Hybrid Autoregressive Inference for Scalable Multi-Hop Explanation Regeneration

Marco Valentino[1,2]**, Mokanarangan Thayaparan**[1,2]**,**
Deborah Ferreira[1]**, André Freitas**[1,2]

[1] University of Manchester, United Kingdom
[2] Idiap Research Institute, Switzerland
marco.valentino@manchester.ac.uk

Abstract

Regenerating natural language explanations in the scientific domain has been proposed as a benchmark to evaluate complex multi-hop and explainable inference. In this context, large language models can achieve state-of-the-art performance when employed as cross-encoder architectures and fine-tuned on human-annotated explanations. However, while much attention has been devoted to the quality of the explanations, the problem of performing inference efficiently is largely under studied. Cross-encoders, in fact, are intrinsically not scalable, possessing limited applicability to real-world scenarios that require inference on massive facts banks. To enable complex multi-hop reasoning at scale, this paper focuses on bi-encoder architectures, investigating the problem of scientific explanation regeneration at the intersection of dense and sparse models. Specifically, we present **SCAR** (for Scalable Autoregressive Inference), a hybrid framework that iteratively combines a Transformer-based bi-encoder with a sparse model of explanatory power, designed to leverage explicit inference patterns in the explanations. Our experiments demonstrate that the hybrid framework significantly outperforms previous sparse models, achieving performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster and scalable to corpora of millions of facts. Further analyses on semantic drift and multi-hop question answering reveal that the proposed hybridisation boosts the quality of the most challenging explanations, contributing to improved performance on downstream inference tasks.

Introduction

Explanation regeneration is the task of retrieving and combining two or more facts from an external knowledge source to reconstruct the evidence supporting a certain natural language hypothesis (Xie et al. 2020; Jansen et al. 2018). As such, this task represents a crucial intermediate step for the development and evaluation of explainable Natural Language Inference (NLI) models (Wiegreffe and Marasović 2021; Thayaparan, Valentino, and Freitas 2020). In particular, explanation regeneration on science questions has been proposed as a benchmark for complex multi-hop and explainable inference (Jansen and Ustalov 2019; Jansen et al. 2016). Scientific explanations, in fact, require the articulation and integration of commonsense and scientific knowl-

edge for the construction of long explanatory reasoning chains, making multi-hop inference particularly challenging for existing models (Clark et al. 2018; Khot et al. 2019). Moreover, since the structure of scientific explanations cannot be derived from the decomposition of the questions, the task requires the encoding of complex abstraction and grounding mechanisms for the identification of relevant explanatory knowledge (Valentino, Thayaparan, and Freitas 2021; Thayaparan, Valentino, and Freitas 2021).

To tackle these challenges, existing neural approaches leverage the power of the self-attention mechanism in Transformers (Devlin et al. 2019; Vaswani et al. 2017), training sequence classification models (i.e., cross-encoders) on annotated explanations to compose relevant explanatory chains (Cartuyvels, Spinks, and Moens 2020; Das et al. 2019; Chia, Witteveen, and Andrews 2019; Banerjee 2019). While Transformers achieve state-of-the-art performance, cross-encoders make multi-hop inference intrinsically inefficient and not scalable to large corpora. The cross-encoder architecture, in fact, does not allow for the construction of dense indexes to cache the encoded explanatory sentences, resulting in prohibitively slow inference time for real-world applications (Humeau et al. 2019).

In this work, we are interested in developing new mechanisms to enable scientific explanation regeneration at scale, optimising, at the same time, quality of the explanations and inference time. To this end, we focus our attention on bi-encoders (or siamese networks) (Reimers and Gurevych 2019), which allow for efficient inference via Maximum Inner Product Search (MIPS) (Johnson, Douze, and Jégou 2019). Given the complexity of multi-hop reasoning in the scientific domain, bi-encoders are expected to suffer from a drastic drop in performance since the self-attention mechanism cannot be leveraged to learn meaningful compositions of explanatory chains. However, we hypothesise that the orchestration of latent and explicit patterns emerging in natural language explanations can improve the quality of the inference while preserving the scalability intrinsic in bi-encoders.

To validate this hypothesis, we present **SCAR** (for Scalable Autoregressive Inference), a hybrid architecture that combines a Transformer-based bi-encoder with a sparse model of explanatory power, designed to capture explicit inference patterns in corpora of scientific explanations. Specifically, SCAR integrates sparse and dense encoders to define
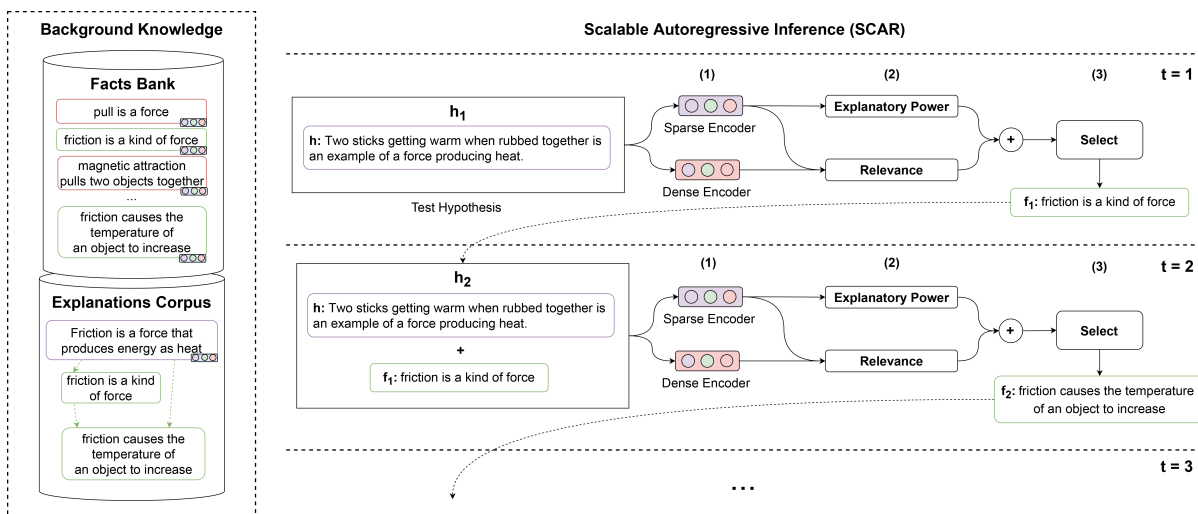
Figure 1: We propose a hybrid, scalable explanation regeneration model that performs inference autoregressively. At each time-step t, we perform inference integrating sparse and dense bi-encoders (1) to compute relevance and explanatory power of sentences in the Facts Bank (2) and expand the explanation (3). The relevance of a fact at time-step t is conditioned on the partial explanation constructed at time $t-1$, while the explanatory power is estimated leveraging inference patterns emerging across similar hypotheses in the Explanations Corpus.

a joint model of relevance and explanatory power and perform inference in an iterative fashion, conditioning the probability of selecting a fact at time-step t on the partial explanation constructed at time-step $t-1$ (Fig. 1). We performed an extensive evaluation on the WorldTree corpus (Jansen and Ustalov 2019), presenting the following conclusions:

1. The hybrid framework based on bi-encoders significantly outperforms existing sparse models, achieving performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster.

2. We study the impact of the hybridisation on semantic drift, showing that it makes SCAR more robust in the construction of challenging explanations requiring long reasoning chains.

3. We investigate the applicability of SCAR on multi-hop question answering without additional training, demonstrating improved accuracy and robustness when performing explainable inference iteratively.

4. We perform a scalability analysis by gradually expanding the adopted facts bank, showing that SCAR can scale to corpora containing millions of facts.

To the best of our knowledge, we are the first to propose a hybrid autoregressive model for complex multi-hop inference in the scientific domain, demonstrating its efficacy for explanation regeneration at scale.

Multi-Hop Explanation Regeneration

Given a scientific hypothesis h expressed in natural language (e.g., *"Two sticks getting warm when rubbed together is an example of a force producing heat"*), the task of explanation regeneration consists in reconstructing the evidence supporting h, composing a sequence of atomic sentences

$E_{seq} = f_1, \ldots, f_n$ from external corpora (e.g., f_1:*"friction is a kind of force"*; f_2:*"friction causes the temperature of an object to increase"*). Explanation regeneration can be framed as a multi-hop abductive inference problem, where the goal is to construct the best explanation supporting a given natural language statement adopting multiple retrieval steps. To learn to regenerate scientific explanations, a recent line of research relies on explanation-centred corpora such as Worldtree (Jansen and Ustalov 2019), which are typically composed of two distinct knowledge sources (Fig. 1):

1. A *facts bank* of individual commonsense and scientific sentences including the knowledge necessary to construct explanations for scientific hypotheses.

2. An *explanations corpus* consisting of true hypotheses and natural language explanations composed of sentences from the facts bank.

Hybrid Autoregressive Inference

To model the multi-hop nature of scientific explanations, we propose a hybrid architecture that performs inference autoregressively (Fig. 1). Specifically, we model the probability of composing an explanation sequence $E_{seq} = f_1, \ldots, f_n$ for a certain hypothesis h using the following formulation:

$$P(E_{seq}|h) = \prod_{t=1}^{n} P(f_t|h, f_1, \ldots f_{t-1}) \quad (1)$$

where n is the maximum number of inference steps and f_t represents a fact retrieved at time-step t from the facts bank. We implement the model recursively by updating the hypothesis h at each time-step t, concatenating it with the partial explanation constructed at time step $t-1$:

$$h_t = g(h, f_1, \ldots f_{t-1}) \quad (2)$$

where $g(\cdot)$ represents the string concatenation function. The probability $P(f_t|h_t)$ is then approximated via an explanatory scoring function $es(\cdot)$ that jointly models *relevance* and *explanatory power* as:

$$es(f_t, h_t) = \lambda \cdot r(f_t, h_t) + (1 - \lambda) \cdot pw(f_t, h) \qquad (3)$$

where $r(\cdot)$ represents the relevance of f_t at time step t, while $pw(\cdot)$ represents the explanatory power of f_t. As shown in recent work (Valentino, Thayaparan, and Freitas 2021), scientific explanations are composed of abstract sentences describing underlying explanatory laws and regularities that are frequently reused to explain a large set of hypotheses. To leverage this feature during inference, we measure the explanatory power $pw(\cdot)$ of a fact as the extent to which it explains similar hypotheses in the explanations corpus. The relevance $r(\cdot)$ is computed through a hybrid model that combines a sparse $s(\cdot)$ and a dense $d(\cdot)$ sentence encoder:

$$r(f_t, h_t) = sim(s(f_t), s(h_t)) + sim(d(f_t), d(h_t)) \qquad (4)$$

with $sim(\cdot)$ representing the cosine similarity between two vectors. In our experiments, we adopt BM25 (Robertson, Zaragoza et al. 2009) as a sparse encoder, while Sentence-BERT (Reimers and Gurevych 2019) is adopted to train the dense encoder $d(\cdot)$.

Explanatory Power

Recent work have shown that *explicit explanatory patterns* emerge in corpora of natural language explanations (Valentino, Thayaparan, and Freitas 2021) – i.e., facts describing scientific laws and regularities (i.e., laws such as gravity, or friction that explain a large variety of phenomena) are frequently reused to explain similar scientific hypotheses. These patterns can be leveraged to define a computational model of explanatory power for estimating the relevance of abstract scientific facts, providing a framework for efficient explainable inference in the scientific domain. Given a test hypothesis h, a sentence encoder $s(\cdot)$, and a corpus of scientific explanations, the explanatory power of a generic fact f_i can be estimated by analysing explanations for similar hypotheses in the corpus:

$$pw(f_i, h) = \sum_{h_k \in kNN(h)}^{K} sim(s(h), s(h_k)) \cdot 1_{E_k}(f_i) \qquad (5)$$

$$1_{E_k}(f_i) = \begin{cases} 1 & \text{if } f_i \in E_k \\ 0 & \text{if } f_i \notin E_k \end{cases} \qquad (6)$$

where $kNN(h) = \{h_1, \ldots, h_K\}$ represents a list of hypotheses retrieved according to the similarity $sim(\cdot)$ between the embeddings $s(h)$ and $s(h_k)$, and $1_{E_k}(\cdot)$ is the indicator function verifying whether f_i is part of the explanation E_k for the hypothesis h_k. Specifically, the more a fact f_i is reused for explaining hypotheses that are similar to h in the corpus, the higher its explanatory power. In this work, we hypothesise that this model can be integrated within a hybrid framework based on dense and sparse encoders, improving inference performance while preserving scalability. In our experiments, we adopt BM25 similarity between hypotheses to compute the explanatory power efficiently.

Dense Bi-encoder

To learn a dense encoder $d(\cdot)$, we fine-tune a Sentence-BERT model using a bi-encoder architecture (Reimers and Gurevych 2019). The bi-encoder adopts a siamese network to learn a joint embedding space for hypotheses and facts in the facts bank. Following Sentence-BERT, we obtain fixed sized sentence embeddings by adding a mean-pooling operation to the output vectors of BERT (Devlin et al. 2019). We employ a unique BERT model with shared parameters to learn a sentence encoder $d(\cdot)$ for both facts and hypotheses. At the cost of sacrificing the performance gain resulting from self-attention, the bi-encoder allows for efficient multi-hop inference through Maximum Inner Product Search (MIPS). To enable scalability, we construct an index of dense embeddings for the whole facts bank. To this end, we adopt the approximated inner product search index (IndexIVFFlat) in FAISS (Johnson, Douze, and Jégou 2019).

Training. The bi-encoder is fine-tuned on inference chains extracted from annotated explanations in the WorldTree corpus (Jansen and Ustalov 2019). Since the facts in the annotated explanations are not sorted, to train the model autoregressively, we first transform the explanations into sequences of facts sorting them in decreasing order of BM25 similarity with the hypothesis. We adopt BM25 since the facts that share less terms with the hypothesis tend to require more iterations and inference steps to be retrieved. Subsequently, given a training hypothesis h and an explanation sequence $E_{seq} = f_1, \ldots, f_n$, we derive n positive example tuples (h_t, f_t), one for each fact $f_t \in E_{seq}$, using $h_t = g(h, f_1, \ldots, f_{t-1})$ as hypothesis. To make the model robust to distracting information, we construct a set of negative examples for each tuple (h_t, f_t) retrieving the top most similar facts to f_t that are not part of the explanation. We found that the best results are obtained using 5 negative examples for each positive tuple. We use the constructed training set and the siamese network to fine-tune the encoder via contrastive loss (Hadsell, Chopra, and LeCun 2006), which has been demonstrated to be effective for learning robust dense representations.

Multi-Hop Inference

At each time-step t during inference time, we encode the concatenation of hypothesis and partial explanation h_t using the dense (Sentence-BERT) and sparse (BM25) encoders separately. Subsequently, we adopt the vectors representing h_t to compute the relevance score $r(\cdot)$ of the sentences in the facts bank (Equation 4). In parallel, the sparse representation (BM25) of the hypothesis h is adopted to retrieve the explanations for the top K similar hypotheses in the explanation corpus and compute the explanatory power $pw(\cdot)$ of each fact (Equation 5). Finally, relevance and explanatory power are combined to compute the explanatory scores $es(\cdot)$ (Equation 3) and select the top candidate fact f_t from the facts bank to expand the explanation at time-step t. After t_{max} steps, we rank the remaining facts considering the explanation constructed a time-step t_{max}.

Model	Approach Description	MAP
Dense Models (Cross-encoders)		
Cartuyvels, Spinks, and Moens (2020)	Autoregressive BERT	**57.07**
Das et al. (2019)	BERT path-ranking + single fact ensemble	56.25
Das et al. (2019)	BERT single fact	55.74
Das et al. (2019)	BERT path-ranking	53.13
Chia, Witteveen, and Andrews (2019)	BERT re-ranking with gold IR scores	49.45
Banerjee (2019)	BERT iterative re-ranking	41.30
Sparse Models		
Valentino, Thayaparan, and Freitas (2021)	Unification-based Reconstruction	**50.83**
Chia, Witteveen, and Andrews (2019)	Iterative BM25	45.76
BM25 (Robertson, Zaragoza et al. 2009)	BM25 Relevance Score	43.01
TF-IDF	TF-IDF Relevance Score	39.42
Hybrid Models (Bi-encoders)		
SCAR	Scalable Autoregressive Inference	**56.22**

Table 1: Results on the test-set and comparison with previous approaches. SCAR significantly outperforms all the sparse models and obtains comparable results with state-of-the-art cross-encoders.

Empirical Evaluation

We perform an extensive evaluation on the WorldTree corpus adopting the dataset released for the shared task on multi-hop explanation regeneration[1] (Jansen and Ustalov 2019), where a diverse set of sparse and dense models have been evaluated. WorldTree is a subset of the ARC corpus (Clark et al. 2018) that consists of multiple-choice science questions annotated with natural language explanations supporting the correct answers. The explanations in WorldTree contain an average of six facts and as many as 16, requiring challenging multi-hop inference to be regenerated. The WorldTree corpus provides a held-out test-set consisting of 1,240 science questions with masked explanations where we run the main experiment and comparison with published approaches. To run our experiments, we first transform each question and correct answer pair into a hypothesis following the methodology described in (Demszky, Guu, and Liang 2018). We adopt explanations and hypotheses in the training-set ($\approx 1,000$) for training the dense encoder and computing the explanatory power for unseen hypotheses at inference time. We adopt `bert-base-uncased` (Devlin et al. 2019) as a dense encoder to perform a fair comparison with existing cross-encoders employing the same model. The best results on explanation regeneration are obtained when running SCAR for 4 inference steps (additional details in Ablation Studies). In line with the shared task, the performance of the system is evaluated through the Mean Average Precision (MAP) of the produced ranking of facts with respect to the gold explanations in WorldTree. Implementation and pre-trained models adopted for the experiments are available online[2].

[1]https://github.com/umanlp/tg2019task

[2]https://github.com/ai-systems/hybrid_autoregressive_inference

Explanation Regeneration

Table 1 reports the results achieved by our best model on the explanation regeneration task together with a comparison with previously published approaches. Specifically, we compare our hybrid framework based on bi-encoders with a variety of sparse and dense retrieval models. Overall, we found that SCAR significantly outperforms all the considered sparse models (+5.39 MAP compared to Valentino, Thayaparan, and Freitas (2021)), obtaining, at the same time, comparable results with the state-of-the-art cross-encoder (-0.85 MAP compared to Cartuyvels, Spinks, and Moens (2020)). The following paragraphs provide a detailed comparison with previous work.

Dense Models. As illustrated in Table 1, all the considered dense models employ BERT (Devlin et al. 2019) as a cross-encoder architecture. The state-of-the-art model proposed by Cartuyvels, Spinks, and Moens (2020) adopts an autoregressive formulation similar to SCAR. However, the use of cross-encoders makes the model computationally expensive and intrinsically not scalable. Due to the complexity of cross-encoders, in fact, the model can only be applied for re-ranking a small set of candidate facts at each iteration, which are retrieved using a pre-filtering step based on TF-IDF. In contrast, we found that the use of a hybrid model allows achieving comparable performance without cross-attention and pre-filtering step (-0.85 MAP), making SCAR approximately 50 times faster (Table 2). The second-best dense approach employs an ensemble of two BERT models (Das et al. 2019). A first BERT model is trained to predict the relevance of each fact individually given a certain hypothesis. A second BERT model is adopted to re-rank a set of two-hops inference chains constructed via TF-IDF. The use of two BERT models in parallel, however, makes the approach computationally exhaustive. We observe that SCAR can achieve similar performance with the use of a single BERT bi-encoder, outperforming each individual sub-component

Model	MAP ↑	Time (s/q) ↓
Autoregressive BERT	**57.07**	9.6
BERT single fact	55.74	18.4
BERT path-ranking	53.13	31.8
SCAR	56.22 (98.5%)	**0.19** (×50.5)

Table 2: Detailed comparison with BERT cross-encoders on the test-set in terms of Mean Average Precision (MAP) and inference time (seconds per question).

in the ensemble with a drastic improvement in efficiency (SCAR is 96.8 times and 167.4 times faster respectively, see Table 2). The remaining dense models (Chia, Witteveen, and Andrews 2019; Banerjee 2019) adopt BERT-based cross-encoders to re-rank the list of candidate facts retrieved using sparse Information Retrieval (IR) techniques. As illustrated in Table 1, SCAR outperforms these approaches by a large margin (+6.77 and +14.92 MAP).

Sparse Models. We compare SCAR with sparse models presented on the explanation regeneration task. We observe that SCAR significantly outperforms the Unification-based Reconstruction model proposed by Valentino, Thayaparan, and Freitas (2021) (+5.39 MAP), which employs a model of explanatory power in combination with BM25, but without dense representation and autoregressive inference. These results confirm the contribution of the hybrid model together with the importance of modelling explanation regeneration in a iterative fashion. In addition, we compare SCAR with the model proposed by Chia, Witteveen, and Andrews (2019) which adopts BM25 vectors to retrieve facts iteratively. We found that SCAR can dramatically improve the performance of this model by 10.46 MAP points. Finally, we measure the performance of standalone sparse baselines for a sanity check, showing that SCAR can significantly outperform BM25 and TFIDF (+13.21 and +16.8 MAP respectively), while preserving a similar scalability (Fig. 3).

Inference Time

We performed additional experiments to evaluate the efficiency of SCAR and contrast it with state-of-the-art cross-encoders. To this end, we run SCAR on 1 16GB Nvidia Tesla P100 GPU and compare the inference time with that of dense models executed on the same infrastructure (Cartuyvels, Spinks, and Moens 2020). Table 2 reports MAP and execution time in terms of seconds per question. As evident from the table, we found that SCAR is 50.5 times faster than the state-of-the-art cross-encoder (Cartuyvels, Spinks, and Moens 2020), while achieving 98.5% of its performance. Moreover, when compared to the individual BERT models proposed by Das et al. (2019), SCAR is able to achieve better MAP score (+0.48 and +3.09), increasing even more the gap in terms of inference time (96.8 and 167.4 times faster).

Ablation Studies

In order to understand how the different components of SCAR complement each other, we carried out distinct ab-

Model	t_{max}	MAP	Time (s/q)
Bi-encoder	1	41.98	0.04
	2	**42.17**	0.08
	3	39.97	0.12
	4	38.34	0.16
	5	37.24	0.19
	6	36.64	0.24
BM25	1	45.99	0.02
	2	47.77	0.04
	3	**48.35**	0.05
	4	48.06	0.07
	5	47.97	0.09
	6	47.66	0.11
Bi-encoder + BM25	1	51.53	0.05
	2	54.52	0.08
	3	55.65	0.14
	4	56.07	0.18
	5	**56.24**	0.22
	6	55.87	0.27
SCAR	1	57.10	0.06
	2	59.20	0.10
	3	59.73	0.15
	4	**60.28**	0.19
	5	59.79	0.24
	6	59.36	0.29

Table 3: Ablation study on the dev-set, where t_{max} represents the maximum number of iterations adopted to regenerate the explanations, and (s/q) is the inference time.

lation studies. The studies are performed on the dev-set since the explanations on the test-set are masked. Table 3 presents the results on explanation regeneration for different ablations of SCAR adopting an increasing number of iterations t_{max} for the inference. The results show how the performance improves as we combine sparse and dense models, with a decisive contribution coming from each individual sub-component. Specifically, considering the best results obtained in each case, we observe that SCAR achieves an improvement of 18.11 MAP over the dense component (Bi-encoder) and 11.93 MAP when compared to the sparse model (BM25). Moreover, the ablation demonstrates the fundamental role of the explanatory power model in achieving the final performance, which leads to an improvement of 4.04 MAP over the Bi-encoder + BM25 model (Equation 3). Overall, we notice that performing inference iteratively is beneficial to the performance across the different components. We observe that the improvement is more prominent when comparing $t_{max} = 1$ (only using the hypothesis) with $t_{max} = 2$ (using hypothesis and first fact), highlighting the central significance of the first retrieved fact to support the complete regeneration process. Except for the Bi-encoder, the experiments demonstrate a slight improvement when adding more iterations to the process, obtaining the best results for SCAR using a total of 4 inference steps. We notice that the best performing component in terms of inference time is BM25. The integration with the dense model,
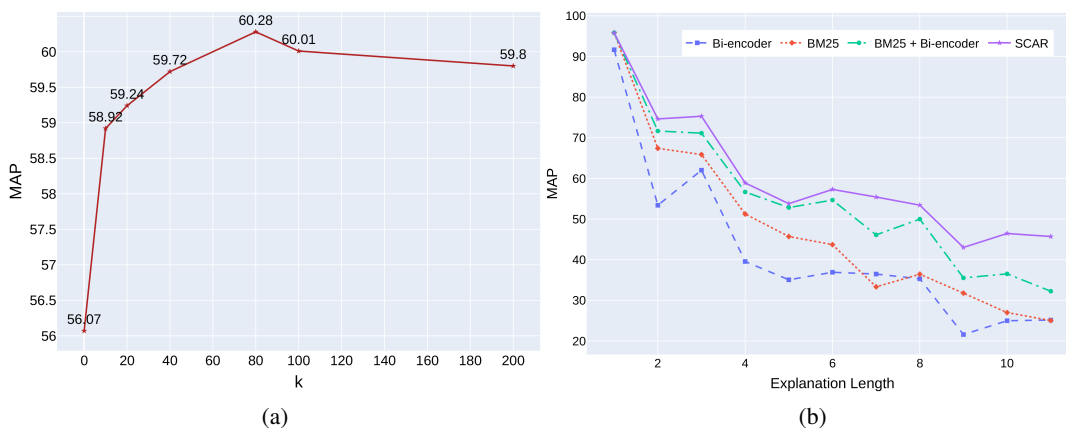
Figure 2: (a) Impact of increasing the number of similar hypotheses K to estimate the explanatory power (Equation 5). (b) Performance considering hypotheses with gold explanations including an increasing number of facts.

in fact, slightly increases the inference time, yet leading to a decisive improvement in terms of MAP score. Even with the overhead caused by the Bi-encoder, however, SCAR can still perform inference in less than half a second per question, a feature that demonstrates the scalability of the approach with respect to the number of iterations. Finally, we evaluate the impact of the explanatory power model by considering a larger set of training hypotheses for its implementation (Figure 2a). To this end, we compare the performance across different configurations with increasing values of K in Equation 5. The results demonstrate the positive impact of the explanatory power model on the inference, with a rapid increase of MAP peaking at $K = 80$. After reaching this value, we observe that considering additional hypotheses in the corpus has little impact on the model's performance.

Semantic Drift

Recent work have shown that the regeneration of scientific explanations is particularly challenging for multi-hop inference models as it can lead to a phenomenon known as *semantic drift* – i.e., the composition of spurious inference chains caused by the tendency of drifting away from the original context in the hypothesis (Khashabi et al. 2019; Xie et al. 2020; Jansen and Ustalov 2019; Thayaparan, Valentino, and Freitas 2020). In general, the larger the size of the explanation, the higher the probability of semantic drift. Therefore, it is particularly important to evaluate and compare the robustness of multi-hop inference models on hypotheses requiring long explanations. To this end, we present a study of semantic drift, comparing the performance of different ablations of SCAR on hypotheses with a varying number of facts in the gold explanations. The results of the study are reported in Figure 2b. Overall, we observe a degradation in performance for all the considered models that becomes more prominent as the explanations increase in size. Such a degradation is likely due to semantic drift. However, the results suggest that SCAR exhibits more stable performance on long explanations (≥ 6 facts) when compared to its individual sub-components. In particular, the plotted results in

Figure 2b clearly show that, while all the models start with comparable MAP scores on explanations containing a single fact, the gap in performance gradually increases with the size of the explanations, with SCAR obtaining an improvement of 13.46 MAP over BM25 + Bi-encoder on explanations containing more than 10 facts. These results confirm the hypotheses that implicit and explicit patterns possess complementary features for explanation regeneration and that the proposed hybridisation has a decisive impact on improving multi-hop inference for scientific hypotheses in the most challenging setting.

Multi-Hop Question Answering

Since the construction of spurious inference chains can lead to wrong answer prediction, semantic drift often influences the downstream capabilities of answering the question. Therefore, we additionally evaluate the performance of SCAR on the multiple-choice question answering task (WorldTree dev-set), employing the model as an explainable solver without additional training. Specifically, given a multiple-choice science question, we employ SCAR to construct an explanation for each candidate answer, and derive the relative candidate answer score by summing up the explanatory score of each fact in the explanation (Equation 3). Subsequently, we consider the answer with the highest-scoring explanation as the correct one. Table 4 shows the results achieved adopting different iterations t for the inference. Similarly to the results on explanation regeneration, this experiment confirms the interplay between dense and sparse models in improving the performance and robustness on downstream question answering. Specifically, we observe that, while the performance of different ablations decreases rapidly with an increasing number of inference steps, the performance of SCAR are more stable, reaching a peak at $t = 3$. This confirms the robustness of SCAR in multi-hop inference together with its resilience to semantic drift.

Model	t = 1	t = 2	t = 3	t = 4
Random	25.00	25.00	25.00	25.00
BM25	**48.23**	39.82	35.84	33.18
Bi-encoder	**54.42**	52.21	50.88	50.00
Bi-encoder + BM25	**59.29**	52.21	47.79	44.69
SCAR	60.62	60.62	<u>**61.06**</u>	57.96

Table 4: Accuracy in question answering using the models as explainable inference solvers without additional training.
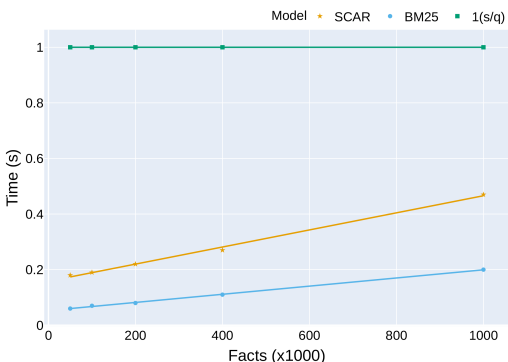

Figure 3: Scalability of SCAR to corpora containing a million facts compared to that of standalone BM25.

Scalability

Finally, we measure the scalability of SCAR on facts banks containing milions of sentences. To perform this analysis, we gradually expand the set of facts in the WorldTree corpus by randomly extracting sentences from GenericsKB[3] (Bhakthavatsalam, Anastasiades, and Clark 2020), a curated facts bank of commonsense and scientific knowledge. To evaluate scalability, we compare the inference time of SCAR with that of standalone BM25, which is widely adopted for Information Retrieval at scale (Robertson, Zaragoza et al. 2009). The results of this experiment, reported in Figure 3, demonstrate that SCAR scales similarly to BM25. Even considering the overhead caused by the Bi-encoder model, in fact, SCAR is still able to perform inference in less than 1 second per question on corpora containing 1 million facts, demonstrating its suitability for real-world scenarios requiring inference on large knowledge sources.

Related Work

Multi-hop inference is the task of combining multiple pieces of evidence to solve a particular reasoning problem. This task is often used to evaluate explainable inference since the constructed chains of reasoning can be interpreted as an explanation for the final predictions (Wiegreffe and Marasović 2021; Thayaparan, Valentino, and Freitas 2020). Given the importance of multi-hop reasoning for explainability, there is a recent focus on resources providing annotated expla-

---

[3] https://allenai.org/data/genericskb

nations to support the inference (Xie et al. 2020; Jhamtani and Clark 2020; Khot et al. 2020; Ferreira and Freitas 2020; Khashabi et al. 2018; Yang et al. 2018; Mihaylov et al. 2018; Jansen et al. 2018; Welbl, Stenetorp, and Riedel 2018). In particular, explanation regeneration on science questions is designed to evaluate the construction of long explanatory chains, in a setting where the structure of the inference cannot be derived from a direct decomposition of the questions (Xie et al. 2020; Jansen and Ustalov 2019; Jansen et al. 2018). To deal with the difficulty of the task, state-of-the-art models leverage the attention mechanism in Transformers (Vaswani et al. 2017), learning to compose relevant explanatory chains via sequence classification models (Cartuyvels, Spinks, and Moens 2020; Das et al. 2019; Chia, Witteveen, and Andrews 2019; Banerjee 2019). The autoregressive formulation proposed in this paper is similar to the one introduced by Cartuyvels, Spinks, and Moens (2020), which, however, perform iterative inference though a cross-encoder architecture based on BERT (Devlin et al. 2019). Differently from this work, we present a hybrid architecture based on bi-encoders (Reimers and Gurevych 2019) with the aim of optimising both accuracy and inference time in explanation regeneration. Our model of explanatory power is based on the work done by Valentino, Thayaparan, and Freitas (2021), which shows how to leverage explicit patterns in corpora of scientific explanations for multi-hop inference. In this paper, we build upon this line of research by demonstrating that models based on explicit patterns can be combined with neural architectures to achieve nearly state-of-the-art performance while preserving scalability. Our framework is related to recent work on dense retrieval for knowledge-intensive NLP tasks, which focuses on the design of scalable architectures with Maximum Inner Product Search (MIPS) based on Transformers (Xiong et al. 2021; Zhao et al. 2021; Lin et al. 2021; Karpukhin et al. 2020; Lewis et al. 2020; Dhingra et al. 2019). Our multi-hop dense encoder is similar to Lin et al. (2021) and Xiong et al. (2021) which adopt bi-encoders for multi-step retrieval on open-ended commonsense reasoning and open-domain question answering. However, to the best of our knowledge, we are the first to integrate dense bi-encoders in a hybrid architecture for complex explainable inference in the scientific domain.

Conclusion

This work presented SCAR, a hybrid autoregressive architecture for scalable explanation regeneration. An extensive evaluation demonstrated that SCAR achieves performance comparable with that of state-of-the-art cross-encoders while being ≈ 50 times faster and intrinsically scalable, and confirmed the impact of the hybridisation on semantic drift and question answering. This work demonstrated the effectiveness of hybrid architectures for explainable inference at scale, opening the way for future research at the intersection of latent and explicit models. As a future work, we plan to investigate the integration of relevance and explanatory power in an end-to-end differentiable architecture, and explore the applicability of the hybrid framework on additional natural language and scientific reasoning tasks, with a focus on real-world scientific inference problems.

References

Banerjee, P. 2019. ASU at TextGraphs 2019 Shared Task: Explanation ReGeneration using Language Models and Iterative Re-Ranking. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 78–84.

Bhakthavatsalam, S.; Anastasiades, C.; and Clark, P. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Cartuyvels, R.; Spinks, G.; and Moens, M.-F. 2020. Autoregressive Reasoning over Chains of Facts with Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6916–6930. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Chia, Y. K.; Witteveen, S.; and Andrews, M. 2019. Red Dragon AI at TextGraphs 2019 Shared Task: Language Model Assisted Explanation Generation. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 85–89.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Das, R.; Godbole, A.; Zaheer, M.; Dhuliawala, S.; and McCallum, A. 2019. Chains-of-Reasoning at TextGraphs 2019 Shared Task: Reasoning over Chains of Facts for Explainable Multi-hop Inference. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 101–117.

Demszky, D.; Guu, K.; and Liang, P. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dhingra, B.; Zaheer, M.; Balachandran, V.; Neubig, G.; Salakhutdinov, R.; and Cohen, W. W. 2019. Differentiable Reasoning over a Virtual Knowledge Base. In *International Conference on Learning Representations*.

Ferreira, D.; and Freitas, A. 2020. Natural Language Premise Selection: Finding Supporting Statements for Mathematical Text. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 2175–2182.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742.

Humeau, S.; Shuster, K.; Lachaux, M.-A.; and Weston, J. 2019. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *International Conference on Learning Representations*.

Jansen, P.; Balasubramanian, N.; Surdeanu, M.; and Clark, P. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2956–2965.

Jansen, P.; and Ustalov, D. 2019. TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 63–77.

Jansen, P.; Wainwright, E.; Marmorstein, S.; and Morrison, C. 2018. WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jhamtani, H.; and Clark, P. 2020. Learning to Explain: Datasets and Models for Identifying Valid Reasoning Chains in Multihop Question-Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 137–150.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 1–1.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

Khashabi, D.; Azer, E. S.; Khot, T.; Sabharwal, A.; and Roth, D. 2019. On the capabilities and limitations of reasoning for natural language understanding. *arXiv preprint arXiv:1901.02522*.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2019. QASC: A Dataset for Question Answering via Sentence Composition. *arXiv preprint arXiv:1910.11473*.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. QASC: A Dataset for Question Answering via Sentence Composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8082–8090.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Lin, B. Y.; Sun, H.; Dhingra, B.; Zaheer, M.; Ren, X.; and Cohen, W. 2021. Differentiable Open-Ended Commonsense

Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4611–4625. Online: Association for Computational Linguistics.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Thayaparan, M.; Valentino, M.; and Freitas, A. 2020. A Survey on Explainability in Machine Reading Comprehension. *arXiv preprint arXiv:2010.00389*.

Thayaparan, M.; Valentino, M.; and Freitas, A. 2021. Explainable Inference Over Grounding-Abstract Chains for Science Questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1–12. Online: Association for Computational Linguistics.

Valentino, M.; Thayaparan, M.; and Freitas, A. 2021. Unification-based Reconstruction of Multi-hop Explanations for Science Questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 200–211. Online: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6: 287–302.

Wiegreffe, S.; and Marasović, A. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Xie, Z.; Thiem, S.; Martin, J.; Wainwright, E.; Marmorstein, S.; and Jansen, P. 2020. WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5456–5473. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Xiong, W.; Li, X.; Iyer, S.; Du, J.; Lewis, P.; Wang, W. Y.; Mehdad, Y.; Yih, S.; Riedel, S.; Kiela, D.; and Oguz, B. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *International Conference on Learning Representations*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Zhao, C.; Xiong, C.; Boyd-Graber, J.; and Daumé III, H. 2021. Multi-Step Reasoning Over Unstructured Text with Beam Dense Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4635–4641. Online: Association for Computational Linguistics.