SPEER: Sentence-Level Planning of Long Clinical Summaries via Embedded Entity Retrieval

Griffin Adams, Jason Zucker, Noemie Elhadad (Columbia University) - ArXiv [Preprint 4 Jan 2024]

Abhibha

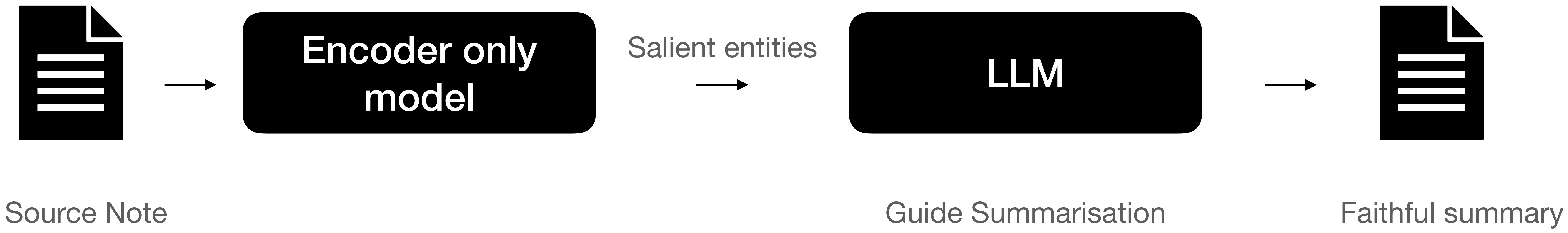
Introduction

- **Motivation:**
  - Hospital course summarization is time-consuming due to the sheer number of clinical concepts covered in admission
  - Frequent copy-pasting of information to generate EHRs leads to entities being entered multiple times -> Note Bloat
- **Challenges:**
  - Generate clinically useful summaries i.e. salient entities are covered.
  - Demonstrate that the entity selection task should be thought of as its own classification task rather than implicitly determined by LLM

Related work

- **LLM Summarisation:**
  - Human evaluation is critical to reveal the efficacy of LLM-generated summaries
- **Guided Summarisation:**
  - Abstractive summ requires three sequential tasks: content selection (extraction), content planning (organization), surface realization (abstraction)
  - Prior work suggests, handling content (entity) selection by a dedicated model outperforms all-in-one approach.
  - Eg: Extractive models can be used to enhance the performance of abstractive model by treating the extract as an auxiliary input with its own encoder -> Gsum
- SPEER interleaves planning and realisation and relies on a separately trained classifier for content selection

Proposed methodology - High level picture



Source Note

Encoder only model

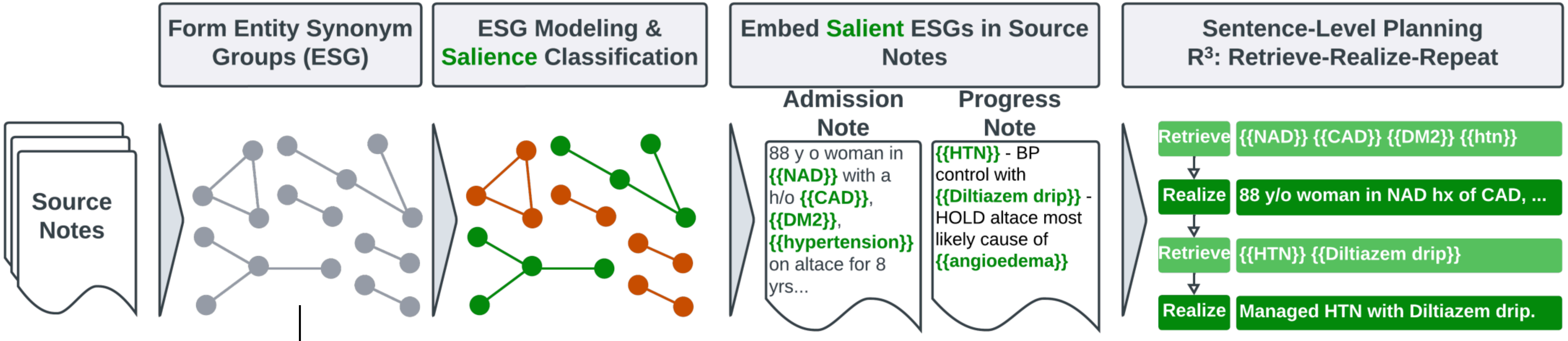
Salient entities

LLM

Guide Summarisation

Faithful summary

Proposed methodology



Form Entity Synonym Groups (ESG)

ESG Modeling & Salience Classification

Embed Salient ESGs in Source Notes

Sentence-Level Planning
R^3: Retrieve-Realize-Repeat

Source Notes

Admission Note

88 y o woman in {{NAD}} with a h/o {{CAD}}, {{DM2}}, {{hypertension}} on altace for 8 yrs...

Progress Note

{{HTN}} - BP control with {{Diltiazem drip}} - HOLD altace most likely cause of {{angioedema}}

Retrieve {{NAD}} {{CAD}} {{DM2}} {{htn}}

Realize 88 y/o woman in NAD hx of CAD, ...

Retrieve {{HTN}} {{Diltiazem drip}}

Realize Managed HTN with Diltiazem drip.

Step 4 & 5

Step 6

Step 7

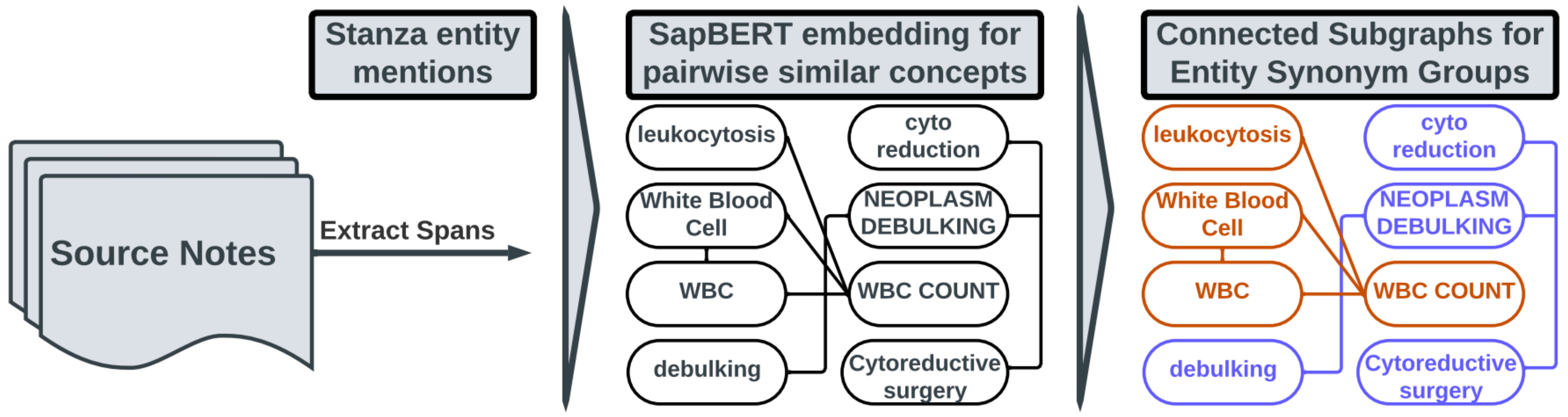


Figure 1: Extracting entities and forming groups of synonymous entities (ESGs). For each admission, we form a set of ESGs from the source notes and content selection is performed by classifying each ESG as salient or not.

Extracting entities

Identifying synonym pairs

Forming ESG's

Step 1

Step 2

Step 3

Proposed methodology

Step 1 - Extracting entities

- Use clinical NER model trained on MIMIC III to extract
  - **Problems:** Diagnoses and symptoms
  - **Tests:** Lab Tests and imaging
  - **Treatment:** Medications and procedures.

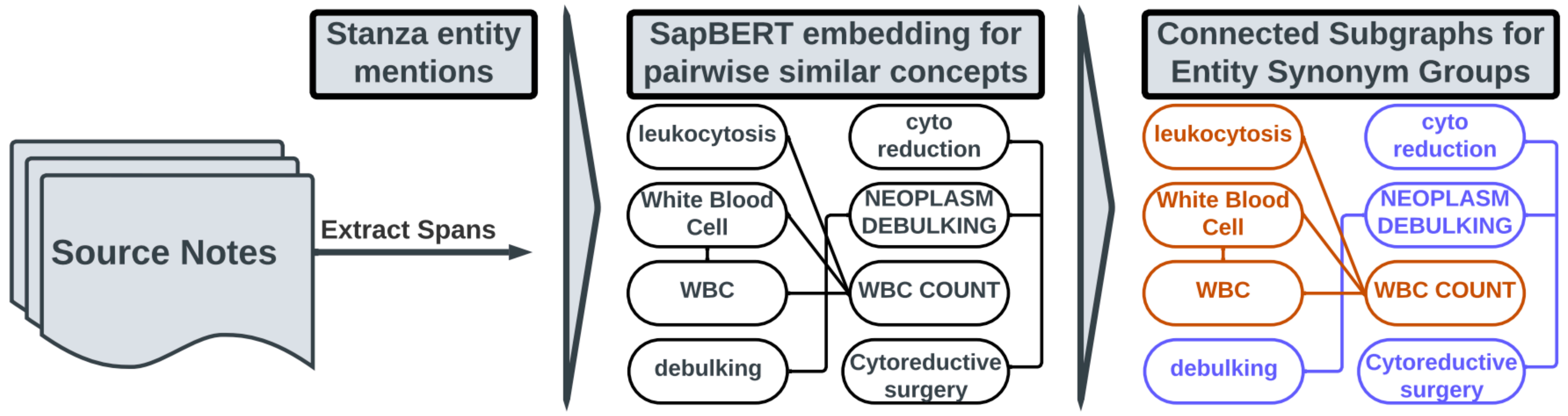


Figure 1: Extracting entities and forming groups of synonymous entities (ESGs). For each admission, we form a set of ESGs from the source notes and content selection is performed by classifying each ESG as salient or not.

Step 1

Proposed methodology

Step 2 - Identifying synonym pairs

- Use similarity in embedding space to identify synonymous clusters of entity spans
- Embed entities using SapBERT - trained tc align synonymous clinical concepts and us cosine similarity to identify synonymous pairs
- Assign labels (unrelated, synonymous) to 1000 pairs
- Use semantic over lexical as WBC -> White blood cell (not lexical but semantically same)

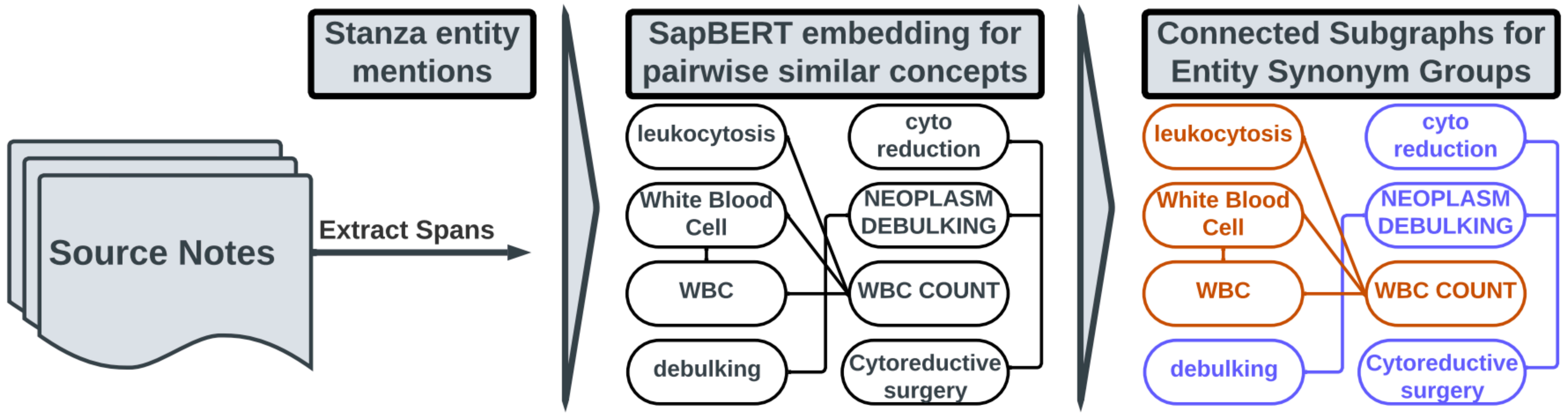


Figure 1: Extracting entities and forming groups of synonymous entities (ESGs). For each admission, we form a set of ESGs from the source notes and content selection is performed by classifying each ESG as salient or not.

Step 2

Proposed methodology

Step 3 - Forming ESG’s

- For each hospital admission, collect all entity mentions and form a graph with one node for each unique entity

- Edge assigned between two mentions i exact match or similarity > 0.75

- Fully connected sub-graphs as ESG’s
  - Reduces entity sparsity
  - Eg: leucocytosis -> condition characterised by high WBC count.

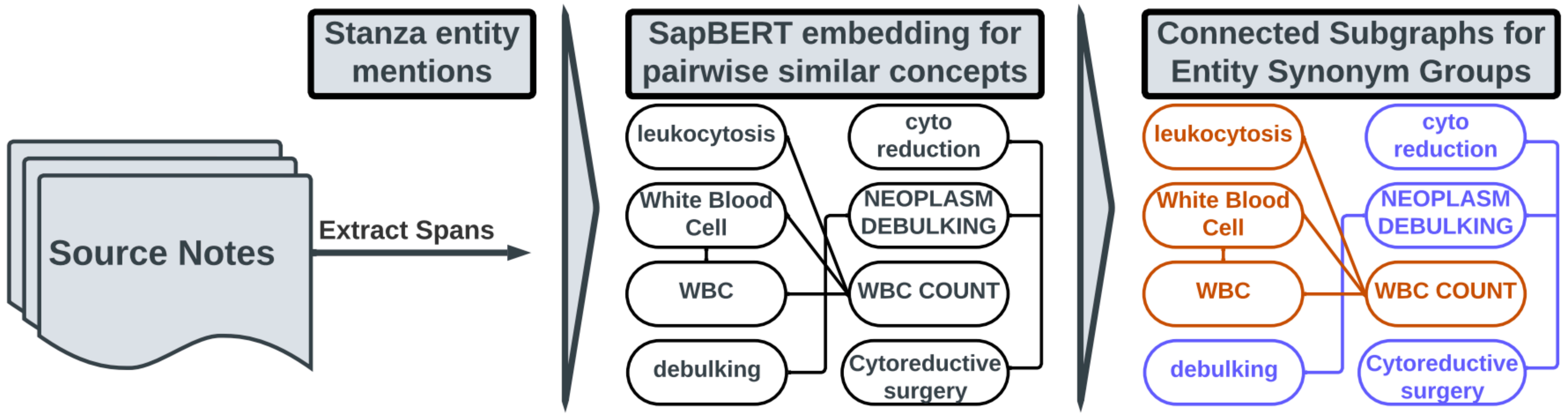


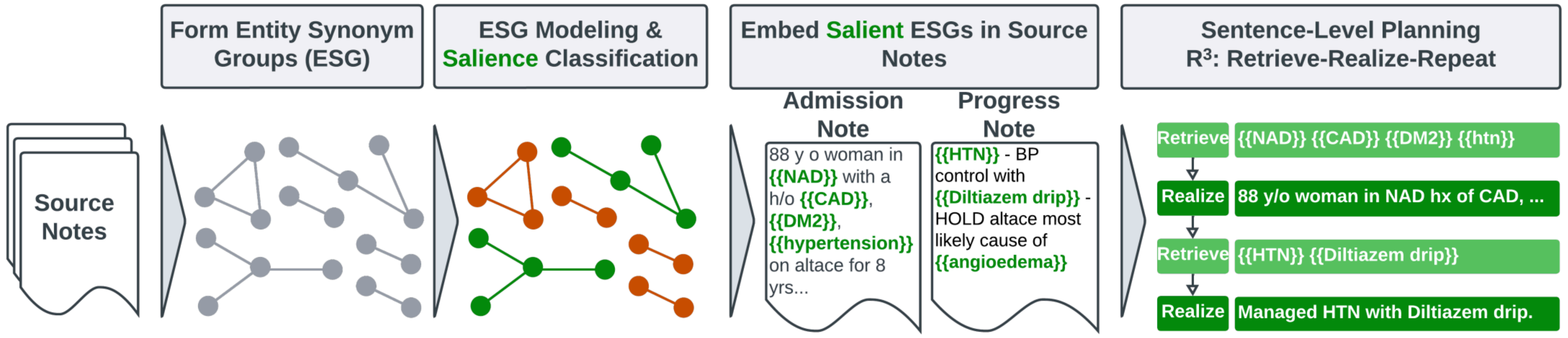
Figure 1: Extracting entities and forming groups of synonymous entities (ESGs). For each admission, we form a set of ESGs from the source notes and content selection is performed by classifying each ESG as salient or not.

Step 3

Proposed methodology

Step 4 - Defining ESG salience

- ESGs are extracted from source notes and based on embedding similarity,
  - ESG is salient if >= 1 spans in the ESG is a synonym of >= 1entity spans extracted from reference summary
- Only 5.7% of the source ESGs are salient -> makes the task difficult.



Form Entity Synonym Groups (ESG)

ESG Modeling & Salience Classification

Embed Salient ESGs in Source Notes

Sentence-Level Planning R^3: Retrieve-Realize-Repeat

Source Notes

Admission Note

88 y o woman in {{NAD}} with a h/o {{CAD}}, {{DM2}}, {{hypertension}} on altace for 8 yrs...

Progress Note

{{HTN}} - BP control with {{Diltiazem drip}} - HOLD altace most likely cause of {{angioedema}}

Retrieve {{NAD}} {{CAD}} {{DM2}} {{htn}}

Realize 88 y/o woman in NAD hx of CAD, ...

Retrieve {{HTN}} {{Diltiazem drip}}

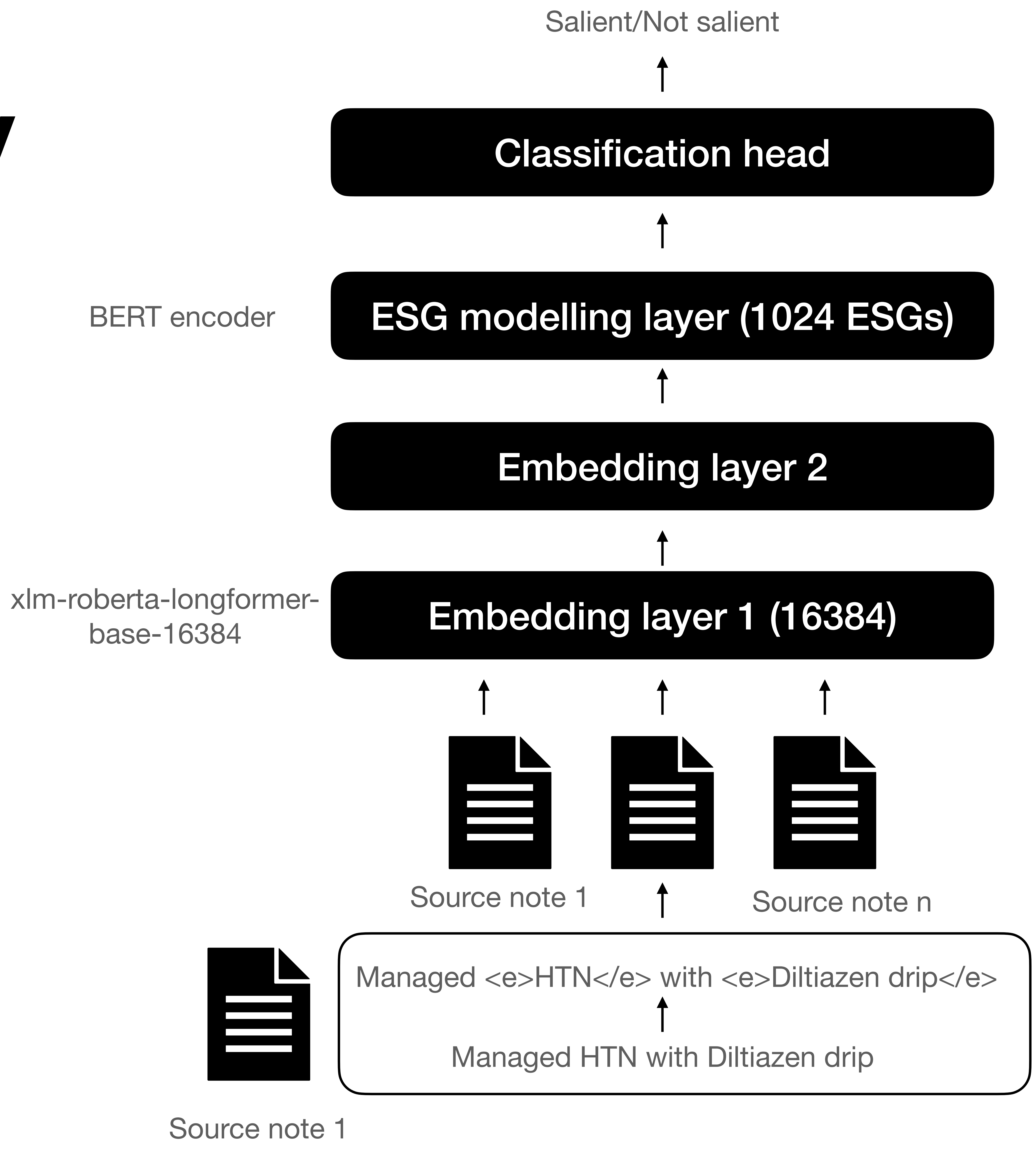
Realize Managed HTN with Diltiazem drip.

Step 4 & 5

Proposed methodology

Step 5- Learning ESG salience

- Use hierarchical token-to-ESG encoder model to perform binary classification
  - Demarcate each entity span with <e> and </e>.
  - Concatenate source notes and encode using Longformer
  - Construct hidden state representations of each entity span by mean pooling embeddings for each token in entity span. (Embedding 1)
  - Mean pool hidden states for all entity spans of the same ESG
  - Frequently mentioned concepts are salient hence learn an embedding for relative frequency. (Embedding 2)
  - Initialise ESG modelling layer i.e fully connected BERT encoder layer


Salient/Not salient
Classification head
BERT encoder
ESG modelling layer (1024 ESGs)
Embedding layer 2
xlm-roberta-longformer-base-16384
Embedding layer 1 (16384)
Source note 1
Source note n
Managed <e>HTN</e> with <e>Diltiazen drip</e>
Managed HTN with Diltiazen drip
Source note 1

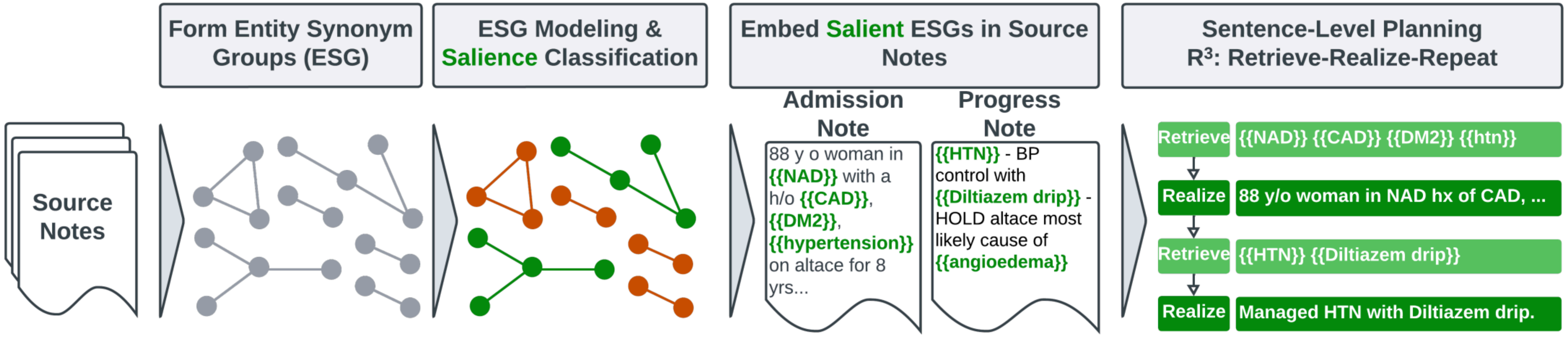
ESG Guided Summarisation

Prompt Guidance - baseline

- Extract salient ESG’s using trained encoder and use those to prompt LLM
- Convert Salient ESGs -> natural language prompt
- **Prompt:** Defining ESG salience, <Randomly shuffled Salient ESGS [PROBLEMS; TREATMENTS; TESTS]>
- ESG classifier can learn the order of ESGs but clinical notes exhibit low coherence
- **Drawbacks:**
    - The model may focus more on the entities and not their actual usage
    - Entity guidance is extensive as there are 100+ entities in one note

ESG Guided Summarisation

SPEER



Form Entity Synonym Groups (ESG)

ESG Modeling & Salience Classification

Embed Salient ESGs in Source Notes

Sentence-Level Planning R^3: Retrieve-Realize-Repeat

Source Notes

Admission Note

88 y o woman in {{NAD}} with a h/o {{CAD}}, {{DM2}}, {{hypertension}} on altace for 8 yrs...

Progress Note

{{HTN}} - BP control with {{Diltiazem drip}} - HOLD altace most likely cause of {{angioedema}}

Retrieve {{NAD}} {{CAD}} {{DM2}} {{htn}}

Realize 88 y/o woman in NAD hx of CAD, ...

Retrieve {{HTN}} {{Diltiazem drip}}

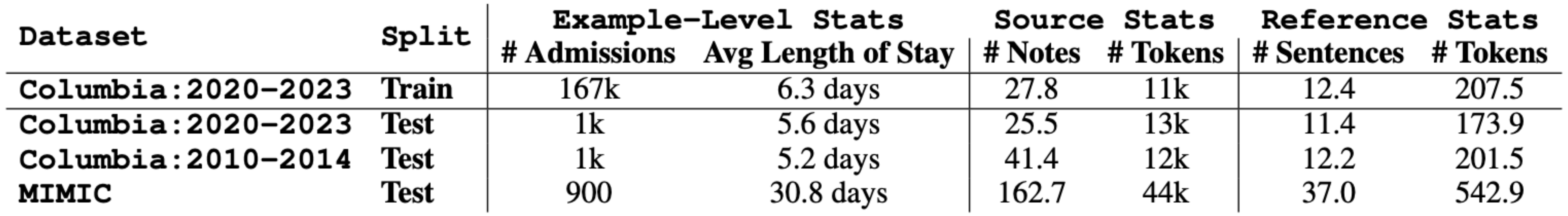
Realize Managed HTN with Diltiazem drip.

Step 6

- Embed salient entities in notes by demarcating each entity span with ‘{{ }}’ tags using the trained encoder.
- Before generating a summary, the LLM generates a list of entities to use and in the order in which they should appear - step 6
- **R3 - Retrieve, Realise, Repeat** - performs retrieval of entities from a fixed set of embedded entities, forms a plan, generates tags {{ }}
    - Encouraging the model to **focus on the specific usage** of an entity
    - **State tracking:** Can keep track on which entities have already been included while generation

Dataset

- Train on single dataset and test on 3
- Training:
    - Train on 167k~ in patient hospital admissions -> 2020 - 2023 (CUIMC)
- Testing:
    - Evaluate on 1000 admissions of Columbia 2020-2023, Columbia 2010-2014 and 900 examples from MIMIC.
- **Note:** MIMIC reference summaries have content which is not mentioned in any of the source notes - reduces scores on reference based metrics.



Dataset	Split	Example-Level Stats		Source Stats		Reference Stats	
		# Admissions	**Avg Length of Stay**	**# Notes**	**# Tokens**	**# Sentences**	**# Tokens**
Columbia:2020-2023	**Train**	167k	6.3 days	27.8	11k	12.4	207.5
Columbia:2020-2023	**Test**	1k	5.6 days	25.5	13k	11.4	173.9
Columbia:2010-2014	**Test**	1k	5.2 days	41.4	12k	12.2	201.5
MIMIC	**Test**	900	30.8 days	162.7	44k	37.0	542.9

Table 2: Statistics for data used for training and evaluating hospital-course summarization models. we use datasets from Columbia University Irving Medical (CUIMC) at two different points of time. We also report scores on MIMIC-III, despite MIMIC having a great deal of unsupported content in reference summaries (Adams et al., 2022).

Metrics

- **Source-Grounded Recall (SGR) -** Focuses on aligning entities mentioned in the model-generated summary with those in the source notes. Helps in the evaluation of relevant entity coverage in the summary.

- **Hallucination Rate (HR) -** Specifically targets and quantifies ungrounded information in the summary and aids in identifying and penalizing fabrications or inaccuracies in the model-generated content.

- **BERTScore-Precision (BSP) -** Measures how well the tokens in the summary align with at least one token in the source notes. Correlates well with fine-grained expert annotations for the faithfulness of hospital course summaries.

- **ClinDistill -** A regression model distilled from several state-of-the-art faithfulness metrics. Provides a sentence-level metric of faithfulness for hospital course summarization.

Experiments

Instruction Templates

Non-Guided

```
[INST]
Generate the BRIEF HOSPITAL COURSE summary.

### Title: Admission Note

DATE: 1/1/2024
NOTE ORDER: 1 of 2
DAY: 1 of 2
ON DAY OF ADMISSION

HPI:
pt is a 90yr old w HTN

### Title: Progress Note

DATE: 1/2/2024
NOTE ORDER: 2 of 2
DAY: 2 of 2
ON DAY OF DISCHARGE

Plan:
pt deemed stable for discharge on ACE
[/INST]
### BRIEF HOSPITAL COURSE:
```

Guided

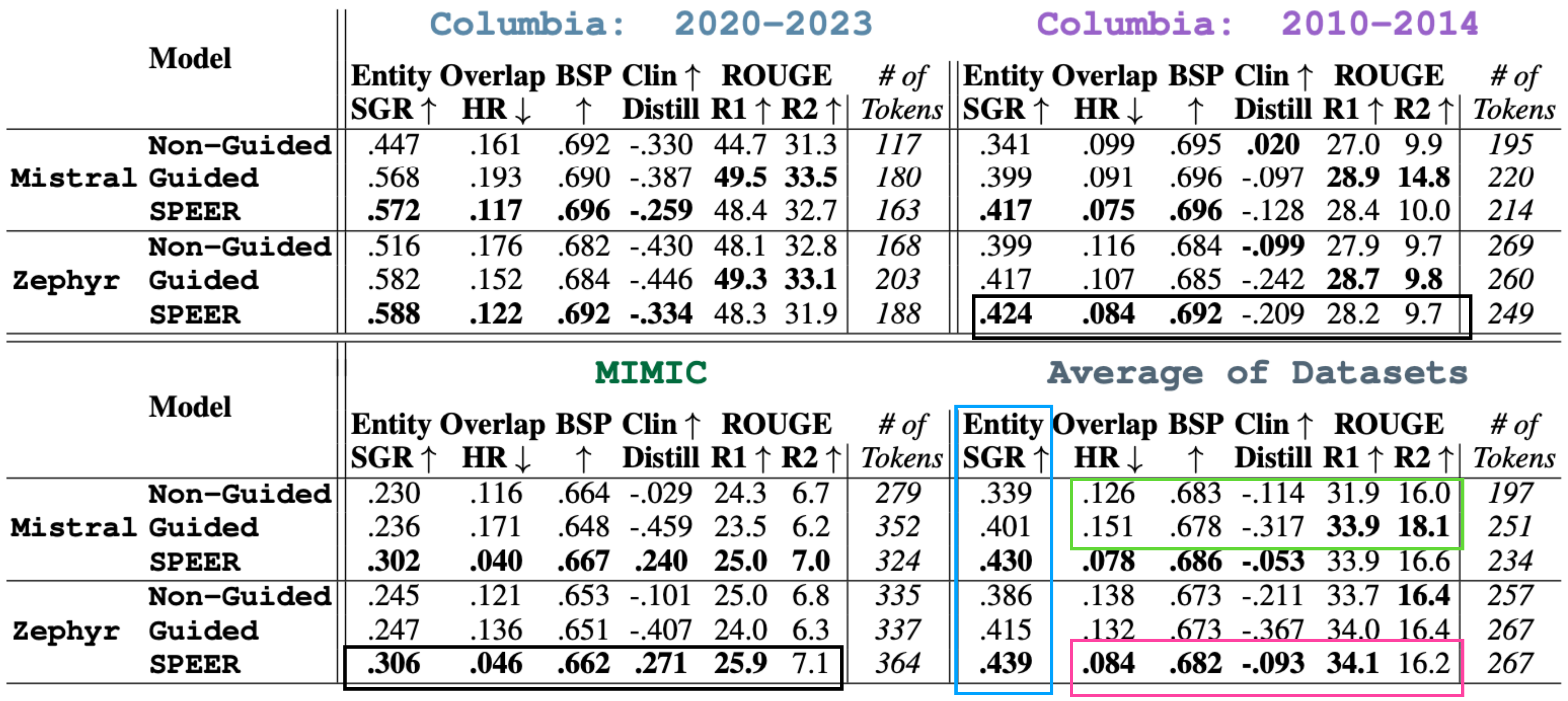
```
[INST]
Generate the BRIEF HOSPITAL COURSE summary using only the medical entities (PROBLEMS, TREATMENTS, and TESTS) provided.
### Title: Admission Note

DATE: 1/1/2024
NOTE ORDER: 1 of 2
DAY: 1 of 2
ON DAY OF ADMISSION

HPI:
pt is a 90yr old w HTN

### Title: Progress Note

DATE: 1/2/2024
NOTE ORDER: 2 of 2
DAY: 2 of 2
ON DAY OF DISCHARGE

Plan:
pt deemed stable for discharge on ACE

### ENTITIES
PROBLEMS:
HTN; Hypertension
TREATMENTS:
ACE; ACE inhibitors
TESTS:
[/INST]
### BRIEF HOSPITAL COURSE:
```

SPEER

```
[INST]
Retrieve a subset of the medical entities in double brackets {{ }} and use them to generate the next sentence of the BRIEF HOSPITAL COURSE summary.
### Title: Admission Note

DATE: 1/1/2024
NOTE ORDER: 1 of 2
DAY: 1 of 2
ON DAY OF ADMISSION

HPI:
pt is a 90yr old w {{ HTN }}

### Title: Progress Note

DATE: 1/2/2024
NOTE ORDER: 2 of 2
DAY: 2 of 2
ON DAY OF DISCHARGE

Plan:
pt deemed stable for discharge on {{ ACE }}
[/INST]
### BRIEF HOSPITAL COURSE:
```

Results & Discussion

- **Implicit versus Explicit Content Selection:** Zephyr and Mistral point values fall well below the precision-recall curves of the classifier.
- Models that rely on entity guidance achieve higher coverage of salient entities than those that do not
- Prompt Guided is surprisingly less faithful than Non-Guided.
- SPEER improves *both* coverage *and* faithful- ness.
- SPEER is more robust to unseen EHRs



Model		Columbia: 2020-2023							Columbia: 2010-2014						
		Entity Overlap		BSP	Clin ↑	ROUGE		# of	Entity Overlap		BSP	Clin ↑	ROUGE		# of
		SGR ↑	HR ↓	↑	Distill	R1 ↑	R2 ↑	*Tokens*	SGR ↑	HR ↓	↑	Distill	R1 ↑	R2 ↑	*Tokens*
Mistral	Non-Guided	.447	.161	.692	-.330	44.7	31.3	*117*	.341	.099	.695	**.020**	27.0	9.9	*195*
	Guided	.568	.193	.690	-.387	**49.5**	**33.5**	*180*	.399	.091	.696	-.097	**28.9**	**14.8**	*220*
	SPEER	**.572**	**.117**	**.696**	**-.259**	48.4	32.7	*163*	**.417**	**.075**	**.696**	-.128	28.4	10.0	*214*
Zephyr	Non-Guided	.516	.176	.682	-.430	48.1	32.8	*168*	.399	.116	.684	**-.099**	27.9	9.7	*269*
	Guided	.582	.152	.684	-.446	**49.3**	**33.1**	*203*	.417	.107	.685	-.242	**28.7**	**9.8**	*260*
	SPEER	**.588**	**.122**	**.692**	**-.334**	48.3	31.9	*188*	**.424**	**.084**	**.692**	-.209	28.2	9.7	*249*

Model		MIMIC							Average of Datasets						
		Entity Overlap		BSP	Clin ↑	ROUGE		# of	Entity Overlap		BSP	Clin ↑	ROUGE		# of
		SGR ↑	HR ↓	↑	Distill	R1 ↑	R2 ↑	*Tokens*	SGR ↑	HR ↓	↑	Distill	R1 ↑	R2 ↑	*Tokens*
Mistral	Non-Guided	.230	.116	.664	-.029	24.3	6.7	*279*	.339	.126	.683	-.114	31.9	16.0	*197*
	Guided	.236	.171	.648	-.459	23.5	6.2	*352*	.401	.151	.678	-.317	**33.9**	**18.1**	*251*
	SPEER	**.302**	**.040**	**.667**	**.240**	**25.0**	**7.0**	*324*	**.430**	**.078**	**.686**	**-.053**	33.9	16.6	*234*
Zephyr	Non-Guided	.245	.121	.653	-.101	25.0	6.8	*335*	.386	.138	.673	-.211	33.7	**16.4**	*257*
	Guided	.247	.136	.651	-.407	24.0	6.3	*337*	.415	.132	.673	-.367	34.0	16.4	*267*
	SPEER	**.306**	**.046**	**.662**	**.271**	**25.9**	7.1	*364*	**.439**	**.084**	**.682**	**-.093**	**34.1**	16.2	*267*

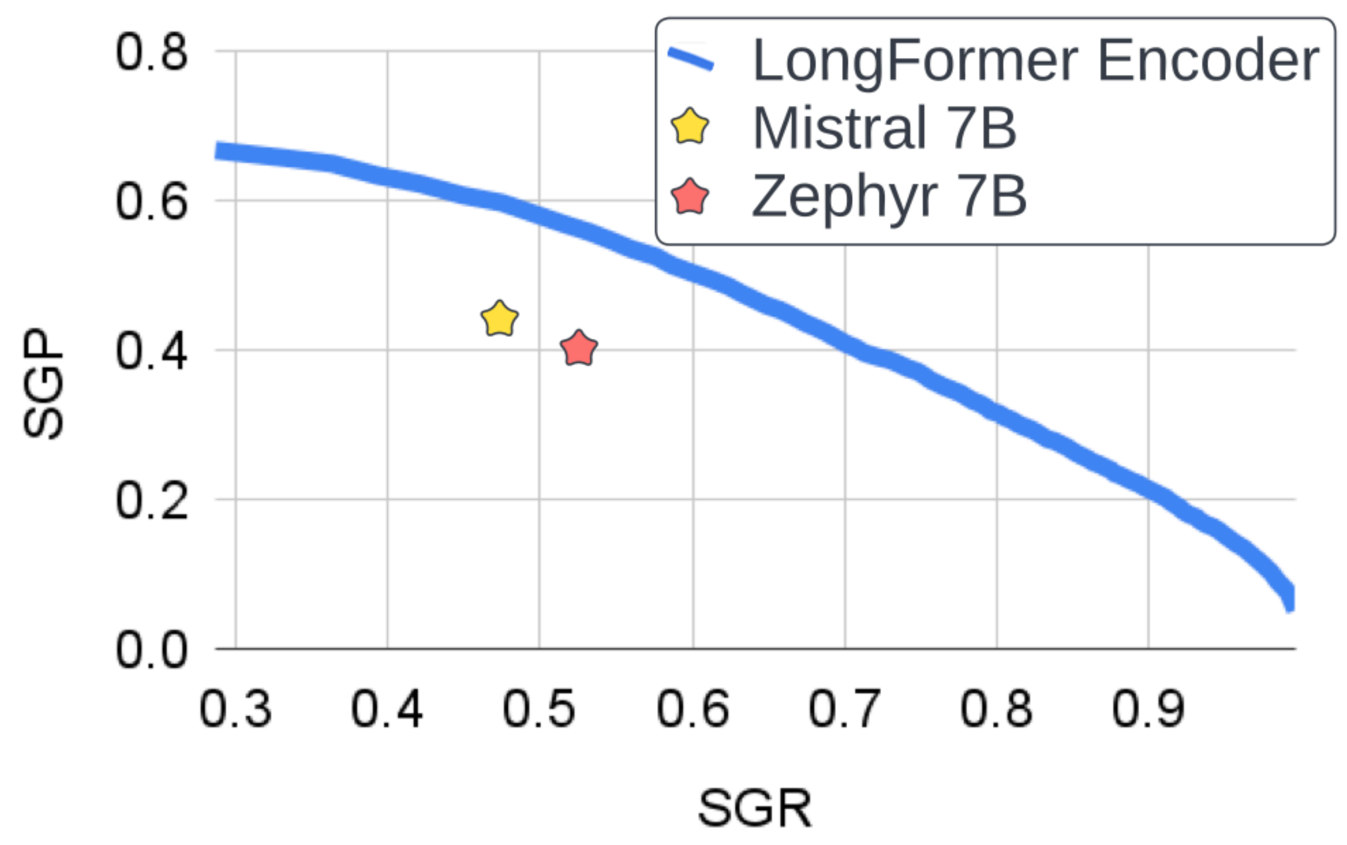
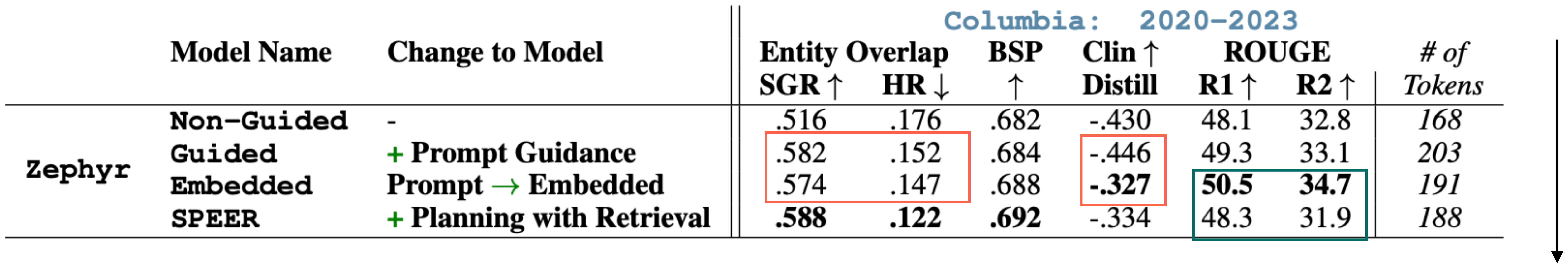


Figure 4: Comparing the entity-level performance (source-guided recall (SGR) and source-guided precision (SGP)) of *explicit* content selection (classifying entities with a LongFormer Encoder) versus *implicit* (autoregressive decoding with Mistral and Zephyr).

Ablations



	Model Name	Change to Model	Columbia: 2020–2023						
			Entity Overlap		BSP	Clin ↑	ROUGE		*# of*
			SGR ↑	HR ↓	↑	Distill	R1 ↑	R2 ↑	*Tokens*
Zephyr	Non-Guided	-	.516	.176	.682	-.430	48.1	32.8	*168*
	Guided	+ Prompt Guidance	.582	.152	.684	-.446	49.3	33.1	*203*
	Embedded	Prompt → Embedded	.574	.147	.688	**-.327**	**50.5**	**34.7**	*191*
	SPEER	+ Planning with Retrieval	**.588**	**.122**	**.692**	-.334	48.3	31.9	*188*

Table 4: From **Non-Guided** to **SPEER**: a step-by-step transition with incremental improvements in faithfulness.

Observe improvements in faithfulness and coverage of salient entities as we transition from the baseline model (**Non-Guided**) to the fully loaded **SPEER** model.

Embedded is **SPEER** without the sentence-level planning. The input is the same (notes with embedded salient ESGs) yet the target output is the summary without planning.

If prompt guidance is replaced with embedded guidance, we achieve a slight decline in SGR, a decrease in hallucinations and an improvement in faithfulness.

Rouge scores decline. Common, yet unsupported, content can artificially boost ROUGE at the expense of faithfulness and coverage.

Conclusion

- First, the researchers explored fine-tuning large language models (LLMs) like Mistral-7B-Instruct and Zephyr-7B-β for the challenging task of hospital-course summarization.

- They found that the process of content selection, which involves deciding which entities to include in the summary, is best achieved by a dedicated salience classifier. This classifier guides the LLM in generating the summary.

- Initially, appending the guidance to the prompt improved the coverage of salient entities but negatively impacted faithfulness. To address this issue and enhance both coverage and faithfulness, they introduced SPEER. It directly retrieves entity guidance from the source notes, resulting in more grounded and complete summaries according to metrics.

Thank you!