Fast Evidence Extraction for Grounded Language Model Outputs

Pranav Mani
Abridge
pranav@abridge.com

Davis Liang
Abridge
davis@abridge.com

Zachary C. Lipton
Abridge
zack@abridge.com

Abstract

Summarizing documents with Large Language Models (LLMs) warrants a rigorous inspection of the resulting outputs by humans. However, unaided verification of generated outputs is time-intensive and intractable at scale. For high-stakes applications like healthcare where verification is necessary, expediting this step can unlock massive gains in productivity. In this paper, we focus on the task of evidence extraction for abstractive summarization: for each summary line, extract the corresponding evidence spans from a source document. Viewing this evidence extraction problem through the lens of extractive question answering, we train a set of fast and scalable hierarchical architectures: EarlyFusion, MidFusion, and LateFusion. Our experiments show that (i) our method outperforms the state-of-the-art by 1.4% relative F1-Score; (ii) our model architecture reduces latency by 4x over a RoBERTa-Large baseline; and (iii) pretraining on an extractive QA corpus confers positive transfer to evidence extraction, especially in low-resource regimes.

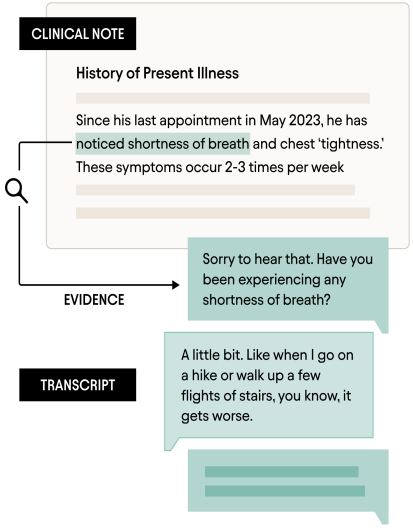

Figure 1: Verifying details in a clinical note requires perusing long conversation transcripts to find substantiating evidence.

1 Introduction

Suppose we train an LLM to summarize a doctor-patient conversation into a clinical note. Such models could save physicians hours each day. However, an auditing step is still requisite. This auditing involves repeatedly diving through a long transcript to find relevant information for every detail that appears in the note (see fig.1). Without an automated mechanism that makes this process efficient, can we really say that we've saved a clinician any time?

Workflows that involve grounded tasks that operate on top of a source document (e.g. summarization, dialogue and translation) (Touvron et al., 2023; Bubeck et al., 2023; Widyassari et al., 2022; Rafailov et al., 2023; Liang et al., 2023) are well suited for LLMs (Krishna et al., 2021; Lehman et al., 2019; Lei et al., 2016; Asan et al., 2020). However, owing to the lingering limitations of these models, humans have remained firmly in the loop, providing last-mile verification of the model's outputs. In these setups, an individual may spend a significant amount of time on verification of LLM-generated first drafts. For grounded tasks, verifying each generated sentence can be broken down into two steps (i) locating a span of text from the much larger source document that has information related to that sentence; (ii) using the obtained span to form conclusions about correctness. With long sources (e.g. hour-long conversation transcripts), it's likely that carrying out the first step of extracting the right span of evidence proves more cumbersome than using the extracted evidence to make conclusions. Furthermore, this problem is exacerbated as the transcript grows in length. Therefore, we present automated Evidence Extraction (EE) as an efficient and scalable way to reduce verification time and fully realize the benefits of workflow automation.

Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 205–218

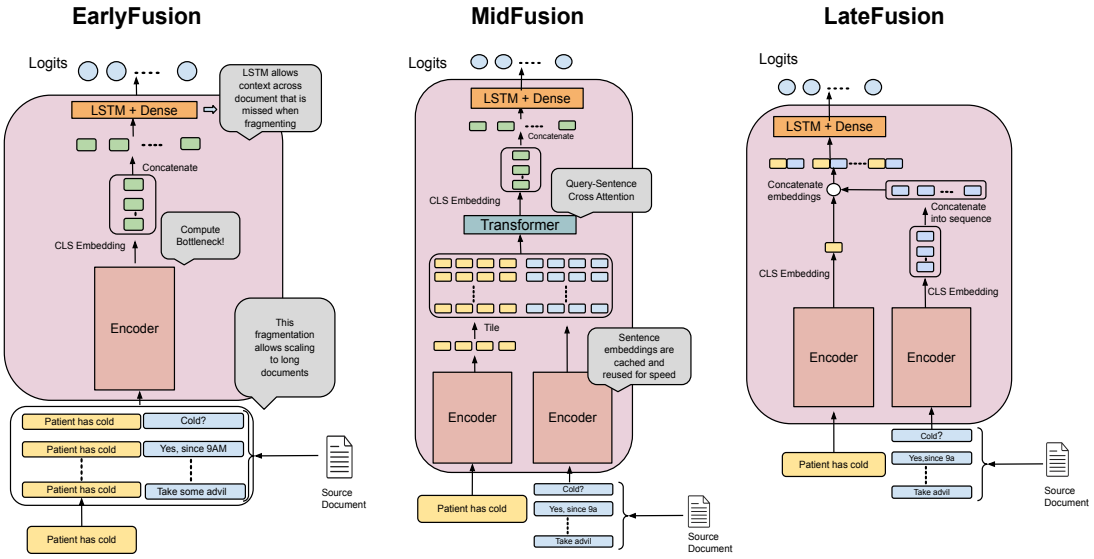

Figure 2: Architectures of the Early, Mid and LateFusion models. Breaking down the source document into sentences helps scale to large documents. In Late and MidFusion architectures the encoders are separated allowing us to cache the embeddings of the sentences of the source. In the MidFusion model, we do not immediately select the CLS embedding but concatenate with the query embeddings for an intermediate transformer step.

In this work, we pose the Evidence Extraction (EE) problem as follows: Given a sentence that requires verification (query) and a source document D that this line should be grounded in, can we identify spans in D (evidence-spans) that contain information relevant to the query? We immediately notice a parallel to Extractive Question Answering (QA): query to question, source to passage, and evidence to answer. This parallel allows us to (i) explore designs for model architectures drawing inspiration from the dual-encoder and cross-encoder families in QA; (ii) explore the benefits that training on QA datasets confer to EE. The latter point proves useful since EE data is hard to come by while ample amounts of QA datasets are available.

In our work, we are focused on exploring simple architectures that are scalable and fast when working with source documents that span beyond thousands of words. For scaling to longer documents, we consider hierarchical architectures that break down a source document into sentences which are encoded independently through RoBERTa-like backbones. We then add document-wide context by concatenating them along the sequence dimension and passing them through an LSTM. For speed, we aim to decouple the encoding of the query and the encoding of the source document. This allows us to amortize the higher cost of computing source document embeddings by caching them for reuse upon subsequent queries on the same source. After the decoupled encoding process, we combine the obtained source and query embeddings in a Late-Fusion step (see Figure 2).

On the flip side, while slow, we find that early fusion of the query string with each sentence in the source is easier to train and performs well due to query-conditioned encoding of the source sentences. We explore an optimal point in the trade-off between performance and throughput and advocate for the use of our proposed MidFusion (MF) architecture that finds an intermediate point to include query-source cross attention. Further, the performance gap between the Late, Mid and Early Fusion models narrows with access to more training data, or in its absence, QA pretraining data. Thus, practitioners can follow the two step strategy of pretraining an MF architecture on QA data followed by finetuning on available EE data.

Our EarlyFusion (EF) model outperforms the State-of-the-Art on the Unified Summarization Benchmark (USB) dataset by 1.4% relative while our MidFusion model following the two step strategy is 5.8x faster while performing within 5% relative F-Score. On our medical dataset, we find the gap between the three models to be far less emphatic due to our access to nearly 0.5M training points. Further, while F-Score reflects the trade-off between precision and recall, we also compute

human-agreement (HA) of displayed evidence using human annotators on our Medical Dataset. We find that the HA of EF is 96%, MF 94%, LF 90% highlighting a gap between the efficacy of these methods under span selection metrics versus human judgement of helpful evidence. As an addition, we collect feedback from two clinicians who used our EE models for verifying LLM generated clinical notes in real clinics. We begin by highlighting relevant prior work in the next section.

2 Related Work

Innovations in better LLM generations are plenty (Lewis et al., 2020; Wallace et al., 2021; Choubey et al., 2021; Wei et al., 2022; Ramprasad et al., 2023; Rafailov et al., 2023). However, our work is situated among post-hoc methods that serve to increase trust in these generations. With the tendency of LLMs to hallucinate (Kalai and Vempala, 2023; Xu et al., 2024) there has been growing interest in post-hoc evaluation of the factuality of LLM generations (Zhang et al., 2021; Manakul et al., 2023; Wei et al., 2024; Goyal and Durrett, 2021; Honovich et al., 2022). Our work considers applications where the aim is not to automatically evaluate each generation but to retrieve supporting material from the source to aid a human with verification. Thus, while scoring the extent of factuality is useful, they cannot replace human spot-checking when an LLM is deployed in a low-risk setting.

While there are similarities with the line of work in Lei et al. (2016); Lehman et al. (2019); Jain et al. (2020) that highlight regions of the input that have correlation with model predictions, they are closer to explaining predictions than explicitly retrieving supporting material. Similar ideas also appear in MultiHop QA works Zhao et al. (2023); Tu et al. (2020); Nishida et al. (2019), but differ in our focus on scale and domain adaptation. The methods in Pruthi et al. (2020) tackle the EE problem in Deep NLP, as we framed it, although they are limited to classification tasks. Further, Kryściński et al. (2019) builds EE and factuality verification models with weak supervision, but their method does not handle cross-sentence dependencies or coreference resolution. More recently, Stammbach (2021); DeHaven and Scott (2023); Krishna et al. (2023); Wadden et al. (2021, 2020) all tackle the EE task, but are distinct given our focus on scalability, speed, and establishing the benefits of QA pretraining for EE. An open-source benchmark for EE is introduced in (Krishna et al., 2023) along with the state-of-the-art methods on this dataset which we compare against.

Domain	# of Examples		
	Train	Valid	Test
Biographies	3740	1875	3642
Landmarks	0	0	211
Disasters	247	122	256
Newspapers	0	0	137
Companies	162	75	156
Schools	220	123	235

Table 1: Number of examples across different domains for the train, validation, and test splits of the Unified Summarization Benchmark (USB) dataset.

3 Methodology

We have a source document D made up of components $u \in U$. Unless mentioned otherwise, u is a sentence (we make explicit when u is a token). An operation (e.g. summarization) on D results in an output O. For each sentence $q \in O$ (e.g summary sentence) we need to find an evidence span $E \subset U$. We refer to q as query. Intuitively, E should have sentences u that contain information relevant to q.

3.1 Proposed Architectures

EarlyFusion Hierarchical Classification For scalability, we first consider a hierarchical architecture that encodes each utterance u_i independently, while adding document-wide context at a later step. This allows us to scale inference to arbitrarily long documents since we batch through the sentences that make up the document. We begin by concatenating the tokens of the query q with the tokens of the i^{th} sentence u_i, separated by a demarcating </s> token. Denote each such query-sentence sequence f_i. Then each f_i is pushed through an encoder backbone (e.g. RoBERTa (Liu et al., 2019)) and the vector corresponding to the CLS token is taken to obtain an embedding r_i. We add document-wide context by concatenating all the sentence embeddings r_is into a sequence and passing this through an LSTM, whose outputs are pushed through a classification head to obtain logits l_i. We consider the sigmoid of the logit $\sigma(l_i)$ to be the score s_i to include u_i in the evidence set E.

LateFusion Hierarchical Classification While processing the document hierarchically allows us

to scale inference to long documents , it does not contribute to faster inference. The main bottleneck is pushing each query-sentence pair through a large backbone like RoBERTa (Liu et al., 2019). If a second query on the same source D originates, we would repeat the entire process. This overhead could be avoided if we independently obtain sentence embeddings for D and reuse them for every query based off of D. Therefore, we consider a late fusion of sentence and query embeddings as follows: Each sentence u_i is pushed through the backbone (e.g.RoBERTa) and the vector corresponding to the `CLS` token is selected as the sentence embedding r_i. These embeddings r_is can be cached. In order to find an evidence set for query q, push the query through the backbone and select the vector corresponding to the `CLS` token as the query embedding r_q. Now concatenate r_q vector and r_i vector to get the late-fused embeddings denoted as, say f_i. Finally, to add document wide context, concatenate these fused embeddings f_is into a sequence which is pushed through an LSTM. Use a classification head on the outputs of the LSTM for this sequence to obtain logits l_i on which we apply a sigmoid to obtain scores s_i for each sentence. For each subsequent query on this source, we can reuse r_is and only need to recompute a single push of the new query through the backbone followed with relatively lightweight LSTM and linear operations.

MidFusion Hierarchical Classification The LateFusion architecture removes several layers of cross-attention between the tokens in the query and the tokens in source sentences that the EarlyFusion architecture enjoys, rendering it much weaker. This leads us to explore where such cross attention could be included while still allowing us to cache the outputs of the backbone model on the source sentences. In the previous architectures, we immediately compress the backbone outputs on the source sentences and the query by simply selecting the `CLS` token's embedding alone. Consider instead that we delay this compression. We could now concatenate the query's token level *embeddings* with each of the source sentence's token level *embeddings* to form a query-sentence sequence, instead of concatenating the query tokens themselves with the source sentence tokens (as we did in EarlyFusion). Formally, push each sentence u_i through the backbone encoder to obtain token level embeddings t_i for each sentence u_i. Cache these embeddings. To find an evidence set for a query q, push the query through the backbone encoder to obtain query embeddings t_q. Now concatenate the query embedding sequence with the token embedding sequence to obtain sequence embeddings $[t_q, t_i]$. These concatenated embedding sequences are passed through a transformer layer, and the outputs of the transformer layer are mean pooled into a single vector r_i. Document wide context is now added by concatenating these r_is into a sequence and operating an LSTM on them, followed through by a classification head. We find that this additional transformer layer before the compression into a single vector with mean-pooling is competitive with the EarlyFusion architecture while still being much faster.

All these architectures are depicted in Figure 2.

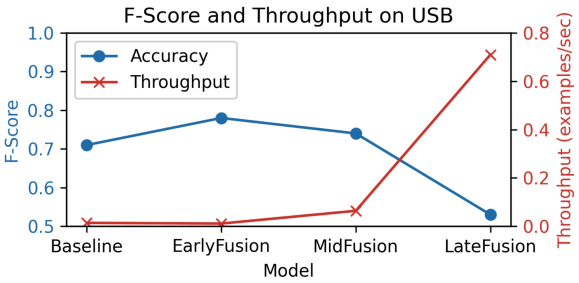

Figure 3: F-Score vs Throughput tradeoff for the three fusion types along with the baseline (adopted from (Krishna et al., 2023)). We use RoBERTa-Large as the backbone encoder. Throughput is computed as an average across the examples in the test split of the USB dataset. We note that the MF model outperforms previous state-of-the-art while having much higher throughput.

3.2 Parallel to Extractive QA

Our Evidence Extraction problem as framed is essentially a span identification problem. Thus, a parallel can be drawn between our task and an extractive QA task (Pearce et al., 2021; Lewis et al., 2019; Xu et al., 2021): query to question, evidence to answer, and source document to passage. An answer in QA tasks is less subjective than evidence and usually has a clearly identified location in the passage. Viewing Evidence Extraction as a harder QA task leads us to explore the benefits of pretraining on QA data. Given the comparatively much higher quantities of QA datasets, we could leverage them for the following reasons:

1. **The need to operate in low-data regimes** Document-Query-Evidence data tuples are scarce. Furthermore, enterprises often update their language models, but re-annotating new EE data each the time the LLM is swapped

Model	F-Score
RoBERTa Large	71.01
T5-Large	77.22
Flan-T5-Large	77.71
Early Fusion (ours)	77.32
Early Fusion++ (ours)	78.80
Mid Fusion (ours)	51.21
Mid Fusion++ (ours)	74.50
Late Fusion(ours)	36.80
Late Fusion++ (ours)	53.06
Llama-13B	5.56
Vicuna-13B	6.65
GPT-3.5-turbo	26.78

Table 2: Results on the USB 4.1 test set. We compute the F-Score at a corpus level by stacking predictions and ground truth for sentences across examples to compute Precision and Recall. ++ indicates models that were first trained on a QA Pretraining Corpus 4.3. The first 3 methods are state-of-the-art from (Krishna et al., 2023).

is impractical. Therefore, the EE model may have to be trained in a low-data regime.

2. **Domain Adaptation gains** Krishna et al. (2023) show that the gains from increasing quantities of in-domain EE train data on OOD test data plateaus. In our experiments we find that pretraining on a related but different task unlocks further domain adaptation gains.

3. **Bi-encoders perform better with more data** Models like our Late and MidFusion models typically converge and perform better when they have access to ample amounts of data. See performance gaps in Table 2 vs Table 4.

We include additional comments and rationale on our methodology in Appendix F.

4 Datasets

4.1 Unified Summarization Benchmark (USB)

The USB dataset (Krishna et al., 2023) is a Wiki-derived benchmark containing annotations for 8 summarization-related tasks. One of those tasks is EE, providing a testbed that is (i) open-source; (ii) presents a low-data regime; (iv) has natural domain splits that allow for testing OOD performance. Dataset statistics are presented in Table 1.

4.2 Medical Dataset

Clinical documentation is one of the leading causes of physician burnout in the United States (Gaffney et al., 2022; Sinsky et al., 2016). Following each encounter, physicians are required to author a SOAP note that covers (S)ubjective (O)bjective (A)ssessment and (P)lan information summarizing the appointment. Traction has been gained by automating the generation of this note using Foundation Models (e.g. see *Abridge AI*). We use a unique corpus containing thousands of recorded clinical conversations (in English) with corresponding SOAP notes created by an annotation workforce trained in SOAP note standards. Composed of 6862 visits of real-life patient-doctor encounters (de-identified to remove PHI information and with full consent), our dataset presents for each visit a trained-worker-scribed transcript, segmented into utterances along with a SOAP note. The conversations are 1.5k words on average. Further, each sentence in the SOAP note is annotated with a supporting evidence span from the conversation. We split the dataset into 5770, 500 and 592 *notes* for train, validation and test splits. Considering each [SOAP note sentence, evidence utterances] tuple as a data point results in 400k train, 50k valid and 50k test samples.This dataset is not open-sourced due to the sensitive nature of the data.

4.3 QA Pretraining Corpus

Our QA Pretraining Corpus is formed by combining three popular Question Answering datasets: SQuAD V1 (Rajpurkar et al., 2016), HotPotQA (Yang et al., 2018), and BioASQ datasets (Krithara et al., 2023). We setup the span-selection problem as sentence classification, to resemble our downstream formulation (Ram et al., 2021). The dataset details are presented in the Appendix in Table 10.

4.4 SynthMed: Synthetically Curated Extractive QA on PubMed Articles

Domain-specific extractive QA datasets are falling out of favor as more focus is given to freeform answer generation. This in tandem with the idea that training on domain-specific extractive QA might be more beneficial than general extractive QA leads us to explore the synthetic generation of domain-specific extractive QA datasets using GPT-4 (Achiam et al., 2023). Given PubMed documents, we prompt GPT-4 to generate QA pairs. We tailor the prompt to focus on challenging questions

Train Domain	Biographies	Companies	Disasters	Landmarks	News	Schools
Mid Fusion	52.15	35.10	31.82	34.84	36.12	40.75
Mid Fusion++	68.08	50.39	42.13	51.01	52.54	48.89

Table 3: Domain Shift experiments on USB dataset. We train the midfusion model on Biographies without (first row) and with (second row) pretraining. We then evaluate its performance on other domains. F-Scores are presented.

Method	Base Model	Precision	Recall	HA
BM25	-	63.01	39.00	65.00
Dual Encoder Retriever	Longformer	79.02	72.10	83.00
Late Fusion (Span Extraction)	RoBERTa-base	74.42	79.06	85.00
Late Fusion (Classification)	RoBERTa-base	75.61	80.29	90.00
Mid Fusion (Classification)	RoBERTa-base	76.37	82.18	94.00
Early Fusion (Classification)	RoBERTa-base	81.29	83.16	96.40

Table 4: Evidence Extraction results on the test split of our Medical Dataset 4.2. We compute the metrics at a character level for better comparison between different granularities and tokenizers. HA (Human Agreement) percentage of examples where the predicted evidence was considered satisfactory by humans.

Method	Precision	Recall	HA
Late Fusion (SE)	65.41	65.72	74.40
Late Fusion (C)	68.21	67.13	76.00
Mid Fusion (C)	71.22	71.98	80.00
Early Fusion (C)	75.27	73.62	84.80

Table 5: EE results on a modified test split of our medical data 4.2 where the queries are modified by applying stochastic rules such as token dropout and reordering. Metrics are computed at a character level. SE: Span Extraction, C: Classification, HA: Human-Agreement. All methods use RoBERTa-base as the backbone.

that have low lexical overlap with the extractive answer, involving multi-hop reasoning, and strictly grounded to the document. We similarly try to generate synthetic Evidence Extraction data but find the generated examples to be of lower quality, often with high lexical overlap between the query and evidence, and sometimes altogether incorrect. For examples and details including the exact prompts used to generate them, refer to Appendix A.

5 Experiment Setup

5.1 General Evidence Extraction

Our first line of experiments aims to test our proposed hierarchical architectures on an open-source benchmark. Accordingly, we use a dataset which contains scope for Evidence Extraction: the USB dataset 4.1. We run experiments with the MidFusion architecture, comparing its domain adaptation performance with and without pretraining. USB provides an organic way to measure domain adaptation capacity by demarcating their data into prespecified domains. We train on the *Biographies* domain and test on the others, providing insights into the benefits of pretraining on out-of-domain data.

We borrow previous state-of-art results on this dataset from (Krishna et al., 2023). We carry out our experiments with RoBERTa-Large (Liu et al., 2019), while adapting the state-of-the-art t5-large (Raffel et al., 2020) and flan-t5-large (Chung et al., 2022) results from (Krishna et al., 2023). We note that there is a discrepancy in the sizes of these models (the t5-large family is at 770M, while RoBERTa-large has 355M parameters with negligible additions from the added LSTM and dense layers) which places us at a disadvantage.

5.2 Medical Evidence Extraction

Our second line of experiments, compares methodologies on our Medical Dataset 4.2. In addition to our hierarchical classification methods, we also include straightforward dual-encoder token-level span selection, as well as LateFusion when posed as span selection. The dual-encoder token-level approach simply encodes the entire transcript using an encoder, and the query using an encoder, concatenates the encodings and classifies start and end tokens for evidence, without any hierarchy involve-

ment. For completeness, we also show the result obtained by a simple BM25 baseline (Robertson et al., 2009). Refer Table 4. For the dual-encoder token-level method we use Longformer (Beltagy et al., 2020) as the choice of backbone for supporting the encoding of long transcripts, while for the hierarchical methods we stick to RoBERTa-base. In addition to Precision and Recall we include an additional metric Human-Agreement (HA) which measures the fraction of examples where a human annotator is satisfied with the conciseness and coverage of the surfaced evidence. It is important to note that Precision and Recall are computed on a test set with 10,000 examples, but Human-Agreement is computed on a random subset of 250 examples since it requires human labor.

In order to test the robustness of our models, we simulate mild distribution shift by adding controlled noise to the queries in the test set. Collaboration with *Abridge* helped us identify realistic noise models that emulate the characteristics of noise observed in hospital systems. These results are presented in Table 5. The noise model is a combination of stochastic token drop in the query, token re-ordering, and inclusion of queries larger than typical of the examples in the dataset.

We also pretrain our models on both generic as well as SynthMed dataset, while testing with and without addition of simulated noise. Refer Table 6.

6 Experiment Results

Naive Baselines are not competitive From Table 4 it is evident that the BM25 model performs much worse than deep learning based alternatives. This puts perspective on the non-trivial nature of the task. Further, conforming with intuition, the BM25 model suffers in recall since rephrasing between the source and query results in lack of a keyword match and requires semantic similarity comparison.

More data helps reduce performance gap In table 2 we see performance leaps as we move from Late to Mid to Early Fusion. However, in table 4 we see that the performance gap while present is not as stark. We attribute this difference to the amount of data available for training. Our Medical Dataset contains hundreds of thousands of data samples while USB contains a few thousand. This also manifests when pretraining on a QA corpus, we see that the gap especially for MidFusion++ model is significantly attenuated, and in distribution shift experiments we see that a further drop in available data when restricted to a single domain drops performance across the board (ref Table 3).

QA pretraining helps Evidence Extraction It is easy to see from Table 6 and Table 2 that QA pretraining confers significant performance boosts despite being a different task. This shows more in low-resource regimes, where MidFusion++ demonstrates similar performance to the full attention models while the boost in performance to EarlyFusion++ seems comparatively modest. Also, QA pretraining has massive impact in robustness as seen in the performance of our models on the simulated OOD medical data (Table 6) as well as domain restricted training on USB (Table 3). While in (Krishna et al., 2023) the authors note that more data for in-domain finetuning does not prove useful, with performance saturating quickly, when faced with a more difficult setting, pretraining on a related task continues to confer large percentage gains in both in-domain and out-of-domain performance. Consistent with their findings, we see that in the EarlyFusion setting, the gains are relatively smaller. Further, as is seen in Table 6, we find that pretraining on domain specific QA data can be more beneficial than training on generic QA datasets especially for niche domains like healthcare.

GPT-4 generated synthetic data is useful From table 6 we see that SynthMidFusion significantly outperforms other types of pretraining methodologies. The pretraining data for this model was curated by prompting GPT-4 as detailed in 4.4. This suggested cheap and efficient ways to lower access to pretraining QA data that is of sufficient quality.

Wide gap between HA and PR-metrics In table 4, 5, 6 we include human evaluation under the column HA (ref 5.2). Evidence relevance as assessed by humans seem to place the model in much better light. This is due to examples where the candidate spans surfaced by the model provide alternate evidence that we consider acceptable under human evaluation but fails to score against the ground truth transcript. The performance gap between Late and EarlyFusion is diluted according to human annotators. Thus, while LateFusion Models are from perfect, they do surface reasonable candidates.

In Appendix D.1 we consider adding document-wide context using a transformer instead of an LSTM. Despite having fewer parameters, LSTMs seem to do better than transformers.

Method	P(M)	R(M)	HA(M)	P(AM)	R(AM)	HA(AM)
MidFusion	76.37	82.18	94.00	71.22	71.98	80.00
MidFusion++	78.38	83.83	95.40	72.73	71.74	81.40
MidFusion++ w DAug.	79.80	83.00	95.20	74.00	75.10	83.80
SynthMidFusion++	81.20	82.41	95.00	72.66	81.10	82.80

Table 6: Evidence Extraction Results on the test splits of our Medical Dataset excluding (M) and including (AM) query augmentation. ++ indicates models that have QA pretraining. MidFusion++ w DAug. corresponds to a MidFusion model where we enabled 10 percent of the medical finetuning data to contain the same query-augmentation strategies as in the AM dataset. SynthMidFusion++ is a MidFusion model pretrained on synthetically generated data 4.4 HA - Human-Agreement 5.2, M: Medical Dataset 4.2, AM: query Augmented Medical Dataset.

Absence of an entity Consider the following line inserted in an LLM generated SOAP note: *"Extremities: No clubbing or cyanosis"* appearing under Physical Exam (PE) section. The PE section is populated this way by default and then changed if an issue is discussed. Here, we need to surface evidence that discussion about clubbing/cyanosis is *not* part of the conversation. This is a failure mode, perhaps for the problem setup itself, since the complete evidence is the entire transcript.

When wrong is it *really* wrong? Often the predicted evidence is reasonable but does not score since it is an alternate source of evidence:

Query: *The patient to continue with the lower dosage of Trulicity if it alleviates the symptoms.*

Predicted Evidence: *"It doesn't cause that but it can make it worse. So, let's change Trulicity to 0.75 mg. It's going to be a dose change. So, use what you have and then we'll go ahead and lower the dosage to make sure that you're doing okay.*

Ground Truth Evidence: *"What you can do is, um, alternate the 1.5 with a 0.75 and you can see if you see a difference in how you feel. And I can give you some of the 0.75 and we'll switch you to the lower dosage because it is true that Trulicity can give you more reflux and if you do have something in your stomach, the bowel issue, it will worsen.*

7 Feedback from Clinicians

With the help of *Abridge* we made our EE model available to two clinicians to aid them in finding evidence for verifying LLM-generated clinical notes from transcripts. We asked them to randomly assign 50 percent of their notes for enabling the aid of our EE model and to carry out the remaining half as usual without this aid (refer to Appendix C for more details on the exact instructions). We then collected feedback:

Feedback.1: *EE dramatically reduces the amount of time required to verify the contents of the AI generated note. Without it, I tend to skim the contents, do keyword searches, and struggle to identify the evidence; this process is frustrating and often negates the time that I saved by not drafting the note myself. I estimate that EE finds the appropriate evidence >75% of the time, and reduces the amount of time needed to review a note from 5 min to 1 min. Moreover, I am more likely to do a comprehensive review of the note when using EE.*

Feedback.2: *The time saved by using EE was consistently 1-2m, almost half the time for a given length, and takes extra cognitive effort without it. Having to scan the whole transcript vs just 3-7 lines of a transcript - huge efficiency booster. I estimate I used EE about a total of 55 times, with 2-3 that may have been close but not quite correct mapping, but minor and corrected when extending the query. In particular, EE makes it easier to check medical terms, reported symptoms, and doses.*[1]

8 Conclusions

In this paper, we described a setup that extracts evidence spans for Language Model outputs on grounded tasks. We presented three hierarchical architectures focused on speed and scalability to long documents, while looking to QA pretraining strategies for boosting performance. We showed that tapping into Extractive QA datasets allows positive transfer even if the curated data is synthetic.

9 Limitations

Some notable limitations:

[1]F.1 is due to an Associate Professor of Medicine, Pulmonary and Critical Care, University of Pittsburg Medical Center and F.2 is due to an MD, University of Pennsylvania, Perelman School of Medicine

1. While the feedback included in 7 is promising, it is not a rigorous clinical study. This paper addresses the first piece of the puzzle: fast and automated EE. A natural future step is to ascertain its impact on reducing the verification burden through formal clinical experiments.

2. It is also possible for users of the EE models to log simple feedback on their satisfaction with the surfaced evidence which could be leveraged to further improve the EE model.

3. The synthetic data generated is of the QA task. While this confers generalization benefits to EE, this choice is also partly a consequence of the relatively poor quality of synthetic EE data that current LLMs generate. In Appendix A we show some examples of synthetically generated EE data even after several iterations of refining the prompts used to generate them. Notably, generated EE queries are often lines copied verbatim from the passage. A future direction is to more comprehensively explore synthetic data generation strategies that might directly yield EE data.

4. An important future step is to explore multilingual capabilities of EE models, with possibilities to have the query and the source be in different languages.

10 Ethics

This study complies with HIPAA guidelines by conducting training and evaluation only on deidentified patient data to ensure privacy and data security. Further, we did not retain or view any patient data when obtaining feedback from clinicians for sec 7. Additionally, all personnel viewing even the deidentified medical data first obtained HIPAA compliance certificates after completing mandatory best-practices online courses.

11 Acknowledgements

The authors would like to thank Elisa Ferracane and John Giorgi for helpful discussions and suggestions. We would also like to thank Dr. Mike Myerburg and Dr. Katherine Choi for their reviews on the impact of using our models in real encounters.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. 2020. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, 22(6):e15154.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *J. ACM*.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Prafulla Kumar Choubey, Alexander R Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Fatema Rajani. 2021. Cape: contrastive parameter ensembling for reducing hallucination in abstractive summarization. *arXiv preprint arXiv:2110.07166*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Mitchell DeHaven and Stephen Scott. 2023. Bevers: A general, simple, and performant framework for automatic fact verification. *arXiv preprint arXiv:2303.16974*.

Alvaro Figueira and Bruno Vaz. 2022. Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15):2733.

Adam Gaffney, Stephanie Woolhandler, Christopher Cai, David Bor, Jessica Himmelstein, Danny McCormick, and David U Himmelstein. 2022. Medical documentation burden among us office-based physicians in 2019: a national study. *JAMA Internal Medicine*, 182(5):564–566.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Adam Tauman Kalai and Santosh S Vempala. 2023. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*.

Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron C Wallace, Jeffrey P. Bigham, and Zachary Chase Lipton. 2023. USB: A unified summarization benchmark across tasks and domains. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972, Online. Association for Computational Linguistics.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. *arXiv preprint arXiv:1904.01606*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*.

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

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. *arXiv preprint arXiv:1905.08511*.

Kate Pearce, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. 2021. A comparative study of transformer-based language models on extractive question answering. *arXiv preprint arXiv:2110.03142*.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. *arXiv preprint arXiv:2101.00438*.

Sanjana Ramprasad, Elisa Ferracane, and Sai P. Selvaraj. 2023. Generating more faithful and consistent soap notes using attribute-specific parameters. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760.

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 14–20. Association for Computational Linguistics.

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*.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9073–9080.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2021. Multivers: Improving scientific claim verification with weak supervision and full-document context. *arXiv preprint arXiv:2112.01640*.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

Peng Xu, Davis Liang, Zhiheng Huang, and Bing Xiang. 2021. Attention-guided generative models for extractive question answering. *arXiv preprint arXiv:2110.06393*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Sen Zhang, Jianwei Niu, and Chuyuan Wei. 2021. Fine-grained factual consistency assessment for abstractive summarization models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 107–116.

Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. 2023. Hop, union, generate: Explainable multi-hop reasoning without rationale supervision. *arXiv preprint arXiv:2305.14237*.

A Synthetic Data

In table 7 and 8 we show some randomly picked examples from our synthetically generated Question Answering and Evidence Extraction datasets respectively. We see a stark difference in the depth of the QA examples versus those of the EE examples, leading us to primarily consider the QA data for pretraining experiments. The following prompts were used to create these examples

1. **QA Prompt**: *"Generate challenging question-answer pairs given a passage, abiding by the following instructions. (i) The answer should be an extractive span from the passage. (ii) Answering the question should require reading and comprehending the full passage but should not require any knowledge not found in the passage. (iii) The question can be in question form or statement form in which case the answer should correspond to evidence from the passage for that statement. (iv) Rewrite the question such that it has low lexical overlap with the answer. (v) Your response should be in JSONL format where each line is a dictionary containing keys 'Statement' and 'Evidence'. Passage:*

2. **EE Prompt**: *Generate statement-evidence pairs given a passage, abiding by the following instructions. (i) The evidence should be an*

extractive span from the passage. (ii) Locating the evidence should require reading and comprehending the full passage but should not require any knowledge not found in the passage. (iii) The statement can be such that there is evidence in the passage to contradict or substantitate it. (iv) Rewrite the statement such that it has low lexical overlap with the evidence. (v) Your response should be in JSONL format where each line is a dictionary containing keys 'Statement' and 'Evidence'. Passage:

B Versions of Software Packages

Numpy 1.24.4, Python 3.10.12, transformers 4.29.0, torch 2.0.0, SpaCy 3.6.0, fuzzywuzzy 0.18.0, openai 1.33.0

C Instructions for Feedback from Clinicians

For obtaining F.1 and F.2 in sec 7 we first created an interface with a simple mechanism to toggle the LateFusion Model from table 4 on and off. They were asked to randomly assign 20 of 40 notes to their usual verification process without EE model assistance, and the remaining 20 with EE model assistance for querying evidence. The random assignment also allows us to remove biases in opinion that may arise if one set is completed first before the other due to the fatigue factor as they get to the tail end of the experiment. The clinicians were then asked to provide feedback paying attention to

1. Any change in average time required to verify a clinical note when using our model as opposed to without

2. An estimate of how many times the model was queried and what fraction of responses was relevant evidence

3. If the use of the model led to identification of errors that would have otherwise passed unseen or impact on confidence in the final note when using our model in the loop.

D Ablations

D.1 LSTM vs Transformer for adding document wide context

For adding context across document (which is important for identifying non-contiguous evidence spans and coreference resolution), our architectures incoporate an LSTM, which is also thematically light-weight in alignment with our efforts for low-inference latency, that operates on the independent sentence embeddings by treating them as a sequence. In this section, we justify our use of the LSTM over transformer layers by considering an ablation. We run experiments on the Unified Summarization Benchmark dataset with a transformer instead of an LSTM in the final step. The results are shown in table 9.

E Details of our QA corpus

In Table 10 we show the number of examples we use in the train, test and validation splits of our QA corpus. The positive to negative class prportion is calculated by considering the ratio of number of sentences that have positive label to the number of sentences that have label zero.

F Comments on Methodology

Here we briefly include some commentary on the methodology and relegate the rest to the analysis of experiments.

Choice of Classification Setup: The task is to produce prediction sets. Therefore, the space of predictions is the power set of U (Tsoumakas and Katakis, 2007; Bates et al., 2021). Predicting a logit and a corresponding *softmax score* across each member in this set is computationally infeasible. Assigning a softmax score across utterances alternatively is interpreted as comparing the relative scores of different utterances making it into E (multiclass) but does not extend an easy interpretation to selecting multiple utterances (multilabel). Therefore, while we do compute logits *with-context* from neighbouring utterances, we proceed to score each utterance using a sigmoid of its logit[2]. An alternative that applies when the set E contains only a single contiguous span of utterances is to identify start and end utterance pointers for this span. We also include modeling of this type where applicable.

LLMs for Verification: In section 4.4 we discuss the prompting of LLMs to curate QA datasets (Li et al., 2023; Figueira and Vaz, 2022). This is different from their application to generate explanations. The key point is that we are interested in

[2]enabling the selection of multiple utterances based on thresholds (set using cross-validation)

Generated Question	Corresponding Answer
Describe the outcome of capsaicin treatment on the obesity and steatohepatitis development in Pemt(-/-) mice.	disruption of the hepatic afferent vagus nerve by capsaicin failed to reverse either the protection against the HFD-induced obesity or the development of HF-induced steatohepatitis in Pemt(-/-) mice.
How does hepatic vagotomy affect hepatic inflammation and ER stress in Pemt(-/-) mice?	HV increased the hepatic anti-inflammatory cytokine interleukin-10, reduced chemokine monocyte chemotactic protein-1 and the ER stress marker C/EBP homologous protein.
Elucidate the method used to validate candidate genes following array analysis.	pyrosequencing and genotyping for putative methylation-associated polymorphisms performed using standard PCR
How many genes showed a significant number of BWC-linked CpGs, and what was this threshold?	four of which showed ≥ 4 BWC-linked CpGs
In what way were subjects paired with the control group in the HS prevalence study?	matched with controls based on age, gender, and race

Table 7: Examples from GPT-4 generated synthetic QA data. This is a random sample and non-cherry picked, but it is possible to see the innate ability of these models to generate quality QA examples for training.

Generated Query	Corresponding Evidence
ILC2s were increased in patients with coexisting asthma among the CRSwNP population.	ILC2s were increased in patients with coexisting asthma (P = 0.03) in the CRSwNP population.
Pemt(-/-) mice are protected from HF-induced obesity when fed a high-fat diet (HFD).	Pemt(-/-) mice are protected from HF-induced obesity; however, they develop steatohepatitis.
A higher chemotherapy effect on lymphocytic infiltration is associated with pCR and better prognosis.	A higher infiltration by CD4 lymphocytes was the main factor explaining the occurrence of pCR, and this association was validated in six public genomic datasets.
Cluster Y is a profile mainly characterized by high CD3 and CD68 infiltration.	Immune cell profiles were analyzed and correlated with response and survival.
A higher infiltration by CD4 lymphocytes predicts pathological complete response to neoadjuvant chemotherapy.	We identified three tumor-infiltrating immune cell profiles, which were able to predict pathological complete response (pCR) to neoadjuvant chemotherapy

Table 8: Examples from GPT-4 generated synthetic EE data. This is a random sample and non-cherry picked, yet it is apparent that these examples consist of statements that have high lexical overlap with sentences in the passage.

outputs that point to locations in a document that a human can quickly verify. While using LLMs in chain-of-thought or self-rationalizing through explanations is a form of interpretability, they do not mitigate the need for a human to verify even those freeform explanations.

Model	Fusion	F-Score
EarlyFusion	LSTM	77.32
EarlyFusion	Transformer	76.61
EarlyFusion++	LSTM	78.80
EarlyFusion++	Transformer	78.12
MidFusion	LSTM	51.21
MidFusion	Transformer	48.87
MidFusion++	LSTM	74.50
MidFusion++	Transformer	74.31
LateFusion	LSTM	36.80
LateFusion	Transformer	39.72
LateFusion++	LSTM	53.06
LateFusion++	Transformer	54.13

Table 9: Ablation Study: We consider the use of Transformer instead of LSTM for the final stage of our hierarchical architecture. Results are shown on the test split of the USB dataset. The F-Score is computed at an utterance level by computing micro precision and recall.

Entity	Value
# Train Samples	180469
# Validation Samples	13006
Positive to Negative Class Proportion	0.073

Table 10: Dataset statistics for our QA Pretraining Corpus, which consists of a mixture of SQuAD, HotpotQA, and BioASQ.