., 2018). Conventional summary evaluation metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), or BERTScore (Zhang et al., 2019) consider lexical and semantic overlap but fail to assess factuality and faithfulness since they are not sensitive to the hallucinated content. Thus, several approaches have been developed to measure the hallucination rate in the text generated by the model. They rely on question-answering methods (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021a; Fabbri et al., 2022), text-entailment (Falke et al., 2019; Goyal and Durrett, 2020; Kryściński et al., 2020), LLMs (Manakul et al., 2023; Friel and Sanyal, 2023; Liu et al., 2023; Mündler et al., 2023), or Information Extraction (Nan et al., 2021b; Goodrich et al., 2019; Dhingra et al., 2019). Given that legal texts contain many references to laws, which can be considered as entities of interest, we use the latter to evaluate generated summaries, personalizing the metrics presented by Nan et al. (2021b).

Entity-Centric Summarization Several approaches focused on directing the model to generate named entities at the intersection of the source

document and the reference summary. Zheng et al. (2020) extract key phrases at the intersection of the document and the reference summary and train LSTM models with a concatenation of the document and the extracted phrases. Liu and Chen (2021); Mao et al. (2020) select named entities instead. Fan et al. (2018); He et al. (2022) propose generated summaries controlled by several aspects, such as the summary subject, keywords or named entities, by concatenating them with the document or by adding them as a prompt in order to guide the model. Xiao and Carenini (2023) on the other hand, use a Global Relevance module to select important entities from the source document and introduce them to the decoder. Zhang et al. (2022) proceed differently, by calculating the coverage rate of entities (Nan et al., 2021b) and introducing a control token based on this rate to guide the generation. In this paper, with the objective of getting as close as possible to the writing style of the legal domain, we investigate the possibility of fine-tuning the model to introduce entities that do not exist in the source document. We concatenate with the source document a set of entities from the reference summary or entities from the dataset related to the document.

Legal Domain The legal domain presents unique challenges due to its complex and highly specialized language, which explains the multitude of dedicated approaches in the literature (Sansone and Sperlí, 2022). Among neural network-based approaches, Chalkidis et al. (2020) proposed LEGAL-BERT. Regarding summarization, models do not only need to summarize content accurately but also to maintain the integrity of legal entities and terminology. Legal-Pegasus[2] is pretrained on legal documents, whereas Legal-LED[3] is specifically fine-tuned on legal data. At the same time, several datasets in legal domain have been created to fine-tune language models for automatic summarization. Among them, one can cite BillSum (Kornilova and Eidelman, 2019), a dataset for summarization of US Congressional and California state bills; EUR-Lex-Sum (Aumiller et al., 2022), a dataset based on manually curated document summaries of legal acts from the European Union law platform (EUR-Lex); and Multi-LexSum (Shen et al., 2022), a collection of expert-authored summaries drawn from ongoing CRLC (*Civil Rights Litigation Clearinghouse*) writing. This diversity in corpus types highlights the potential adaptability of summarization models to various legal contexts, ranging from legislative documents to civil rights summaries. Most of the proposed datasets and fine-tuned models are extractive and available in English, with very few existing in French. The approach we propose in this paper is abstractive, and evaluated on two different languages, French and English.

4 Legal Summary Specificities

Our aim here is to study the characteristics of reference summaries in terms of entities of interest as well as abstractions. To this end, we analyzed two legal datasets for summarization: EUR-Lex-Sum and a private corpus of French legal news.

4.1 Datasets Description

The EUR-Lex-Sum dataset (Aumiller et al., 2022) contains 1,504 legal acts documents of 12,000 tokens in average, retrieved using web scraping from the European Union law platform website[4]. Associated reference summaries are composed of 800 tokens in average. The dataset is divided into training and test sets, containing 1,316 and 188 documents respectively.

We also used a private business dataset, named Légibase composed of 8,485 legal and regulatory documents of French local authorities and public administrations[5]. Each document includes (a) a title, (b) a text (body/content), (c) a set of associated metadata, notably the document topic, and (d) a manually-produced summary. Each information is written by experts in the field, whose aim is to keep the legal news up to date. These experts focus on what is new and evolving in French law, often starting with a presentation of the background of the law, the code and its various clauses, followed by a presentation of the evolution. Each summary sets out the general context of the document content and the news described within it. Documents and reference summaries are respectively composed of 8,325 and 1,110 tokens in average. The train and test sets are composed of 9,353 and 851 documents.

4.2 Entities of Interest

We considered 4 types of entities: Organizations, Locations, Persons and Law entities (e.g., Regulation 2019/279, Decision 1147/366, Directive 2016/879 in English, or loi n° 93-22, article 57 du

[2] https://huggingface.co/nsi319/legal-pegasus

[3] https://huggingface.co/nsi319/legal-led-base-16384

[4] https://eur-lex.europa.eu

[5] https://www.legibase.fr/ An example document and its associated reference summary are presented in Appendix C.

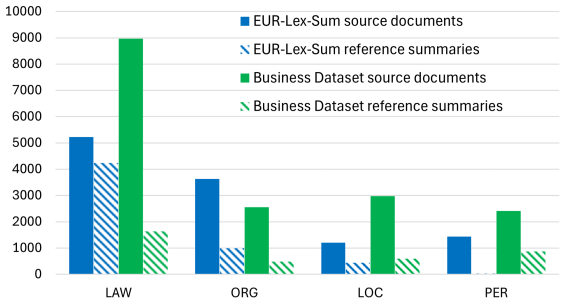

Figure 3: Distribution of entity types across source documents and reference summaries in the two datasets: EUR-Lex-Sum and Légibase. Distinct occurrences are considered.

Code civil, décret n° 89-11 in French). As Law entities are written with very specific and repetitive patterns, we used a set of regex to extract them from the text. All other entities were extracted using a NER model fine-tuned on the Indian Court Judgements dataset[6].

Figure 3 presents the distribution of the extracted entities across the datasets. Law entities are, not surprisingly, by far the most present, both in source documents and in reference summaries. Another interesting result is that Persons are barely present in the reference summaries of EUR-Lex-Sum.

4.3 Abtractiveness of Reference Summaries

Table 1 shows an analysis of faithful entities (F_{r_i}), as well as in-dataset and out-dataset abstractive entities ($A^{in}_{r_i}$ and $A^{out}_{r_i}$). We observe that a large proportion of the entities present in the summaries are abstractive (66% and 68% for EUR-Lex-Sum and Légibase respectively). This confirms that legal experts rely heavily on their external knowledge when writing summaries, in order to contextualize the main entities of the source document and provide a history of them. This external knowledge may or may not be found in other documents in the dataset, as we can see from columns $A^{in}_{r_i}$ and $A^{out}_{r_i}$.

5 Selection of abstractive entities in the source dataset (Q1)

Our objective is to generate a summary as similar as possible to the reference summary, which therefore contains as many entities as possible from the set of abstractive entities. We propose a method based on an entity graph constructed from the dataset, to identify for each document, the set of entities that most closely matches the abstractive entities of the corresponding reference summary.

[6] https://huggingface.co/opennyaiorg/en_legal_ner_trf

5.1 Graph Construction

We constructed two different graphs: G_{D_T} and G_{D_I}, containing respectively the sets of entities of our training and testing datasets. G_{D_T} (see Figure 4 for an example) is constructed as follows: (i) the nodes set is composed of documents $d_i \in D_T$, entities E_{d_i} of each d_i and the subject of d_i[7]; (ii) edges represent 1) the similarity between documents, evaluated with a cosine similarity between embeddings of documents[8], 2) the presence of an entity in a document, and 3) a link between a document and a subject. The construction of G_{D_I} is similar.

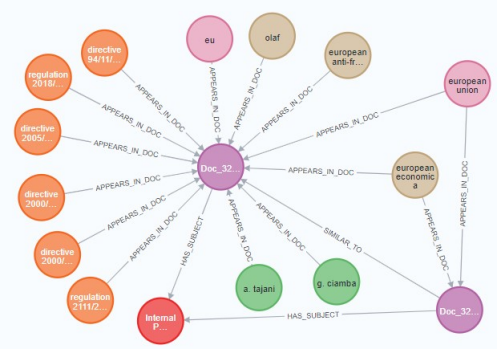

Figure 4: Extract of G_{D_T} for the EUR-LEx-Sum dataset. Purple nodes are documents, red ones are Subjects, while orange, brown and green ones are respectively Law, Organization and Person entities. The two documents are semantically close and share the same subject 'Internal Market'. They also share the same entity named 'European Economic Committee'.

5.2 Entity Selection Strategies

We defined different strategies $S_{i,j}$ to select entities from the created graphs. The idea is, for each document d_i, to retrieve entities from related documents: for $S_{i,1}$ we extract entities connected by a maximum of 1, 2, or 3 hops to d_i; for $S_{i,2}$ we extract entities of documents sharing the same subject than d_i; for $S_{i,3}$ we extract entities of documents semantically similar to d_i; and finally for $S_{i,4}$ we extract entities of documents sharing the same subject than d_i and being semantically similar to d_i. Note that entities of d_i are systematically excluded from the resulting sets.

5.3 Experiments and Results

In order to determine which selection strategy would be most effective, we measured the inter-

[7] The subjects of EUR-LEX-Sum documents are obtained by scrapping the associated web pages, while those of Légibase are given with the documents.

[8] https://huggingface.co/facebook/bart-large-mnli

	EUR-Lex-Sum						Légibase					
	Total			Avg			Total			Avg		
Entity Type	$\#F_{r_i}$	$\#A^{in}_{r_i}$	$\#A^{out}_{r_i}$	$\#F_{r_i}$	$\#A^{in}_{r_i}$	$\#A^{out}_{r_i}$	$\#F_{r_i}$	$\#A^{in}_{r_i}$	$\#A^{out}_{r_i}$	$\#F_{r_i}$	$\#A^{in}_{r_i}$	$\#A^{out}_{r_i}$
Law	1,978	1,108	1,906	3.72	2.37	2.37	667	964	486	1.10	1.12	1.06
Org	340	309	457	1.67	1.71	1.32	222	266	140	1.04	1.05	1.01
Loc	233	247	95	1.68	2.14	1.06	351	245	89	1.12	1.11	1.01
Per	1	0	26	0	0	0	253	621	368	1.04	1.13	1.05

Table 1: Total number of faithful and abstractive entities in the reference summaries of EUR-Lex-Sum and Légibase. The average number of entities per reference summary is also reported.

		EUR-Lex-Sum			Légibase			
Strategy		Law	Org	Loc	Law	Org	Loc	Per
$S_{i,1}$	R	0.36	0.39	0.63	0.47	0.50	0.66	0.39
	P	0.21	0.08	0.18	0.29	0.12	0.22	0.11
$S_{i,2}$	R	0.33	0.31	0.56	0.44	0.47	0.61	0.40
	P	0.19	0.07	0.16	0.25	0.18	0.24	0.15
$S_{i,3}$	R	0.21	0.18	0.27	0.25	0.20	0.18	0.12
	P	0.13	0.05	0.09	0.15	0.09	0.10	0.06
$S_{i,4}$	R	0.08	0.07	0.16	0.06	0.10	0.07	0.04
	P	0.07	0.02	0.06	0.04	0.08	0.09	0.03

Table 2: Recall and precision of the intersection between the abstractive entities of the reference summaries and entities obtained by each of our retrieval strategies.

Experiment	Training	Inference
Oracle FH	$d_i \cup A^{in}_{r_i}, d_i \in D_T$	$d_i \cup A^{in}_{r_i}, d_i \in D_I$
Setting 0	$d_i \in D_T$	$d_i \in D_I$
Setting 1	$d_i \cup A^{in}_{r_i}, d_i \in D_T$	$d_i \cup E_{s_{i,1}}, d_i \in D_I$
Setting 2	$d_i \cup A^{in}_{r_i} \cup E_{s_{i,1}}, d_i \in D_T$	$d_i \cup E_{s_{i,1}}, d_i \in D_I$
Setting 3	$d_i \cup E_{s_{i,1}}, d_i \in D_T$	$d_i \cup E_{s_{i,1}}, d_i \in D_I$

Table 3: Settings used for experiments

section of their respective sets with the set of abstractions in the reference summaries. Results are shown in Table 2. None of the strategies strongly maximize the intersection with abstractions. The intersection of $S_{i,4}$ is quite low, particularly for the Organizations and Locations, leading us to conclude that abstractive entities are widely dispersed in the corpus, and not necessarily linked to d_i. We chose $S_{i,1}$ as the best compromise for the rest of our experiments.

6 Entity-driven summarization (Q2)

Given the set of entities retrieved from the graph, we experimented various fine-tuning settings, by varying the input provided to the model. The quality of the generated summary was then evaluated using entity-oriented metrics.

6.1 Experimental Setup

Models As documents in the EUR-Lex-Sum dataset are very long (12k tokens on average), we evaluated our approach with a classic Long Encoder-Decoder model (LED)[9] as well as with BART (Lewis et al., 2020)[10] using a Long Document Transformer framework (Beltagy et al., 2020) with 4 epochs. Checkpoints were retrieved from the HuggingFace platform. Both models led to similar conclusions in terms of behaviour. We decided to report only LED results in the paper for space limitation reasons. BART results can be found in Appendix D as supplementary material[11].

Fine-Tuning Format We prepended the set of entities to the source document, using the special token **[ENTITYSET]** to introduce the entities and separated them with the token |. The source document was introduced using the special token **[DOCUMENT]**. For EUR-Lex-Sum, we considered the 3 entity types: Law, Organization and Location. Person entities were removed since they are not present in the reference summaries (see Figure 3). For Légibase, the 4 entities were used.

Fine-Tuning Settings We considered 3 different settings to fine-tune the model (settings notations are summarized in Table 3), with the objective to evaluate how the entities provided during fine-tuning affect the model's performance during inference, and more broadly, to assess whether the model's behavior varies based on the set of entities used.

- **Setting 1:** During the training phase, we provide the model with the source document d_i and the $A^{in}_{r_i}$ set. During inference, we concatenate the source document and 20 entities of $E_{s_{i,1}}$, retrieved with $S_{i,1}$. These entities are chosen randomly, maintaining a balance between the different entity types (law, org, loc, and per).
- **Setting 2:** During training, the model receives the source document, prepended by 20 entities from $A^{in}_{r_i}$ completed by some randomly chosen entities of $E_{S_{i,1}}$. For inference, only the source document and 20 entities of $E_{S_{i,1}}$ are provided.

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

[10] https://huggingface.co/hyesunyun/update-summarization-bart-large-longformer

[11] We did not use the Legal-Pegasus model as it is limited to 1024 tokens. The Legal-LED model has also been discarded due to non-satisfactory results.

Systems	EUR-Lex-Sum								Légibase							
	F ↑		C ↑		FH ↑		$\overline{\text{FH}}$ ↓		F ↑		C ↑		FH ↑		$\overline{\text{FH}}$ ↓	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
ECC	**0.33**	**0.21**	0.09	0.34	0.20	0.07	-	0.70	0.47	**0.06**	**0.30**	0.16	**0.11**	**0.09**	-	0.82
CTRLSum	0.29	0.19	0.09	0.31	0.18	0.06	-	0.71	0.44	0.05	0.27	0.18	0.10	0.07	-	0.85
Mixtral	0.25	0.12	0.04	0.19	0.10	0.02	-	0.76	0.36	0.02	0.20	0.11	0.04	0.00	-	0.97
Oracle FH	*0.33*	*0.21*	*0.12*	*0.35*	*0.38*	*0.11*	-	*0.60*	*0.59*	*0.05*	*0.30*	*0.44*	*0.50*	*0.07*	-	*0.80*
Setting 0	0.29	0.17	0.05	0.24	0.19	0.04	-	0.73	0.53	0.04	0.25	0.12	0.09	0.04	-	0.94
Setting 1	0.30	0.15	0.08	0.25	0.23	0.05	-	0.78	0.54	0.02	0.22	0.17	0.10	0.07	-	0.93
Setting 2	**0.33**	0.20	**0.10**	0.34	**0.30**	**0.08**	-	**0.68**	0.63	0.05	0.28	**0.36**	0.10	0.01	-	**0.81**
Setting 3	0.32	0.20	**0.10**	**0.35**	0.29	**0.08**	-	**0.68**	**0.64**	0.05	0.29	**0.36**	0.09	0.01	-	**0.81**

Table 4: Results on EUR-Lex-Sum and Légibase, using Entity-Level metrics : Coverage (C), Faithfulness (F), Factual hallucination (FH) and Non-Factual hallucination rate ($\overline{\text{FH}}$). Results in bold and underlined are respectively the best and second best results for a given metric.

- Setting 3: During training, the model receives only 20 entities of the $E_{S_{i,1}}$ set and the source document. For inference, the same elements are provided.
The size of the entities set (20 entities) added to each document was chosen on the basis of preliminary experiments on EUR-lex-Sum.

6.2 Entity-Level Metrics

To evaluate the generated summary g_i in terms of faithful and hallucinated entities according to the source document d_i and reference summary r_i, we extended the metrics of Nan et al. (2021b):

- Coverage rate (Recall, Precision) in relation to the source document:

$$R(C) = \frac{\#(E_{g_i} \cap E_{d_i})}{\#E_{d_i}}, \quad P(C) = \frac{\#(E_{g_i} \cap E_{d_i})}{\#E_{g_i}}$$

- Faithfulness rate:

$$R(F) = \frac{\#(E_{g_i} \cap F_{r_i})}{\#F_{r_i}}, \quad P(F) = \frac{\#(E_{g_i} \cap F_{r_i})}{\#E_{g_i}}$$

- Factual hallucination rate (in relation to the abstractive entities of the reference summary):

$$R(FH) = \frac{\#(E_{g_i} \cap A^{in}_{r_i})}{\#A^{in}_{r_i}}, \quad P(FH) = \frac{\#(E_{g_i} \cap A^{in}_{r_i})}{\#E_{g_i}}$$

- Non-factual hallucination rate:

$$P(\overline{FH}) = \frac{\#(E_{g_i} - E_{r_i})}{\#E_{g_i}}$$

6.3 Baselines

We used the following state-of-the-art baselines for comparison with our approach: (1) **ECC** (Zhang et al., 2022) is a model training approach for improving the factuality of generated summaries based on the coverage metric; (2) **CTRLSum** (He et al., 2022) is a training approach to guide the model to include a set of pre-selected keywords and entities in the generated summary; and (3) **Mixtral 8 * 7b** (Jiang et al., 2024) is a LLM, chosen to compare the experiments conducted with PLMs against a LLM without fine-tuning. Additionally, we also report (i) **Setting 0**, in which we run a basic fine-tuning of the model, where only the source document is provided as input, for both training and inference. (ii) **Oracle FH**, in which the set of abstractive entities $A^{in}_{r_i}$ is given at training and inference time. Resulting summaries can be considered as Oracle summaries, i.e., summaries maximizing the R(FH) and P(FH) metrics. They are the best possible summaries that can be obtained by our approach considering factual hallucinations.

6.4 Results

Results are presented in Table 4. The high results obtained by Oracle FH summaries show that guiding the summary generation with entities allows to obtain better summaries in terms of faithfulness, factual and non-factual hallucinations. This is also true whatever the considered setting (1, 2 or 3), since all settings are better than setting 0, in which no entities are included. Another conclusion is that the quality of entities given as input of the model strongly impacts its performances (comparison of settings 1, 2 and 3). Giving only the set of abstractive entities to the model during the training phase (setting 1) does not allow to reach the best performances. Indeed, if some noise is added during training (settings 2 and 3), and consequently if the training is closer to the inference configuration, results are improved. Regarding SOTA baselines, we can see the ECC and CTRLSum models outperform Mixtral, probably because they are entity-oriented, unlike the latter. ECC is the most competitive baseline, with particularly good results regarding faithfulness and coverage (F and C metrics). This is not surprising since the approach is optimized on the F metric. Finally, we observe

	Wrt the Source Document					Wrt the Reference summary								
	EUR-Lex-Sum			Légibase		EUR-Lex-Sum					Légibase			
Systems	Bert-S ↑	Q-E ↑	F-CC ↑	Bert-S ↑	F-CC ↑	Bert-S ↑	Q-E ↑	F-CC ↑	R-L ↑	B ↑	Bert-S ↑	F-CC ↑	R-L ↑	B ↑
Reference Summaries	0.85	0.35	0.20	0.84	0.30	-	-	-	-	-	-	-	-	-
ECC	0.86	0.33	0.35	0.84	0.37	0.84	**0.38**	0.36	0.22	0.40	0.85	0.34	0.19	0.31
CTRLSum	0.83	0.36	0.31	0.83	0.35	0.85	0.33	0.34	0.21	0.40	0.84	0.34	0.18	0.31
Mixtral	0.83	0.29	0.22	0.82	0.32	0.84	0.30	0.29	0.18	0.37	0.83	0.27	0.14	0.26
Oracle FH	*0.86*	*0.36*	*0.33*	*0.84*	*0.39*	*0.86*	*0.39*	*0.39*	*0.26*	*0.44*	*0.87*	*0.39*	*0.20*	*0.36*
Setting 0	0.83	0.27	0.24	0.84	0.29	0.83	0.28	0.27	0.20	0.39	0.83	0.28	0.16	0.29
Setting 1	0.85	0.33	0.31	0.83	0.36	0.86	0.34	0.34	0.23	0.39	0.84	0.36	0.16	0.33
Setting 2	**0.88**	**0.37**	**0.39**	**0.86**	0.38	**0.87**	0.35	0.39	0.21	0.41	**0.86**	**0.40**	**0.20**	**0.35**
Setting 3	0.87	0.35	**0.39**	0.85	**0.41**	0.86	0.37	**0.40**	**0.24**	**0.42**	**0.86**	0.39	0.19	**0.35**

Table 5: Performance of the different systems (baselines, Oracle FH, settings) evaluated using metrics Bert-Score (reported in the table as Bert-S), QuestEval (Q-E), FactCC (F-CC), Rouge-L (R-L) and Bleu (B). QuestEval has not been reported for Légibase as it is not available in French.

similar performance gains on both datasets, despite their language differences. One notable exception is the FH metric on Légibase, which decreases with settings 2 and 3 (compared to setting 1). Possible explanations include either a difference in behavior for French language, or that the number of entities provided as input no longer allows the model to hallucinate factually. This will be the subject of future experiments on the French language.

7 Evaluating the Overall Quality of the Generated Summaries (Q3)

The experiments presented in previous sections showed that injecting entities into the model allows us to improve entity-centric metrics. In this section, we aim at assessing the overall quality of the generated summaries.

7.1 Contextual Metrics

To evaluate the quality of summaries at token-level, we report the well-known **ROUGE-L** (Lin, 2004) and **BLEU** (Papineni et al., 2002) metrics. We also use 3 contextual metrics: **Bert-Score** (Zhang et al., 2019) which compares contextual embeddings of sentences, **FactCC** (Kryściński et al., 2020) which assesses the factual consistency of the generated summaries using textual entailment, and **QuestEval** (Scialom et al., 2021b), a question generation and answering metric. As reference summaries are slightly different from source documents in terms of content, all metrics are evaluated both against source documents and reference summaries.

7.2 Discussion and results

Results are reported in Table 5. ECC remains the strongest baseline, but is outperformed by our approach on all contextual metrics. This shows that our approach, although entity-oriented, is able to provide high-quality summaries. Whether the generated summaries are compared with the documents or with the reference summaries, we observed that settings 2 and 3 outperform all other systems. This is also true compared to Oracle FH. Adding noise to the training data thus allows the model to better generalize. This confirms our conclusions from Table 4. A last conclusion concerns the metrics evaluated wrt the source document. Reference summaries, when evaluated wrt the source document obtain weak results (first line of Table 5). Metrics evaluated against source documents are thus not relevant to compute on our datasets, since reference summaries of the latter are quite different from the source documents.

All these results motivate us to further explore the approach of guiding the summary generation model with additional entity-oriented information.

8 Case Study

In this section, we present some examples of generated summaries using baselines and our model, in order to provide a qualitative analysis. Those examples are showed in Table 5 for EUR-Lex-Sum. Examples of Légibase are present in Appendix C.

These examples illustrate that all approaches generate factual entities (in green), but also hallucinations (in red). The latter are often due to errors in Regulations or Laws numerical identification code. For instance, Regulation 2017/1939 is generated instead of Regulation 2017/1938 (Figure 5–Our generated summary). The same can be observed with Regulation 525/2013 instead of 347/2013 (ECC Generated Summary). Another type of hallucinations is also present at the level of entity headings: Decision (EU) No 994/2010 is prompted in Mistral Generated Summary instead of Regulation (EU) No 994/2010. Moreover, although

Source Document : Official Journal of the European Union. L 280/1. **REGULATION 2017/1938** OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2017 concerning measures to safeguard the security of gas supply and repealing **Regulation No 994/2010** (Text with **EEA** relevance). THE **EUROPEAN PARLIAMENT** AND THE **COUNCIL OF THE EUROPEAN UNION**, Having regard to the Treaty on the Functioning of the European Union **(TFEU)** [...]. Those objectives should be achieved through the most cost-effective measures and in such a way that gas markets are not distorted.(4) Union law, in particular **Directive 2009/72/EC** of the European Parliament and of the Council (3), **Directive 2009/73/EC** of the European Parliament and of the Council (4), **Regulation (EC) No 713/2009** of the European Parliament and of the Council (5), **Regulation No 714/2009** of the European Parliament and of the Council (6), **Regulation No 715/2009** of the European Parliament and of the Council (7) and **Regulation (EU)No 994/2010** of the European Parliament and of the Council [...] [12]

Reference Summary : Gas supply security in the EU Gas supply security in the EU SUMMARY OF: **Regulation 2017/1938** aiming to safeguard an uninterrupted supply of gas throughout the EU WHAT IS THE AIM OF THIS REGULATION? It aims to strengthen EU energy security by helping prevent potential supply disruptions, and to respond to them when they occur, to ensure that households and other vulnerable consumers are always supplied. [...] MAIN DOCUMENT **Regulation (EU) 2017/1938** of the European Parliament and of the Council of 25 October 2017 concerning measures to safeguard the security of gas supply and repealing **Regulation (EU) No 994/2010** (OJ L 280, 28.10.2017, pp. 1-56) RELATED DOCUMENT Communication from the Commission the European Parliament, the Council, the European Economic and Social Committee, the Committee of the Regions and the European Investment Bank – Framework Strategy for a Resilient Energy Union with a Forward-Looking Climate Change Policy (COM(2015) 80 final, 25.2.2015) **Regulation (EU) No 347/2013** of the European Parliament and of the Council of 17 April 2013 on guidelines for trans-European energy infrastructure and repealing **Decision No 1364/2006/EC** and amending **Regulation (EC) No 713/2009**. The **European Network of Transmission System Operators (ENTSOG)** will carry out a 4-yearly EU-wide simulation of gas supply and infrastructure disruption scenarios [...]

Baseline - ECC Generated Summary : Protecting the gas supply of the **European Union** SUMMARY OF: **Regulation (EU) 2017/1938** — measures to safeguard the security of gas supply WHAT IS THE AIM OF THE REGULATION? It aims to ensure that all the necessary measures are taken to safeguard gas supply across the **EU**, in particular to protected customers, in the event of difficult climatic conditions or disruptions of the supply. KEY POINTS The regulation applies to all EU countries except those that are not party to the Treaty on the Functioning of the European Union **(TFEU)**. [...] FROM WHEN DOES THE REGOLATION APPLY? It has applied since 28 October 2017. BACKGROUND The **EU** has a gas market that operates smoothly. It operates smoothly and is the best guarantee of its security, to reduce the exposure of individual EU countries to the harmful effects of disruptions of gas. For more information, see: 'Protect the **EU** gas market' on the **European Commission**'s website' (European Commission). MAIN DOCUMENT **Regulation(EU) (2017/1938)** and **Regulation (EU) No 525/2013** of the Council of 23 and 24 october 2014 concerning measures to protect the gas supplies of the Union (OJ L 280, 28.10.2017, pp. 1-2).

Baseline - Mistral Generated Summary : **Regulation (EU) 2017/1938** — measures to safeguard the security of gas supply and repealing **Decision (EU) No 994/2010** WHAT IS THE AIM OF THE REGULATION? It aims to ensure that all the necessary measures are taken to ensure the security and stability of the gas market in the **EU**, in particular to protected customers. The regulation applies to gas supply to all **EU** countries. The legislation applies to all gas supply from the gas grid to the neighbouring **European Union** country. It does not apply to natural gas (gas) or electricity generation. It applies only to gas supplied from the single largest supplier. In the event of a major disruption of the supply of gas to a country, it must be taken into account in the decision-making process. To protect the security, the regulation requires that gas supply be shared by all **EU** governments, including the **European Council** and The **European Commission** [...]

Generated Summary (Setting 2) : Guaranteeing gas supply in the event of a major disruption Guarantaining gas supply within the **EU** SUMMARY OF: **Regulation (EU) 2017/1938** — measures to safeguard the security of gas supply and repealing **Regulation(EU) No 994/2010** WHAT IS THE AIM OF THE REGULATION? It aims to ensure that EU countries' national authorities and the **European Commission** have a coordinated [...] MAIN DOCUMENT **Regulation EU/2017/1939** of the **European Parliament and of the Council** of 25 October. The regulation forms part of the Energy Union package, which aims to make energy secure, affordable and sustainable through closer cooperation between EU countries. Last updated: 22.04.2022(2.11) The date of expunishment of the Treaty on the Functioning of the Union **(TFEU)** and its subsequent amendments to Articles 1 and 2 and of **Regulation (EC) No 347/2013** of the **Parliament**, **Council**, **European Commission**, **European Network of Transmission System Operators for Gas (ENTSog)**.

Figure 5: Example of a legal document from EUR-Lex-Sum with its reference summary (excerpts) produced by experts, as well as 3 generated summaries (Setting 2, ECC and Mixtral). Entities are marked in bold. In the reference summary, faithful entities are colored in green. Abstractive entities of the reference summary are colored in dark blue when present in the dataset (in-dataset abstractions) whereas the light blue one is an out-dataset abstraction. In the generated summary, faithful entities are also colored in green. The entity colored in red is a non-factual hallucination whereas the yellow one is a factual hallucination.

ECC is our strongest baseline, it still generates similar hallucinations as the other models.

Finally, we can see that guiding the generation model to generate abstrative entities is successful on these examples. Our model is the only one to generate factual hallucinations (in yellow), which are themselves present as abstractive entities (in light blue) in the reference summaries. Moreover, focusing on entities does not affect the overall quality of the generated summary in terms of readibility. This supports the analysis conducted in Section 7.

9 Conclusion

In this paper, we showed the specificities of legal corpora in terms of abstraction in reference summaries produced by legal experts, which requires an adaptation of state-of-the-art summarization approaches. We introduced a new method for selecting factual hallucinations on an entity-level based on a graph created from the source dataset, and proposed different training and inference settings for guiding the summary generation model with these entities. Our experiments on both French and English datasets show that this method significantly reduces non-factual entity hallucinations, increases coverage metrics and factual hallucinations while improving the overall quality of the generated summaries. To the best of our knowledge this is the first entity-centric summarization method using only the original dataset as a way to control factual and non-factual hallucinations.

10 Limitations

Although promising, our approach has some limitations. First, we were unable to provide a human evaluation of our approach. Since legal language is particularly complex, collaboration with legal experts is necessary. We chose to use Légibase to work directly with the experts who wrote the articles and summaries. For time and organizational reasons, we were not able to evaluate a significant number of generated summaries in order to present the results of this evaluation, but this is currently in progress.

A second limitation of our approach is that it focuses exclusively on entity-level hallucinations. Although we showed that the overall quality of our generated summaries is good, we could extend our work to relation-level hallucinations, using the same framework with facts (subject, predicate, object) to ensure that the introduced entities and terms are correctly placed (Huang et al., 2020). Additionally, we aim to expand our hallucination metrics to distinguish between intrinsic and extrinsic hallucinations at relation/information level. Intrinsic hallucination can be defined as terms or concepts from the source information that are misrepresented in the summary, making them unfaithful to the original source. On the other hand, extrinsic ones can be defined as information that cannot be verified from the source content. Another future work is to extend our external knowledge to ontologies and knowledge graphs in order to retrieve more facts from multiple sources (Wang et al., 2022; Dziri et al., 2021), and not only from the datasets used in this study.

A third limitation concerns the strategies used to find abstractive entities in the source dataset. The results with these heuristics are promising but showed that abstractive entities are disseminated in all the datasets. Graph learning strategies will probably be very useful (Ribeiro et al., 2022).

Finally, another limitation is the non-use of current large language models (LLMs). We compared our approach with the Mixtral 8 * 7b (Jiang et al., 2024), but used it with a basic prompt, without fine-tuning. Therefore, the next natural step of this research work is to fine-tune LLMs such as LLaMA 2 (Touvron et al., 2023) and Mixtral.

11 Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014740 made by GENCI.

References

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. Eur-lex-sum: A multi-and cross-lingual dataset for long-form summarization in the legal domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Meng Cao, Yue Dong, and Jackie Chi Kit 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.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

Ethan Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

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.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

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

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *ArXiv*, abs/2310.18344.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. Ctrlsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

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

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

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 (ACL'20)*, page 7871–7880.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106.

Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. 2022. Faithful abstractive summarization via fact-aware consistency-constrained transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1410–1419, New York, NY, USA. Association for Computing Machinery.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv e-prints*, pages arXiv–2010.

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.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324, pages 141–147.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Feng Nan, Cicero dos Santos, Henghui Zhu, Patrick Ng, Kathleen Mckeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021a. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021b. Entity-level factual consistency of abstractive text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021a. Questeval: Summarization asks for fact-based evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang,

and Patrick Gallinari. 2021b. Questeval: Summarization asks for fact-based evaluation. In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: real-world summaries of civil rights lawsuits at multiple granularities. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Guan Wang, Weihua Li, Edmund Lai, and Jianhua Jiang. 2022. Katsum: Knowledge-aware abstractive text summarization. *arXiv preprint arXiv:2212.03371*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wen Xiao and Giuseppe Carenini. 2023. Entity-based spancopy for abstractive summarization to improve the factual consistency. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249.

Changmeng Zheng, Yi Cai, Guanjie Zhang, and Qing Li. 2020. Controllable abstractive sentence summarization with guiding entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5668–5678.

A Scientific Artifacts

This work would not be possible without many open-source scientific artifacts, including pytorch (Paszke et al., 2019), transformers (Wolf et al., 2020), Neo4j (Miller, 2013), Numpy (Harris et al., 2020), nltk (Bird et al., 2009), the language models used for training and inference: LED (Beltagy et al., 2020) and BART (Lewis et al., 2020), the different baselines used: ECC (Zhang et al., 2022), CTRLSum (He et al., 2022), Mixtral (Jiang et al., 2024).

B Notations

Notation	Description
$\mathcal{D}_T$	Source documents of the train split
$\mathcal{D}_I$	Source documents of the inference split
$\mathcal{D}$	The total set of source documents such as $\mathcal{D} = \mathcal{D}_T \cup \mathcal{D}_I$
$\mathcal{R}$	Reference summaries
$\mathcal{G}$	Generated summaries
d_i	Source document such as $d_i \in \mathcal{D}$
r_i	Reference summary of d_i such as $r_i \in \mathcal{R}$
g_i	Generated summary of d_i such as $g_i \in \mathcal{G}$
E_{d_i}	Entities of d_i
E_{r_i}	Entities of r_i
E_{g_i}	Entities of g_i
F_{r_i}	Faithful entities of r_i, such as $F_{r_i} \subseteq E_{r_i}$ and $F_{r_i} = E_{r_i} \cap E_{d_i}$
$A^{in}_{r_i}$	Abstractive entities of the reference summary r_i that can be found in the dataset
$A^{out}_{r_i}$	Abstractive entities of the reference summary r_i that cannot be found in the dataset
A_{r_i}	The overall set of the abstractive entities of r_i such as $A_{r_i} = A^{in}_{r_i} \cup A^{out}_{r_i}$ and $A^{in}_{r_i} \cap A^{out}_{r_i} = \emptyset$
$H^f_{g_i}$	Factual hallucinations in g_i
$H^{nf}_{g_i}$	Non factual hallucinations in g_i
H_{g_i}	The overall set of hallucinations, such as $H_{g_i} = H^f_{g_i} \cup H^{nf}_{g_i}$ and $H^f_{g_i} \cap H^{nf}_{g_i} = \emptyset$
$E_{S_{i,j}}$	Entities selected by strategy $S_{i,j}$ such as i refers to the couple of source document, reference summary (d_i, r_i) and $j \in [1, 4]$ (cf. table 2)
G_{D_T}	Entity graph of the train split
G_{D_I}	Entity graph of the inference split

Table 6: Key Notation Summary

C Légibase Example

Source Document : Affectation d'un lieu autre que la mairie pour la célébration des mariages Disposant parfois de salles trop petites pour une célébration dans de bonnes conditions, ou de lieux prestigieux, les maires revendiquaient la possibilité de célébrer des mariages dans un lieu autre que celui de la mairie. Depuis le 4 mars 2017, ils peuvent donc affecter tout autre bâtiment communal situé sur leur territoire pour la célébration d'unions, à condition de recueillir l'autorisation préalable du procureur de la République « en lui transmettant son projet de décision d'affectation, accompagné de tous documents utiles » (**CGCT**, **art. R. 2122-35**) lui permettantde s'assurer « que la décision du maire garantisse les conditions d'une célébration solennelle, publique et républicaine [et] que les conditions relatives à la bonne tenue de l'état civil sont satisfaites » (**CGCT**, **art. L. 2121-30**). Le procureur de la République dispose de deux mois pour faire connaître son opposition motivée, sauf si les éléments transmis lui paraissent insuffisants pour la formuler. Dans ce cas, le délai est prorogé d'un mois par le procureur, qui en informe le maire. En cas de non-opposition, ce dernier peut donc affecter un autre bâtiment que celui de la maison commune, en communiquant au procureur copie de sa décision.Conditions de délégation des fonctions du maire en qualité d'officier d'état civil. Le **décret no 2017-270** précise aussi les conditions de délégation des actes d'état civil aux fonctionnaires de la commune. Ainsi, le maire peut dorénavant « déléguer à un ou plusieurs fonctionnaire titulaires de la commune tout ou partie des fonctions qu'il exerce en tant qu'officier d'état civil, sauf celles prévues à **l'article 75 du Code civil** ».Pour rappel, cet article réserve la cérémonie du mariage dans son ensemble, et notamment la lecture des **articles 214 et 215 du Code civil**, à l'officier d'état civil seul, c'est-à-dire le maire, ses adjoints et uniquement sous réserve de délégation spécifique du maire en cas d'empêchement de celui-ci, les conseillers municipaux... Autrement dit, un fonctionnaire de la commune ne peut obtenir du maire une délégation pour lire ces articles. Le décret précise également que lorsque le maire délègue ses fonctions, « les actes dressés dans le cadre des fonctions ainsi déléguées » doivent comporter la seule signature du délégué.Sources : **Décret no 2017-270 du 1er mars 2017** relatif à la délégation des fonctions d'officier de l'état civil exercées par le maire et au lieu de célébration des mariages **Loi no 2016-1547 du 18 novembre 2016** de modernisation de la justice du XXIe siècleC.

Reference Summary : Le 1er mars 2017, le Gouvernement a pris le **décret n° 2017-270** en application des dispositions de la **loi du 18 novembre 2016** de modernisation de la justice du XXIe siècle relatives à la délégation des fonctions d'officier de l'état civil du maire et du lieu de célébration des mariages. La **loi n° 2016-1547** offre désormais, par son **article 49** (codifié à **l'art. L. 2121-30** du **CGCT**), la possibilité aux maires de célébrer des mariages dans un autre bâtiment communal que celui de la mairie, ainsi que la faculté de déléguer plus largement leurs fonctions. Ces dispositions sont entrées en vigueur le 4 mars 2017 et sont codifiées aux **articles R. 2122-10 et R. 2122-11** du Code général des collectivités territoriales.

Baseline - ECC Generated Summary : Les maires peuvent affecter un bâtiment communal depuis le 1er mars 2017, autre que la mairie, pour célébrer des mariages, à condition d'obtenir l'autorisation du procureur de la République, qui dispose de deux mois pour s'y opposer (prorogeable d'un mois si les pièces transmises sont insuffisantes). Cette autorisation vise à garantir les conditions de solennité, de publicité et la conformité à la tenue de l'état civil. En cas de non-opposition, le maire peut formaliser sa décision et en informer le procureur. Le (**décret no 2017-270** permet également au maire de déléguer certaines fonctions d'état civil aux fonctionnaires de la commune, sauf celles relatives à la cérémonie de mariage, notamment la lecture des (**articles 214 du code civil** et **décret 216 du Code civil**, qui reste réservée au maire, à ses adjoints ou, en cas de délégation spécifique, à des conseillers municipaux.

Baseline - Mixtral Generated Summary : Les maires peuvent célébrer des mariages dans un lieu autre que la mairie, à condition que celui-ci soit situé sur le territoire communal et que le procureur de la République approuve cette décision **CGCT**, (**art. R. 2121-370**). Le procureur dispose de deux mois pour examiner le projet et peut proroger ce délai d'un mois en cas de documents insuffisants. Une fois l'autorisation obtenue, le maire doit communiquer sa décision au procureur. Le **décret n° 2017-270** introduit également la possibilité pour le maire de déléguer certaines de ses fonctions d'officier d'état civil à des fonctionnaires communaux, à l'exception des cérémonies de mariage, qui restent sous la responsabilité exclusive du maire, de ses adjoints ou, en cas de délégation spécifique, des conseillers municipaux. Les actes établis par les fonctionnaires délégués doivent porter leur seule signature.

Generated Summary (Setting 2) : Depuis le 4 mars 2017, les maires peuvent affecter un bâtiment communal autre que la mairie pour célébrer les mariages, à condition d'obtenir l'autorisation préalable du procureur de la République (**CGCT**, **article R. 2122-11**). Cette autorisation garantit le respect des conditions de solennité, de publicité et de conformité à la tenue de l'état civil (**CGCT**, **art. L. 2121-30**). Le procureur dispose de deux mois pour s'opposer à cette décision, avec une possible prorogation d'un mois si les éléments fournis sont insuffisants. En l'absence d'opposition, le maire peut procéder à l'affectation du lieu alternatif. Le **décret no 2017-270** encadre aussi la délégation des fonctions d'état civil aux fonctionnaires de la commune. Le maire peut déléguer certaines fonctions, à l'exception de celles relatives à la cérémonie de mariage (C. civ., **article. 65**), notamment la lecture des **articles 214 et 215**, qui reste réservée au maire, à ses adjoints ou aux conseillers municipaux dûment délégués en cas d'empêchement.

Figure 6: Example of a legal document from our Légibase Dataset with its reference summary (excerpts) produced by experts, as well as 3 generated summaries (Setting 2, ECC and Mixtral). Entities of interest are marked in bold. In the reference summary, faithful entities are colored in green. Abstractive entities of the reference summary are colored in dark blue when present in the dataset (in-dataset abstractions) whereas the light blue one is an out-dataset abstraction. In the generated summary, faithful entities are also colored in green. The entity colored in red is a non-factual hallucination whereas the yellow one is a factual hallucination.

D BART Results

Systems	EUR-Lex-Sum								Légibase							
	F ↑		C ↑		FH ↑		$\overline{\text{FH}}$ ↓		F ↑		C ↑		FH ↑		$\overline{\text{FH}}$ ↓	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
Oracle FH	*0.33*	*0.25*	*0.09*	*0.42*	*0.35*	*0.12*	-	*0.57*	*0.60*	*0.12*	*0.33*	*0.21*	*0.22*	*0.10*	-	*0.76*
Setting 0	0.25	0.21	0.06	0.38	0.19	0.06	-	0.77	0.50	0.08	0.28	0.16	0.05	0.08	-	0.83
Setting 1	0.29	0.18	0.07	0.28	0.25	0.07	-	0.74	0.51	0.09	0.26	0.16	0.12	0.06	-	0.86
Setting 2	0.32	0.23	0.08	0.39	0.26	0.08	-	0.67	0.58	0.10	0.31	0.20	0.17	0.09	-	0.77
Setting 3	0.32	0.25	0.08	0.43	0.24	0.08	-	0.65	0.61	0.10	0.29	0.18	0.17	0.08	-	0.79

Table 7: Results of Bart on EUR-Lex-Sum and Légibase, using Entity-Level metrics: Coverage (C), Faithfulness (F), Factual hallucination (FH) and Non-Factual hallucination rate ($\overline{\text{FH}}$)