
Scaling Down: Multi-Hop Information Retrieval in Resource-Constrained Environments

Nikolay Staroverov¹

Abstract

While Multi-Hop Question Answering is a foundational task in Natural Language Processing, the current approaches rely on Large Language Models, which we characterize as computationally prohibitive for local deployment. In this paper, we propose a data-centric approach to Multi-Hop Retrieval, utilizing a 5.38M parameter model with 195.5ms of CPU latency, which occupies only 20.5MB of disk space. On standardized numerical benchmarks, it achieves 0.61 on MRR and 0.56 on Recall@1, outperforming the non-iterative sBERT baseline by factors of $1.72\times$ and $2.75\times$, respectively. Through the Broken Chain experiment, we demonstrate that such models have the ability to implicitly learn sequential dependencies, previously often observed in larger-scale architectures. We also utilize the Explicit Positional Encoding (EPE) tokens as a way to ground the model's output in long-sequence environments. Furthermore, we characterize EPE as a structural regularizer, demonstrating its ability to mitigate the Sequence Tax in small recurrent architectures.

1. Introduction

Multi-Hop Question Answering is foundational to NLP, as real-world information is typically distributed across separate sources. Retrieval-Augmented Generation (RAG) emulates this synthesis by grounding neural models in external evidence.

The majority of current research in this field focuses on applying Large Language Models (LLMs), specifically large Transformers, to this domain. However, this raises concerns, particularly due to the energy efficiency and size of such models. As most Transformers require notable hardware, it makes them increasingly hard to run on consumer-grade GPUs, and nearly impossible to run on consumer-grade

CPUs. Without the ability to run local inference, many users rely on external LLM providers to access them, which raises privacy concerns and requires significant financial overhead if those models are being used via an API.

In order to introduce a way of solving this problem, we apply a micro-parameter model to the task of Multi-Hop Retrieval. The proposed model is a 5.38M GRU with an attention mechanism and a Pointer-Generator mechanism. We re-define the task into a retrieval problem: in Multi-Hop Retrieval, the model does not have to generate the answer from scratch, but instead generates structured sequences of keywords to navigate a Knowledge Base with the help of downstream BM25 (Robertson et al., 1994) retrieval to retrieve answering sentences. We propose a data-centric approach to this problem, using Top-8 IDF compression, sentence-level shuffling, Query Expansion and Explicit Positional Encoding (EPE) as our main techniques.

The primary contributions of this paper are as follows:

1. We demonstrate that a 5.38M parameter model can achieve high recall (Recall@1 = 0.5619) on Multi-Hop Retrieval, significantly outperforming unassisted neural baselines (sBERT), while maintaining sub-200ms CPU latency.
2. We provide empirical evidence (via the "Broken Chain" experiment) that such micro-parameter architectures are capable of learning implicit logical dependencies without explicit prompting techniques.
3. We show that Explicit Positional Encoding can be used to mitigate the Sequence Tax observed in long-sequence environments and to provide grounding on Out-of-Distribution (OOD) data.

Code, pre-trained model weights and preprocessed datasets are available at <https://github.com/winkelee/Scaling-Down-Multi-Hop-Retrieval>.

2. Related Work

The majority of previous work has focused on Query Decomposition (Wolfson et al., 2020) or establishing explicit

¹Independent Researcher. Correspondence to: Nikolay Staroverov <nikolay_staroverov@proton.me>.

reasoning via various prompting techniques in Large Language Models (LLMs) (Wei et al., 2023) (Zhou et al., 2023). However, instead of engaging in explicit Query Decomposition or end-to-end Question Answering, we utilize a more resource-efficient approach for smaller neural networks: Information Retrieval.

Furthermore, datasets like *MuSiQue* (Trivedi et al., 2022), *2WikiMultihopQA* (Ho et al., 2020) and *HotpotQA* (Yang et al., 2018) have become the standard for training and assessing various models on Multi-Hop Question Answering capabilities. Most current approaches utilize the strategy of "Scaling Up" the model size in order to achieve better performance. Our work diverges from this strategy, instead investigating the limits of "Scaling Down", proving that a model within a microscopic parameter count can achieve notable performance on Multi-Hop Retrieval. We also provide empirical evidence that such small models have the ability to implicitly learn sequential dependency chains within their output, which are necessary for the retrieval of all supporting sentences.

In addition, See et al. (See et al., 2017) applied Pointer-Generator networks to the task of abstractive summarization of long contexts with a significant vocabulary (50,000). In our work, we utilize them to work with a constrained vocabulary of 8,192 tokens and to generate high-IDF keyword anchors linked to a specific sentence rather than generating natural language.

3. Methodology

3.1. The Multi-Hop Retrieval Task

Our proposed algorithm involves a question Q and a context C as the model’s input data, and a plan T as the output data. $T = \langle s_1, s_2 \dots s_k \rangle$ is a sequence of s_i , where each s_i is a sequence of keywords corresponding to a specific sentence $s_i^* \in S_{\text{supp}}$ in C . Here, S_{supp} is the set of all supporting sentences for Q in C . Each s_i in T is delimited by a "<STEPn>" token.

At inference, we define two sources for information retrieval: \mathcal{D} (a set of all paragraphs available for retrieval, also called the "General Knowledge Base") and \mathcal{S} (a set of individual sentences available for retrieval, also called the "Sentence Knowledge Base"). Notably, $\mathcal{S} = \bigcup \mathcal{D}$.

During training, the datasets provide the required C . However, in a production system it would be necessary to define a method for retrieving such C from \mathcal{D} . We suggest a hybrid approach to the retrieval of C , combining OkapiBM25 (Robertson et al., 1994) and the Max-Sim function as a re-ranking step:

$$\text{MaxSim}(Q, D) = \frac{1}{|\text{tok}(Q)|} \sum_{w_q \in \text{tok}(Q)} \max_{w_d \in \text{tok}(D)} \text{sim}(v_{w_q}, v_{w_d}), \quad (1)$$

where Q denotes the received query and $D \in D_{\text{primary}}$ represents an individual paragraph from the primary set of top- N_{primary} paragraphs retrieved via OkapiBM25. Here, $\text{tok}(Q)$ and $\text{tok}(D)$ are sets of their respective tokens. Then $v_{w_q}, v_{w_d} \in \mathbb{R}^d$ denote the d -dimensional FastText (Bojanowski et al., 2017) vector representations of words w_q and w_d , retrieved from $\text{tok}(Q)$ and $\text{tok}(D)$ respectively. After this re-ranking step, we pick top- $N_{\text{secondary}}$ paragraphs. While N_{primary} can be set arbitrarily, it is recommended to set $N_{\text{secondary}}$ within the bounds of $4 \leq N_{\text{secondary}} \leq 6$, to ensure the alignment of the inference context length with the distribution of training data.

As the model’s output T is a sequence of sequences of keywords, we need to find the sentence of origin of each $s_{i,j}$. After performing Query Expansion on each s_i by adding the keywords from Q to the end of s_i , each s_i acts as a query for the retrieval space \mathcal{S} . As searching the global space \mathcal{S} can yield better-quality answering sentences, it also increases the risk of retrieving out-of-context distractors. In order to ground the searching space to the given context, it is possible to swap \mathcal{S} for \mathcal{S}_{all} , where \mathcal{S}_{all} is the set of all sentences present in C .

3.2. Dataset

MuSiQue (Trivedi et al., 2022), *2WikiMultihopQA* (Ho et al., 2020) and *HotpotQA* (Yang et al., 2018) were chosen as the source datasets for the construction of the training and testing datasets. Each sample $S = \{Q, C, T\}$, where Q is a question, C is a context and T is the target output. As such, for each S :

- Q is a standard, natural-language question.
- C is the union of D_{supp} and D_{dist} , where D_{supp} is the set of all paragraphs containing one or more supporting sentences, and D_{dist} is the set of n_{dist} paragraphs containing no supporting sentences, where $n_{\text{dist}} = \min(2, |D_{\text{all}}| - |D_{\text{supp}}|)$. All paragraphs in C are separated by the special token "<SEP>". EPE is used by appending a special token of type "<P n-Sm>" at the start of every sentence in every paragraph, where n is the number of the paragraph and m is the number of the sentence. Every paragraph is assigned a random place in the context. Sentence-level shuffling of ground truth sentences is also applied to every paragraph in C in order to mitigate any heuristic biases.
- T is the union of all preprocessed supporting sentences from S_{supp} in D_{supp} . Every sentence in S_{supp} undergoes

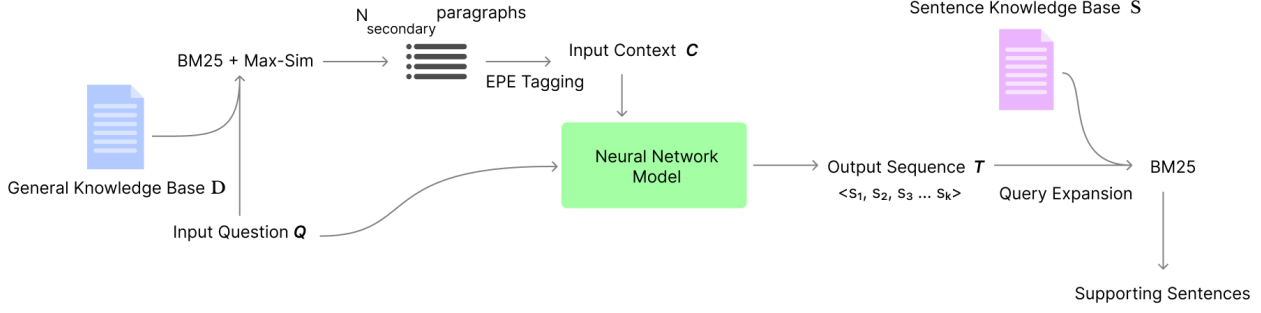


Figure 1. System architecture for Multi-Hop Retrieval. The process includes initial context lookup, EPE tagging, generation of the output, query expansion and the retrieval of the final supporting sentences.

four preprocessing steps: stopword removal, Top-8 IDF (Inverse Document Frequency) sampling, prepending the “<STEP n >” token (where n is the ordinal position of the sentence in T), prepending the EPE token of the respective sentence of the source paragraph in D_{supp} .

The length of T can vary and corresponds to the size of S_{supp} . To acquire tokens from the final strings, word-level embeddings are used.

The fixed vocabulary V , $|V| = 8,192$, was constructed using the most frequent tokens across the entire training corpus (Q, C, T), using T before applying the Top-8 IDF compression.

3.2.1. EXPLICIT POSITIONAL ENCODING (EPE)

As mentioned above, special EPE tokens of type “<P n -S m >” were used in construction of both the input context C and the targets T . These tokens uniquely mark the start of each sentence in each paragraph, spanning the range from “<P 0-S 0>” (the first sentence of the first paragraph) to “<P 7-S 14>” (the 15th sentence of the 8th paragraph). The aim of EPE is to enable the model to first extract the EPE token, and then copy the subsequent keywords.

3.2.2. KEYWORD MATCHING AND TOP-N IDF KEYWORDS

As recurrent neural networks are vulnerable to error accumulation over long sequences, we implement target compression. If $P(w_t)$ is the probability of generating the correct token at timestep t , the cumulative probability of at least one error (P_{err}) increases monotonically with sequence length:

$$P_{\text{err}} = (1 - \prod_t P(w_t)). \quad (2)$$

Additionally, we hypothesize that a small GRU-based neu-

ral network might experience difficulties generating longer sequences due to the lack of sufficient hidden size.

Our proposed solution is to shorten the target output lengths by retaining only verbs, proper nouns, nouns, adjectives and numbers. In addition to this stopwords removal, we calculate IDF scores for each token w_i in our dataset as in (3), where $|\mathcal{D}|$ is the number of all unique paragraphs in the training corpus and $|\mathcal{D}|_{w_i}$ is the number of unique paragraphs in the training corpus that contain the word w_i . After this calculation, only the 8 highest scoring words are retained as the content of an s_i .

$$\text{IDF}_{w_i} = \log\left(\frac{|\mathcal{D}| + 1}{|\mathcal{D}|_{w_i} + 1}\right) + 1. \quad (3)$$

Note that we use the *smoothed* IDF formula in order to prevent division by zero.

3.3. Model Architecture

The proposed architecture is composed of 5 primary modules: Question and Context Encoders (bi-directional single layer GRUs), Decoder (GRU), Attention Mechanism and a Pointer-Generator mechanism.

Both encoders work similarly. Each encoder receives a series of tokens w_1, w_2, \dots, w_k and produces a sequence of hidden states h_i of size $h = 256$ in 1 direction and $2h = 512$ in two directions, as well as the final hidden state h_f . To reduce the parameter count of the model, 2 separate projection layers have been put in place for each encoder (for a total of 4), projecting h_i and h_f from 512 to 256, as shown in (4).

$$\tilde{h}_t = W_{\text{pr}} \begin{bmatrix} \vec{h}_t \\ \overleftarrow{h}_t \end{bmatrix} + b_{\text{pr}}, \quad W_{\text{pr}} \in \mathbb{R}^{h \times 2h}, \quad b_{\text{pr}} \in \mathbb{R}^h. \quad (4)$$

A separate linear layer of type (4) is used once again to project the concatenation of h_t from both encoders from $2h$ to h .

At each decoding time step t , the embedding of the previous token w_{t-1} (“<BOS>” at t_0) is concatenated with the attention vector a_t and passed to the decoder alongside previous decoder hidden state s_{t-1} . During training, teacher forcing is used, so w_t is the previous ground truth token, while during inference w_t represents the previous token generated by the model. The attention distribution a_t is calculated as in the original Pointer-Generator paper (See et al., 2017):

$$e_i^t = \mathbf{v}^T \tanh(W_h h_i + W_s s_{t-1} + w_c c_i^t + b), \quad (5)$$

$$a^t = \text{softmax}(e_i^t), \quad (6)$$

where \mathbf{v}^T , W_h , W_s , w_c and b are learnable parameters. Afterwards, we compute the attention vector by producing a weighted sum of tokens with the attention distribution:

$$h_t^* = \sum_i a_i^t h_i. \quad (7)$$

Following the Pointer-Generator paper (See et al., 2017), we define the coverage vector c_i^t as the sum of attention distributions over all previous decoder time steps:

$$c^t = \sum_{t'=0}^{t-1} a^{t'}. \quad (8)$$

The coverage vector allows us to use coverage loss (9) more efficiently by providing necessary information about already-attended-to tokens.

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t). \quad (9)$$

One of the main ways to lower the model’s parameter size and its computational complexity, is to reduce its vocabulary size. While Byte-Pair Encoding (BPE) reduces vocabulary size, it significantly extends decoding horizons, worsening the P_{err} from Equation 2. We therefore utilize word-level tokenization. To handle Out-of-Vocabulary (OOV) tokens, we use a Pointer-Generator mechanism, which allows us to compute a single value $p_{\text{gen}} \in [0, 1]$ (generation probability) like in (10), where W_h , W_s , W_x and b are learnable parameters. Due to the size of our vocabulary being smaller by a factor of $6.1 \times$ than in the Pointer-Generator paper, we add a small learnable bias to the p_{gen} calculation, to help it copy more of the original input.

$$p_{\text{gen}} = \sigma(W_h h_t^* + W_s s_t + W_x x_t + b). \quad (10)$$

The value of p_{gen} is later used to compute the probability for each specific word w , $P(w)$, which make up the final vocabulary distribution, as shown in (11)

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t. \quad (11)$$

3.4. Training

To train the model, a composite loss function \mathcal{L} is used, defined as:

$$\mathcal{L} = \frac{1}{L} \sum_{t=1}^L (\mathcal{L}_{\text{NLL}}^t + \lambda \cdot \text{covloss}_t), \quad (12)$$

where L denotes the number of tokens in the output sequence T , λ is the *coverage loss coefficient* and $\mathcal{L}_{\text{NLL}}^t$ is the *Negative Log-Likelihood loss* of token w_t at timestep t , defined as

$$\mathcal{L}_{\text{NLL}}^t = -\log P(w_t). \quad (13)$$

The hyperparameter λ regulates the severity of the repetition penalty. Compared to the Pointer-Generator paper, where $\lambda = 1$, our maximum value of λ was set to 0.1 during epochs 6-17. This lenient value of λ ensures the model’s ability to attend to the same token even in the presence of high-ranking IDF tokens in the input context C .

4. Results & Experiments

4.1. Benchmarks

In order to evaluate the Main Model and the ablations, we use three numerical benchmarks:

1. **Mean Reciprocal Rank (MRR)** describes the quality of the downstream BM25 retrieval based on the model’s outputs:

$$\text{MRR} = \frac{1}{|S_{\text{supp}}|} \sum_{s \in S_{\text{supp}}} \frac{1}{\text{rank}(s)}. \quad (14)$$

For each step s_i the $\text{rank}(s)$ is taken of the highest-ranking unique $s_{\text{supp}} \in S_{\text{supp}}$.

2. **Recall@k** measures the cumulative coverage of the supporting sentences by T in top-k results from the downstream BM25 retrieval.

$$\text{Recall@k} = \frac{|\bigcup_{i=1}^T \text{top}_k(s_i) \cap S_{\text{supp}}|}{|S_{\text{supp}}|}. \quad (15)$$

3. **Precision@k** quantifies the *Query Utility*, representing the fraction of generated steps that successfully retrieve at least one supporting sentence within the top k results.

$$\text{Precision@}k = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbf{1}(\text{top}_k(s_i) \cap S_{\text{supp}} \neq \emptyset). \quad (16)$$

4. **F1@k** provides the harmonic mean of Precision and Recall, which serves as the primary metric for retrieval efficiency, as it balances comprehensive coverage against query redundancy, described as:

$$\text{F1@}k = 2 \cdot \frac{\text{Precision@}k \cdot \text{Recall@}k}{\text{Precision@}k + \text{Recall@}k}. \quad (17)$$

Notably, all the scores mentioned above are averaged in order to get the final results.

The dataset used for the evaluation of all models is a modified version of the validation dataset, where for each sample $S = \{Q, C, S_{\text{all}}, S_{\text{supp}}\}$ Q is a question, C is the input context, $S_{\text{all}} = \bigcup D_{\text{all}}$ and S_{supp} is defined as in Section 3.1. D_{all} is the set of all paragraphs provided by the original dataset in relation to the question.

4.2. Baseline

We evaluate against two standard Information Retrieval baselines: Okapi BM25 (Robertson et al., 1994) and Sentence-BERT (Reimers & Gurevych, 2019) (sBERT; *all-MiniLM-L6-v2*). These baselines represent the current solutions to the task of Multi-Hop Retrieval. As both baselines are non-iterative, we treat the question Q as s_i in Equations (14), (15) and (16). To compute the scores of the BM25 baseline, the BM25 algorithm is applied to Q and the benchmark scores are computed based on the retrieved sentences. For the sBERT baseline, Q and S_{all} are encoded into 384-dimensional vectors via sBERT and the $\text{top}_k(s_i)$ are received via the cosine similarity function. The performance of the baseline models on numerical benchmarks can be seen in Table 1.

Both the algorithmic and the neural baselines achieve a comparable level of performance on the proposed task. The evidence suggests that such performance is the result of their non-iterative nature: the models cannot construct multiple s_i , leading to a significant disadvantage on an inherently multi-step task. A small difference between the numerical results of both baselines can be observed, with sBERT achieving superior results.

4.3. Results

The multi-step neural approach to the task of Multi-Hop retrieval significantly outperforms both the non-iterative neu-

Table 1. The Performance of the Baselines as compared to the Main model

METRIC	OKAPI BM25	sBERT	MAIN MODEL
MRR	0.3516	0.3582	0.6145
R@1	0.1973	0.2040	0.5619
R@5	0.5062	0.5139	0.8536
P@1	0.6752	0.7086	0.6960
P@5	0.9481	0.9471	0.9792
F1@1	0.3042	0.3155	0.6083
F1@5	0.6356	0.6430	0.8991

ral baseline (sBERT) and the algorithmic baseline (BM25) (Table 1). It achieves an improvement over sBERT by a factor of **1.72** \times on MRR, **2.75** \times on Recall@1 and **1.93** \times on F1@1. Since MRR shows a score of over 0.60, we can safely conclude that the model’s output often places any answering sentence at ranks 1 or 2. Recall@1 also confirms that 56.2% of the sentences receive first place, compared with 20.4% achieved by the baseline neural approach.

Comparing the Main model and baselines on the Precision@1 and Precision@5 metrics, it is evident that the our model acts as a highly-specialized retrieval planner. By generating concise sub-queries s_i within its output T , it maximizes answer coverage, unlikely to inflate or undermine the step count.

The model’s efficiency is demonstrated by its 5.38M parameters and 20.5MB footprint, along with the average CPU latency of 195.5ms (± 38.3 ms, measured with batch size = 1) per inference. This supports the claim of the model’s sustainability in resource-constrained environments, proving that the task of Multi-Hop Retrieval can be executed comfortably at the micro-parameter limit.

4.4. The Broken Chain Experiment: Analyzing Dependency Chains

To evaluate if the architecture relies on implicit sequential dependencies, we remove the first ground-truth sentence from C and S_{supp} . We assume that a subsequent performance collapse confirms that successfully retrieving s_2^* is dependent upon the successful retrieval of s_1^* . The results achieved after running this test on a modified version of the benchmarking dataset, with the first ground truth sentences being absent from both C and S_{supp} can be observed in Table 2.

The observed collapse in various evaluation metrics provides empirical evidence that the architecture learns sequential logical dependencies. We conclude that for this 5.38M parameter model, successful multi-hop retrieval is conditioned on the presence of the full logical chain.

Counterintuitively, the model under the Broken Chain con-

Table 2. The Broken Chain Experiment

METRIC	BROKEN CHAIN	MAIN MODEL
MRR	0.5132	0.6145
R@1	0.5758	0.5619
R@5	0.8407	0.8536
P@1	0.5354	0.6960
P@5	0.9061	0.9792
F1@1	0.5421	0.6083
F1@5	0.8554	0.8991

dition achieves an increase on Recall@1. This is a statistical artifact of reduced target cardinality ($|S_{\text{supp}}|$) in the ablated environment. Following the Equation (15), it is evident that it was significantly easier for the model under the Broken Chain condition to achieve comparable Recall@1 with the Main Model. If the architecture functioned as a parallel keyword matcher, it would be more resilient to the removal of intermediate facts within the context C . The observed collapse in 6 out of 7 performance metrics confirms that the model’s retrieval is logically autoregressive.

5. Ablation Studies

Ablation studies in Sections 5.1, 5.2 and 5.3 required the training of the models on data preprocessed differently from the method described in 3.2. All models have been trained using the same training schedule as the Main Model, with coverage loss $\lambda = 0.1$ being added from the start of the 6th epoch. We trained the models with the learning rate of 1×10^{-4} until the start of the 12th epoch, where they entered a fine-tuning phase of epochs 12-17 with the learning rate 5×10^{-5} . The reported results show the best-performing epoch of each model. The results for ablations in Sections 5.1 and 5.2 can be found in Table 3.

5.1. Explicit Positional Encoding

To test the hypothesis that the use of EPE tokens is necessary to teach the model to formulate logically dependent outputs, we conduct an ablation study, removing the EPE tokens from the output T and the input context C (Table 3). As output sequence length is a crucial detail in tiny recurrent neural networks, we train two Non-EPE models, different in the average length of each s_i , performing Top-8 and Top-9 IDF compressions, with the Main Model being trained on 9 tokens per s_i (Top-8 IDF + 1 EPE).

Contrary to the hypothesis, the Non-EPE models achieve superior numerical results on all proposed metrics. To assess their ability to construct dependency chains, we perform the Broken Chain experiment on the Non-EPE models (Table 3).

Following the definition of the Broken Chain experiment,

all models experience significant output quality degradation, proving that all models are succeeding in implicitly learning dependency chains within their outputs. This experiment suggests that the ability of the model to learn such dependencies is a property of the architecture and the data-centric approach itself.

Furthermore, it is evident that two problems, namely the *Payload Tax* and *Sequence Tax* are impacting the Main Model. We define the *Payload Tax* as the difference in model’s performance based on the informational density of the outputs. The ratio of *lexical tokens* (tokens available to be used for the downstream BM25) to *EPE tokens* (tokens providing metadata information, not useful for downstream retrieval) in the Main Model can be described as 8 : 1, compared to 9 : 0 of the Non-EPE Top-9 model. As the latter model uses more tokens to encode its output and all of those tokens are used during the secondary retrieval, it achieves results superior to both IDF-8 Non-EPE and the Main Model.

The *Sequence Tax*, which we define as the difference in model’s performance based on the length of each s_i , is also evident. The Non-EPE Top-9 model suffers a decrease in its Precision@1 of 1.5% relative to the Non-EPE Top-8 model. The model with Explicit Positional Encoding performs worst on Precision@1, experiencing an output quality degradation, being under the effect of both *Sequence Tax* and *Payload Tax*, increasing the length of s_i while keeping the amount of lexical tokens.

5.2. Top-8 IDF Compression

We hypothesize that Top-K IDF compression, described in Section 3.2, acts as a counter-mechanism to the *Sequence Tax*, described previously with Equation (2). To assess the importance of Top-K IDF compression, we perform an ablation study, training an EPE and Non-EPE model on uncompressed training data, aiming to compare the performance of models with and without EPE on longer s_i (Table 3).

The EPE Non-IDF model achieves superior performance among the absolute majority of numerical benchmarks, compared to the Main and Non-EPE Non-IDF model. This shows that on higher $|s_i|$, EPE is able to stabilize the model’s performance, focusing its attention on distinct parts of C . The difference in MRR, Recall@1 and Precision@1 between the Non-EPE Non-IDF model and the EPE Non-IDF model suggests that the lower $|s_i|$ grants the Non-EPE Non-IDF model an advantage in generating the first s_i , leading to a higher average precision score. However, the same $|s_i|$ grants the model a significant disadvantage in generating the other $k - 1$ s_i due to the lack of grounding, leading to an inferior performance on Recall@k and MRR.

We additionally run the Broken Chain experiment on EPE Non-IDF and Non-EPE Non-IDF models to compare their

Table 3. Explicit Positional Encoding Ablations and the Broken Chain (BC) Experiments. We define two categories of ablated models: IDF and Non-IDF. Bold is used to mark the best score in one category, and bold italic marks the overall best score per evaluation.

MODEL	MRR	R@1	R@5	P@1	P@5	F1@1	F1@5
<i>Standard Evaluation</i>							
MAIN	0.6145	0.5619	0.8536	0.6960	0.9792	0.6083	0.8991
NON-EPE 8	0.6459	0.6031	0.8776	0.7583	0.9845	0.6557	0.9168
NON-EPE 9	0.6606	0.6225	0.8835	0.7471	0.9825	0.6648	0.9199
EPE NON-IDF	0.6461	0.6034	0.8689	0.7203	0.9749	0.6449	0.9070
NON-EPE NON-IDF	0.6136	0.5717	0.8561	0.7412	0.9789	0.6301	0.8996
<i>The Broken Chain Experiment</i>							
BC MAIN	0.5132	0.5758	0.8407	0.5354	0.9061	0.5421	0.8554
BC NON-EPE 8	0.4567	0.5615	0.8600	0.4933	0.9181	0.5115	0.8723
BC NON-EPE 9	0.4964	0.6193	0.8767	0.5668	0.9140	0.5765	0.8797
BC EPE NON-IDF	0.5236	0.6115	0.8613	0.5679	0.9092	0.5748	0.8696
BC NON-EPE NON-IDF	0.4898	0.5573	0.8429	0.5678	0.9163	0.5449	0.8606

failure modes. The data shows both models experiencing significant degradation of output quality, especially on MRR and Precision@k, which later affects F1@k scores. We observe an increase in numerical performance on Recall@1 for the EPE Non-IDF model, which can be explained with reduced target cardinality, as stated in Section 4.4.

Comparing the failure modes of compressed and uncompressed models on the Broken Chain experiment, it is evident that all tested models experience output quality degradation. This affirms the central hypothesis proposed with the experiment: all tested models depend on the successful retrieval of s_i^* to retrieve the true s_{i+1}^* . This provides empirical evidence that micro-parameter recurrent networks possess the capacity to learn implicit sequential dependency chains when trained on appropriately structured data.

Crucially, the models trained with EPE exhibit a less catastrophic MRR collapse relative to their ungrounded counterparts. This demonstrates that while all tested models implicitly internalize sequential dependencies, EPE provides a stabilizing mechanism, rendering the output more resilient to contextual disruption and logical breaks.

5.3. Sentence-Level Shuffling

In the training data, the majority of supporting sentences are contained in the first few sentences of a given supporting paragraph. To mitigate learned heuristics, we apply sentence-level shuffling to supporting paragraphs in C by removing all sentences $s \in S_{\text{supp}}$ and re-inserting them at novel indices. The aim of this technique is to preserve the original logical flow of the paragraph without exposing the model to heuristic biases of the training data. To assess the contribution of this technique, we remove the shuffling from the training data and train another model, evaluating it on a dataset with shuffling applied (Table 4).

Table 4. The Sentence-Level Shuffling Ablation Study

METRIC	NO SHUFFLING	NO QE	MAIN
MRR	0.4805	0.5988	0.6145
R@1	0.4059	0.5616	0.5619
R@5	0.8141	0.8378	0.8536
P@1	0.5040	0.6621	0.6960
P@5	0.9688	0.9083	0.9792
F1@1	0.4380	0.5978	0.6083
F1@5	0.8675	0.8568	0.8991

The catastrophic difference between the performance of the Main Model and ablated model makes it evident that such shuffling during training is a necessity to prevent learned heuristics.

5.4. Query Expansion (QE)

The downstream retrieval task is highly dependent on the quality of keywords the model outputs. To provide the BM25 algorithm with more relevant keywords for retrieval, we perform Query Expansion on each s_i in T by concatenating s_i with a stopwords-free version of the original query. We hypothesize that such concatenation will lead to higher performance on the Multi-Hop Retrieval task. To evaluate the contribution of this technique, we compare the results of the Main Model with and without such Query Expansion (Table 4).

The data shows decrease in the model’s performance can be seen across all metrics without Query Expansion, specifically a 2.6% decrease in MRR and a 5.1% decrease in Precision@1. Recall@1 and F@1 scores have also decreased following the removal of Query Expansion. These results demonstrate the advantage granted to the model by applying Query Expansion to its s_i .

6. Discussion

6.1. Dependency Chains and the Broken Chain Experiment

Section 4.4 demonstrates that the model’s generation is fundamentally autoregressive: the success of a step s_i in retrieving a target fact is dependent on the successful retrieval of the preceding $i - 1$ facts. This pattern was established among all tested models, as shown in Sections 5.1 and 5.2, suggesting that the capacity to learn these logical dependencies is a property of the recurrent architecture with properly formatted training data.

These findings demonstrate that the construction of such logical dependency chains is not exclusively an emergent property of larger-scale models, and that smaller models, such as the proposed 5.38M architecture, also possess the ability to implicitly learn sequential dependency chains. We attribute these capabilities to the architecture itself, as well as the data preprocessing, especially the stopwords elimination: by offloading the burden of natural language generation, we focus the model’s limited capacity on information-dense keyword sequences, which enables the architecture to maintain a stable logical trajectory during generation.

6.2. Explicit Positional Encoding

Sections 5.1 and 5.2 show that Explicit Positional Encoding introduces non-trivial changes into the model’s performance: it grants the model an advantage of output grounding on longer s_i , and a disadvantage described by the *Payload Tax* and the *Sequence Tax* on shorter s_i . Furthermore, as shown in the Broken Chain experiments in Sections 5.1 and 5.2, the EPE models are more resilient to incomplete and broken logical pathways, making it more likely to achieve superior performance in OOD environments. A more detailed model comparison can be seen in Figure 2.

It is evident that the Non-IDF model with Explicit Positional Encoding achieves a superior performance compared to the Non-IDF Non-EPE model. We therefore characterize EPE tokens as a tool for output grounding on high $|s_i|$, as the most significant performance gain can be observed in long-sequence environments. The use of EPE tokens in short-sequence environments can be characterized strictly as a trade-off: trading a notable performance degradation for OOD resilience.

Figure 2 additionally suggests the existence of an efficiency breaking point: as EPE imposes a disadvantage on short s_i and grants an advantage on extended s_i , there must exist such $|s_i|$, at which the negative impact of the Sequence Tax starts to outweigh the positive impact of the Payload Tax. This theoretical $|s_i|$ serves as the point at which the use of Explicit Positional Encoding starts to grant an advantage to the model.

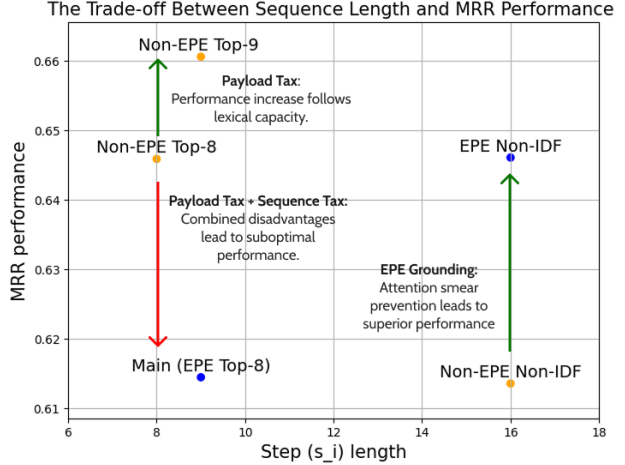


Figure 2. A visual representation of MRR performance in relation to $|s_i|$. In short-sequence environments (left), ungrounded baselines outperform the EPE model due to higher lexical density (Payload Tax). In long-sequence environments (right), EPE tokens act as a structural grounder, mitigating the Sequence Tax and reversing the performance collapse observed in ungrounded architectures.

7. Conclusion

We have presented a data-centric approach to the task of Multi-Hop Retrieval, proposing a small neural network of 5.38M parameters, which occupies 20.5MB of space and achieves an average CPU latency of 195.5ms (± 38.3 ms) per inference. The Main Model outperforms the sBERT baseline by factors of $1.72\times$ (MRR), $2.75\times$ (Recall@1) and $1.93\times$ (F1@1). Through the Broken Chain experiment, we demonstrated that such micro-scale models possess the ability to learn implicit sequential dependencies. Lastly, we demonstrate that Explicit Positional Encoding (EPE) stabilizes the model output in long-sequence environments by providing a discrete grounding mechanism.

Our findings suggest that data-centric methods provide an alternative pathway to the construction of logical dependencies in micro-parameter models through data alignment, as we achieve notable results on MRR and Recall@1 without using the Transformer architecture. The achieved result is a model that is capable of constructing implicit multi-hop dependency chains in a manner that is efficient enough for on-device deployment, ensuring both low latency and user privacy.

While EPE tokens are one way to mitigate the Sequence Tax, future work could focus on applying modern architectures, like xLSTM (Beck et al., 2024), and their exponential gating algorithms to the small-scale task of Multi-Hop Retrieval. Furthermore, future research is needed in order to determine the limits of implicit dependency chains in such micro-parameter architectures.

References

- Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. xlstm: Extended long short-term memory, 2024. URL <https://arxiv.org/abs/2405.04517>.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information, 2017. URL <https://arxiv.org/abs/1607.04606>.
- Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.coling-main.580>.
- Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL <https://aclanthology.org/D19-1410/>.
- Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. Okapi at trec-3. In *Text Retrieval Conference*, 1994. URL <https://api.semanticscholar.org/CorpusID:41563977>.
- See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks, 2017. URL <https://arxiv.org/abs/1704.04368>.
- Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 2022.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Wolfson, T., Geva, M., Gupta, A., Gardner, M., Goldberg, Y., Deutch, D., and Berant, J. Break it down: A question understanding benchmark, 2020. URL <https://arxiv.org/abs/2001.11770>.
- Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., and Chi, E. Least-to-most prompting enables complex reasoning in large language models, 2023. URL <https://arxiv.org/abs/2205.10625>.

A. Supplemental Materials: Qualitative Analysis of Models on Out-of-Distribution Data

To further investigate the behaviour of the models in environments with Out-of-Distribution data, we perform a qualitative analysis on 4 models: Main, Non-EPE Top-8, EPE Non-IDF, Non-EPE Non-IDF.

Below, we present 6 qualitative examples from the Out-of-Distribution dataset. This dataset was created by assembling cleaned paragraphs and, subsequently, their sentences, from the website of TU Wien. In each example, we present the question, context, human-labeled truth and the outputs of several models for comparison. The search space \mathcal{S} was chosen to be all unique sentences present in the corpus. All tokens in model outputs are marked with **green** if they are present in a supporting sentence, **orange** if they come from the most-referenced distractor sentence per s_i and with **red** if they come from another distractor sentence per s_i . This coloring allows us to analyze failure modes: if the model outputs more **orange** than **red**, then it is making largely **logical failures** and not exhibiting catastrophic **attention smear**, and vice versa. You will see retrieved sentences coded with **blue** if they are partial matches to a ground-truth sentence (contain the necessary information, but not the exact words). Some fields may be truncated to preserve correct formatting.

Context (Truncated):

... <P0-S3> it is also an advantage if you enjoy mathematics, as many scientific theories in electrical engineering are expressed using mathematical concepts. <P0-S4> for first-year students, there is a course to learn how to use measurement devices, as well as a mathematics refresher course to ensure everyone starts with the same foundation. <P0-S5> programming is taught during the course of the studies. <P0-S6> in addition to a basic technical understanding, the ability to think “outside the box” is definitely an asset. ... <P2-S2> however, in order to begin studying business informatics, you will need to pass a selection procedure with a ranking test that is explained in detail (process, test examples, material, etc.). <P2-S3> to help you get started with the programme, we also offer a number of events before the official start of the programmes in the winter semester. <SEP> <P3-S0> studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work. <P3-S1> no specific prior knowledge is required. <P3-S2> however, creativity, analytical thinking, and strong spatial awareness are highly beneficial. <P3-S3> equally important are the ability to work in a team, communicate ideas effectively, and engage with complex social, technical, and cultural contexts. ... <SEP> <P4-S0> the environmental engineering programme is designed for anyone who not only wants to understand engineering and the natural sciences, but also to apply them purposefully in order to shape our environment sustainably. ...

Question: Studying what programme requires you to be interested in spatial thinking, yet no prior knowledge?

Human-labeled S_{supp} : studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work. no specific prior knowledge is required.

Main Model: <STEP1> <P0-S4> studying measurement devices mathematics refresher ensure everyone starts <STEP2> <P3-S2> equally analytical thinking strong spatial awareness environment beneficial <EOS>

Non-EPE Top-8: <STEP1> studying important strong course studies <STEP2> studying architecture strong strong understand engineering apply apply shape <STEP3> equally requires communicate test learning offer start programmes contexts <EOS>

EPE Non-IDF: <STEP1> <P0-S3> studying enjoy mathematics scientific theories electrical engineering expressed mathematical concepts <STEP2> <P3-S0> studying architecture requires strong interest spatial thinking design related questions conceptual work <EOS>

Non-EPE Non-IDF: <STEP1> equally students ability bottom material number .). events started test conceptual .). <STEP2> test begin technical understanding ability think box devices asset <STEP3> test begin programme business informatics need pass selection procedure ranking test explained detail process test examples material .). help started programme offer number events programmes start <EOS>

Main Model Retrieved Sentences: for first-year students, there is a course to learn how to use measurement devices, as well as a mathematics refresher course to ensure everyone starts with the same foundation. however, creativity, analytical thinking, and strong spatial awareness are highly beneficial.

Non-EPE Top-8 Retrieved Sentences (Truncated - Supporting Only): studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work. ...

EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): ...studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work.

Non-EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work. ...

sBERT: studying architecture requires a strong interest in spatial thinking, design-related questions, and conceptual work.

Figure 3. The baseline provides only one of the supporting sentences for this question. Both Non-EPE models exhibit significant attention smear, with the Non-EPE Non-IDF exhibiting a catastrophic concentration failure. The Main Model generally makes logical failures, yet mostly focuses its attention on one sentence per s_i . The EPE Non-IDF model successfully keeps its attention on one of the supporting sentences, showing the grounding capabilities of EPE on long sequences.

Context (Truncated):

... <P0-S1> in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria, but they are required to apply for a registration certificate (anmeldebescheinigung) within four months of their arrival in vienna. ... <P1-S9> if you have a permanent residence in austria, you may be eligible to register for student self-insurance with the austrian health insurance fund (ögk), even if you are not an austrian citizen. ... <SEP> <P2-S0> although housing in vienna is generally more affordable than in other major european capitals, the market has become increasingly competitive in recent years, which can make finding accommodation challenging and time-consuming for international students. ... <P2-S2> always be careful which listings you respond to. <P2-S3> vienna has 23 districts, and you will need to officially register your residence once you arrive in the city. ... <P3-S1> for this reason, the selection procedure does not assess programming skills and does not require any prior experience with computers. ...

Question: Once an EU citizen arrives in the city with 23 districts to study at TU Wien, within how many months must they apply for a registration certificate?

Human-labeled S_{supp} : vienna has 23 districts, and you will need to officially register your residence once you arrive in the city. in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria, but they are required to apply for a registration certificate (anmeldebescheinigung) within four months of their arrival in vienna.

Main Model: <STEP1> <P3-S1> careful tu certificate certificate certificate wien wien certificate <STEP2> <P3-S0> eu tu arrives registration wien certificate wien arrives <STEP3> <P3-S1> eu arrives wien arrives citizen eu wien wien <EOS>

Non-EPE Top-8: <STEP1> registration certificate respond registration apply certificate anmeldebescheinigung anmeldebescheinigung vienna <STEP2> vienna residence wien register <STEP3> studying vienna wien wien market competitive finding accommodation challenging consuming <EOS>

EPE Non-IDF: <STEP1> <P3-S0> eu citizen arrives born march 27 1966 american professional basketball player <STEP2> <P3-S1> housing citizen eu wien tu city countries wien wien <STEP3> <P2-S0> careful vienna affordable european european capitals market competitive recent years finding accommodation challenging time consuming international students <EOS>

Non-EPE Non-IDF: <STEP1> general eu citizens citizens liechtenstein norway switzerland need residence permit study austria required apply registration certificate anmeldebescheinigung months arrival vienna <STEP2> permanent eu austria country health <EOS>

Main Model Retrieved Sentences (Truncated - Supporting Only): in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria ...

Non-EPE Top-8 Retrieved Sentences (Truncated - Supporting Only): in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria ...

EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria ...

Non-EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria ...

sBERT: in general, eu citizens, as well as citizens of liechtenstein, norway and switzerland do not need a residence permit to study in austria, but they are required to apply for a registration certificate (anmeldebescheinigung) within four months of their arrival in vienna.

Figure 4. The baseline provides only one of the supporting sentences for this question. The Main Model fails to concentrate its attention on one sentence, while the Non-EPE Top-8 exhibits a stronger attention focus. The EPE Non-IDF shows a catastrophic hallucination, outputting information not present in the input, while the Non-EPE Non-IDF model successfully concentrates its attention on one of the supporting sentences in the first step. The Non-EPE Top-8 and EPE Non-IDF models both retrieve all supporting sentences (truncated), while the Main Model repeats one supporting sentence three times (truncated). The Non-EPE Non-IDF model repeats one supporting sentence two times (truncated).

Context (Truncated):

<P0-S0> the bachelor’s programme “informatik” (hereafter: informatics) focuses on the processing, analysis, and use of data, as well as the development of software and computer systems. . . . <P0-S4> informatics also contributes to making our society safer. . . . <P2-S2> however, to begin a degree in computer engineering, you must pass a selection procedure that includes a ranking test, which is explained in detail (procedure, sample questions, content, etc.). <P2-S3> to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester. <SEP> <P3-S0> there are no special requirements for successfully completing the programme beyond the general requirements for university studies. <P3-S1> you do not need any prior experience in programming or with specific tools. . . . <P3-S3> to help you get off to a good start, we also offer a series of preparatory courses before the official start of the winter semester. . . .

Question: What is the Bachelor’s Programme Informatics focused on, and what kind of preparatory events are offered before the winter semester?

Human-labeled S_{supp} : the bachelor’s programme “informatik” (hereafter: informatics) focuses on the processing, analysis, and use of data, as well as the development of software and computer systems. to help you get off to a good start, we also offer a series of preparatory courses before the official start of the winter semester.

Main Model: <STEP1> <P0-S4> winter contributes theoretical making programming <STEP2> <P3-S1> bachelor bachelor informatics informatics informatics kind informatics informatics <EOS>

Non-EPE Top-8: <STEP1> special requirements successfully completing programme general requirements university studies <STEP2> special requirements completing completing programme requirements requirements studies <STEP3> semester programme informatics specialisations processing .). data preparatory semester <EOS>

EPE Non-IDF: <STEP1> <P3-S0> winter semester born march 18 1952 american football player <STEP2> <P3-S1> informatics programme american rock band formed <STEP3> <P3-S0> programme - informatics semester <EOS>

Non-EPE Non-IDF: <STEP1> special requirements completing programme general requirements university studies <STEP2> begin degree computer engineering pass selection procedure includes ranking test explained detail procedure sample questions content .). help good start offer series preparatory events official start winter semester <STEP3> begin requirements solutions programme programs combines <EOS>

Main Model Retrieved Sentences: to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester. the bachelor’s programme “wirtschaftsinformatik” (hereafter: business informatics) combines informatics and business.

Non-EPE Top-8 Retrieved Sentences: there are no special requirements for successfully completing the programme beyond the general requirements for university studies. there are no special requirements for successfully completing the programme beyond the general requirements for university studies. to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester.

EPE Non-IDF Retrieved Sentences: to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester. to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester. to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester.

Non-EPE Non-IDF Retrieved Sentences: there are no special requirements for successfully completing the programme beyond the general requirements for university studies. to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester. the bachelor’s programme “wirtschaftsinformatik” (hereafter: business informatics) combines informatics and business.

sBERT: to help you get off to a good start, we also offer a series of preparatory events before the official start of the winter semester.

Figure 5. The Main Model fails to concentrate its attention on one sentence, while the Non-EPE Top-8 model concentrates its attention on two sentences in two s_i . EPE Non-IDF and Non-EPE Non-IDF follow the same pattern. All models retrieve only one semantically close sentence to the supporting sentences.

Context (Truncated):

<P0-S0> a master's programme is an accredited programme of study which serves to enhance and supplement students' scientific/academic training. ... <P0-S4> although it is called a master's programme, students are granted the title of diplomingenieur (dipl.-ing.) <P0-S5> upon successful completion of the programme. <P0-S6> students will receive certificates for this academic title in both german and english. <P0-S7> the english version confirms that the title of diplomingenieur is equivalent to a master of science (msc). ... <SEP> <P1-S0> at tu wien, after completing a bachelor's degree in computer engineering, you can enrol in one of the following master's programmes to specialise further: master's programme automation and robotic systems (german-taught) ... <P1-S2> computer engineering at tu wien: professor explains! <SEP> <P2-S0> at tu wien, after completing a bachelor's degree in business informatics, you can enrol in one of the following english-taught master's programmes to specialise further: master's programme biomedical engineering (interfaculty) master's programme business informatics master's programme computational science and engineering (interfaculty) master's programme data science master's programme digital civil engineering science (interfaculty) master's programme embedded computing systems <P3-S1> data protection declaration. . . .

Question: What is the general purpose of Master's programmes at TU Wien and What academic title is awarded upon their completion?

Human-labeled S_{supp} : a master's programme is an accredited programme of study which serves to enhance and supplement students' scientific/academic training. although it is called a master's programme, students are granted the title of diplomingenieur (dipl.-ing.)

Main Model: <STEP1> <P3-S1> data tu tu <STEP2> <P2-S0> tu completing completing enrol enrol enrol biomedical biomedical <STEP3> <P2-S0> tu wien informatics informatics specialise biomedical interfaculty <EOS>

Non-EPE Top-8: <STEP1> data master declaration computer computer wien professor students explains <STEP2> wien enrol enrol specialise interfaculty interfaculty computing computing computing <STEP3> computing programme interfaculty enrol interfaculty biomedical interfaculty informatics computational <EOS>

EPE Non-IDF: <STEP1> <P2-S0> tu wien completing bachelor degree business informatics enrol following english taught master programmes specialise master programme biomedical engineering interfaculty <STEP2> <P3-S1> tu master purpose master master master academic master master <EOS>

Non-EPE Non-IDF: : <STEP1> tu wien completing bachelor degree business informatics enrol following english taught master programmes specialise master programme biomedical engineering interfaculty master programme business informatics master programme computational science <STEP2> english wien confirms bachelor english master master master tu master master <STEP3> tu wien master bachelor possible computational master english automation master master automation master systems <EOS>

Main Model Retrieved Sentences (Truncated): master's programmes at tu wien generally last four semesters, corresponding to 120 ect. . . .

Non-EPE Top-8 Retrieved Sentences (Truncated): engineering at tu wien: professor explains! . . .

EPE Non-IDF Retrieved Sentences: at tu wien, after completing a bachelor's degree in business informatics, you can enrol in one of the following english-taught master's programmes to specialise further: master's programme biomedical engineering (interfaculty) master's programme business informatics master's programme computational science and engineering (interfaculty) . . .

Non-EPE Non-IDF Retrieved Sentences (Truncated): . . . the english version confirms that the title of diplomingenieur is equivalent to a master of science (msc). . . .

sBERT: upon successful completion of the programme.

Figure 6. This specific example poses as a "messy" one: it contains multiple distractor sentences of extreme length. No models retrieve supporting sentences and it is evident that all models struggle heavily with repetition. The models mostly focus their attention on the logn distractor sentences, which they do successfully, making a logical mistake rather than an attention smear one in the majority of the cases. The Non-EPE Non-IDF model exhibits attention smear in the third step.

Context (Truncated):

<P0-S0> in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research. . . . funds raised related to eu, fwf < . . . <P0-S3> according to the principles behind university research at tu wien, the focus of attention is on societal benefit from research projects, not the possible commercial exploitation of the research findings. . . . <P1-S2> in the area of basic research, researchers from tu wien have raised countless grants from the european research council (erc). . . . <SEP> <P2-S0> in the area of basic research, the fwf (austrian research fund) and calls on the vienna science an technology fund (wwtf) are of particular note. <P2-S1> in addition to successes with the wwtf funding channel and individual fwf projects, tu wien – . . . (sfb), national research networks (nfn), doctoral programmes (dk) . . . (peek). . . . <P2-S3> in the field of applied research, tu wien is represented in the ffg (austrian research promotion agency) as part of the comet programme and is the leading austrian university in the cd laboratories run by the christian doppler forschungsgesellschaft (cdg - christian doppler research association). <SEP> <P3-S0> all international students are required to have valid health insurance while studying at tu wien. . . .

Question: Out of the approximately EUR 91 million attracted in 2020 for externally financed project research, funding comes from national funds like FWF and WWTF, but from which European council have researchers at TU Wien raised countless grants in the area of basic research?

Human-labeled S_{supp} : in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research. in the area of basic research, the fwf (austrian research fund) and calls on the vienna science an technology fund (wwtf) are of particular note. in the area of basic research, researchers from tu wien have raised countless grants from the european research council (erc).

Main Model: <STEP1> <P3-S1> eu funds fwf 2020 externally externally wwtf fwf <STEP2> <P3-S0> basic funds researchers researchers researchers basic basic basic <STEP3> <P3-S1> eu tu wwtf researchers wien countless grants grants <EOS>

Non-EPE Top-8: <STEP1> principles research tu wien focus attention societal benefit projects projects possible findings findings <STEP2> eu 91 eu fwf fund ffg doppler forschungsgesellschaft cdg doppler <STEP3> peek wien wien comet fwf sfb nfn dk peek <EOS>

EPE Non-IDF: <STEP1> <P3-S1> european council financed research funding fwf wwtf <STEP2> <P3-S1> european researchers <STEP3> funding european research tu national researchers researchers tu . raised <STEP4> raised 2020 council 2020 researchers funds tu . european wien research research research basic <EOS>

Non-EPE Non-IDF: : <STEP1> according principles university research tu wien focus attention societal benefit research projects possible commercial exploitation research findings <STEP2> international tu required funding health health studying successful national funds funds <EOS>

Main Model Retrieved Sentences (Truncated - Supporting Only): in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research. . . .

Non-EPE Top-8 Retrieved Sentences (Truncated - Supporting Only): . . . in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research. . . .

EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research. in the area of basic research, researchers from tu wien have raised countless grants from the european research council (erc). . . .

Non-EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): . . . in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research.

sBERT: in 2020, tu wien attracted approx. eur 91 million of funding from national and international sources for externally financed project research.

Figure 7. The baseline retrieves only one supporting sentence for this sample. The Main Model exhibits repetition, but retrieves 2 out of 3 supporting sentences (truncated), which is also the case with the EPE Non-IDF model. Both Non-EPE models concentrate their attention on distractor sentences, successfully retrieving only 1 supporting sentence.

Context (Truncated):

<P0-S0> according to a 2017 study conducted by the austrian institute of strategic analysis (isa) and meinungsraum.at, and led by political scientist peter filzmaier, almost half of those asked had experienced cyber bullying and hate posts in the past, with that figure rising to two thirds among the younger respondents. <P0-S1> those affected stated they had mostly ignored these kinds of posts, but that they often felt helpless, while three quarters did not know where to turn for help. ... <P2-S3> aside from gaining professional qualifications, students also develop their social skills and personalities while at university. ...

Question: While a 2017 study by the Austrian Institute of Strategic Analysis (ISA) showed many experienced cyber bullying, and the organisation ZARA opened an advice centre for help, which internal committee at TU Wien serves as a point of contact for students and professors seeking help regarding equal opportunities?

Human-labeled S_{supp} : according to a 2017 study conducted by the austrian institute of strategic analysis (isa) and meinungsraum.at, and led by political scientist peter filzmaier, almost half of those asked had experienced cyber bullying and hate posts in the past, with that figure rising to two thirds among the younger respondents. within tu wien, any students, professors or project workers seeking help – regardless of their hierarchical level within the university – have the following points of contact available to them: the committee on equal treatment (akg) (working group on equal opportunities), the student psychological counselling service, the occupational psychology service and the works council.

Main Model: <STEP1> <P0-S0> 2017 isa meinungsraum filzmaier cyber bullying posts thirds <STEP2> <P3-S0> master internal wien wien wien professors professors professors <EOS>

Non-EPE Top-8: <STEP1> 2017 gaining professional qualifications students develop social skills personalities university <STEP2> scientist asked asked austrian institute supplement analysis students <STEP3> tu asked students austrian institute strategic analysis isa meinungsraum <EOS>

EPE Non-IDF: <STEP1> <P0-S0> according 2017 study conducted austrian institute strategic analysis isa meinungsraum . led political political scientist peter filzmaier half asked experienced cyber bullying hate posts past figure rising thirds younger respondents <STEP2> <P0-S1> political stated ignored kinds posts felt helpless quarters know turn help <STEP3> <P3-S0> felt helpless filzmaier born july 19 1977 american singer songwriter <EOS>

Non-EPE Non-IDF: : <STEP1> led political scientist peter filzmaier half asked experienced cyber bullying hate posts past figure rising thirds younger respondents <STEP2> following wien students professors professors born helpless january 3 1964 american isa isa <EOS>

Main Model Retrieved Sentences (Truncated - Supporting Only):: ...within tu wien, any students, professors or project workers seeking help – regardless of their hierarchical level within the university ...

Non-EPE Top-8 Retrieved Sentences (Truncated - Supporting Only): ... according to a 2017 study conducted ...

EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): according to a 2017 study conducted ...

Non-EPE Non-IDF Retrieved Sentences (Truncated - Supporting Only): according to a 2017 study conducted ...

sBERT: according to a 2017 study conducted by the austrian institute of strategic analysis (isa) and meinungsraum.at, and led by political scientist peter filzmaier, almost half of those asked had experienced cyber bullying and hate posts in the past, with that figure rising to two thirds among the younger respondents.

Figure 8. The baseline retrieves only one supporting sentence. The Main Model succeeds in focusing its attention a supporting step, as do EPE Non-IDF and Non-EPE Non-IDF. The Non-EPE Top-8 model focuses its attention on a distractor sentence. The Main Model retrieves 2 out of 3 supporting sentences (truncated), while other models retrieve 1 (truncated). The EPE and Non-EPE Non-IDF models exhibit significant hallucinations in their last steps.