Closed domain knowledge injection into general Text-to-Text model

Stefano di Terlizzi

Contributing authors: stefanoditerlizzi96@gmail.com;

Abstract

Dialogue generation has been successfully learned from scratch by neural networks, but tends to produce the same general response, e.g., "what are you talking about?", in many conversations. To reduce this homogeneity, external knowledge is applied as an additional condition to diversify a model's output. The project aims at injecting specific domain knowledge and evaluating the IR system and the model response with and without the injection.

Keywords: Text-to-text generative models, Hallucination, Closed domain injection, Text generation evaluation

1 Introduction

One limitation of Large Language Model (LLM) content generation is hallucination, or false assertions in the generated text. Traditionally, fine-tuning has been a goto solution to mitigate hallucination in generative models. However, this approach often demands additional computational resources and extensive training, limiting its practicality. Alternatively, an emerging strategy involves injecting domain-specific knowledge directly into the model during the inference process, thus enhancing the model's understanding of the given context and reducing the likelihood of hallucination. In this study, we delve into the efficacy of closed domain injection as a solution to address hallucination in text-to-text generative models. Specifically, we explore the feasibility of augmenting input prompts with domain-specific information obtained from external knowledge retrieval systems. By integrating such knowledge directly into the model's input, we aim to enhance its contextual understanding and improve the relevance and coherence of generated text. To evaluate the effectiveness of this approach, we conduct experiments using two distinct text-to-text generative models, differing in the number of parameters: one with 77 million parameters (flan-t5-small

by Google) and another with 783 million parameters (flan-t5-large by Google). By comparing the performance of these models with and without closed domain injection, we seek to elucidate the impact of knowledge enhancement on mitigating hallucination across varying model complexities. Through our investigation, we aim to provide insights into the potential of closed domain injection as a practical and efficient strategy for improving the robustness and accuracy of text-to-text generative models in real-world applications.

2 Problem Setup and Experiments

Before delving into the details of our experiments and the methods employed, it is imperative to establish the problem context and the experimental framework. In this section, we outline the dataset used for training and evaluation, introduce the information retrieval system utilized for enhancing prompts, detail the methods employed for prompt enhancement, and elucidate the evaluation metrics utilized to assess the performance of our approach. Now, let's proceed to delve deeper into each component of our problem setup and experimental design.

2.1 Dataset

The dataset utilized in this study comprises structured information sourced from various domains. Each entry in the dataset consists of contextual information, questions, and corresponding target answers. A sample entry from the dataset is presented below:

```
2
3
              "idx": "/wiki/Knox_Cunningham#P39#0",
              "question": "Which position did Knox Cunningham hold from May 1955 to Apr 1956?",
4
 5
              "context": "Knox Cunningham Sir Samuel Knox Cunningham , 1st Baronet , ...",
               "targets": [
 6
                   "Ulster Unionist MP for South Antrim"
 7
 8
               "level": "easy"
9
10
          },
11
12
     ]
```

It is important to note that our dataset contains duplicates in the contexts, meaning that multiple questions may refer to the same context, resulting in duplicates.

2.2 Information retrieval system for enhancing prompts

To enhance prompts, an information retrieval system is employed, leveraging the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The system utilizes a preprocessed corpus of contexts and questions for vectorization. We choose to organize the contexts as a set of unique contexts due to the presence of duplicate contexts associated with different answers. By organizing the contexts as a set, we ensure that each unique context is considered only once during training and evaluation. This approach prevents redundancy in the dataset and ensures that the models learn from

diverse contextual information without being biased towards repetitive contexts. Subsequently, a TF-IDF vectorizer is fitted on the combined corpus. An excerpt of the code implementation is shown below:

```
from sklearn.feature_extraction.text import TfidfVectorizer

contexts = list(set([e["context"] for e in data]))

dataForFit = contexts + [e["question"] for e in data]

vectorizer = TfidfVectorizer(tokenizer=nltk.word_tokenize)

vectorizer.fit(dataForFit)

corpusTFIDF = vectorizer.transform(contexts)
```

2.3 Enhancing methods

The enhancement process involves injecting domain-specific knowledge into the input prompt using the TF-IDF-based information retrieval system. The system identifies relevant context passages based on cosine similarity scores between questions and contexts. Quantile filtering is applied to refine the selection of relevant sentences from the identified context. The enriched prompt is then utilized for text generation. The core functionalities of the enhancement method are illustrated in the provided code snippet.

```
for i,e in enumerate(data):
              #find the context with the IR system
 2
 3
              answer = vectorizer.transform([e["question"]])
              arr = cosine_similarity(answer, corpusTFIDF)[0]
 4
 5
              mx = np.amax(arr)
              contextIndex = (np.where(arr == mx))[0][0]
 6
              #split the context into sentences and score them by cosine similarity
              sentences = sent tokenize(contexts[contextIndex])
 9
10
              sent = vectorizer.transform(sentences)
              sentSimilarity = [(i, score) for i, score in enumerate(cosine_similarity(answer, sent)[0])]
11
12
13
              #quantile filtering
              quantiles = [0, 0.5, 0.75, 0.90]
14
              quant = [np.quantile([e[1] for e in sentSimilarity], q) for q in quantiles]
              for i,q in enumerate(quant):
16
17
                  #quantile filtering and reduce to a string the result
                  searchOn = reduce(
18
                      lambda x, y: f'{x} {y}',
19
                      [ el for i,el in enumerate(sentences) if i in [ e[0] for e in sentSimilarity if e[1]>=q ] ]
20
21
22
                  #generate the response enanching the input
23
                  #with the result of previus computation + the question
                  response = text2textModel(
24
                      tokenizer=mTokenizer,
25
                      model=mModel.
26
27
                      input_text = f'{searchOn}. {e["question"]}'
28
                  response = response.replace('<pad> ', '').replace('</s>', '')
29
30
                  similarityResponseTarget = cosine_similarity(
31
```

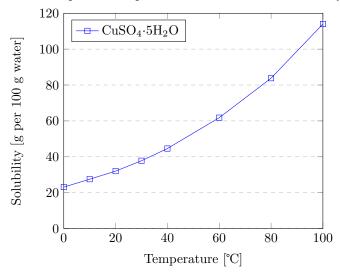
```
vectorizer.transform([response]),
32
                      vectorizer.transform(e["targets"])
33
                  )[0]
34
35
                  finalResult.append({
36
37
                      'question': e["question"],
                      'correctContext': contexts[contextIndex] == e["context"],
38
                      'scoreContextRetrieval': mx,
39
                      'quantileLimitSentences': quantiles[i],
40
                      'targetWithSimilarityScore': [ (el, similarityResponseTarget[i]) for i,el in enumerate(e['targets']) ],
41
42
                      'response': response
43
```

2.4 Evaluation methods

The evaluation methodology relies on cosine similarity metrics to assess the quality and relevance of generated responses compared to target answers. Each response generated by the model is evaluated based on its cosine similarity score with the corresponding target answer. The evaluation process aims to quantify the degree of semantic similarity between generated responses and ground truth answers.

3 Results

Temperature dependence of CuSO₄·5H₂O solubility



4 This is an example for first level head—section head

4.1 This is an example for second level head—subsection head

4.1.1 This is an example for third level head—subsubsection head

Sample body text. Sample body text.

5 Equations

Equations in LATEX can either be inline or on-a-line by itself ("display equations"). For inline equations use the \$...\$ commands. E.g.: The equation $H\psi = E\psi$ is written via the command \$\mathbb{H} \psi = E \psi\\$.

For display equations (with auto generated equation numbers) one can use the equation or align environments:

$$\|\tilde{X}(k)\|^{2} \leq \frac{\sum_{i=1}^{p} \|\tilde{Y}_{i}(k)\|^{2} + \sum_{j=1}^{q} \|\tilde{Z}_{j}(k)\|^{2}}{p+q}.$$
 (1)

where,

$$D_{\mu} = \partial_{\mu} - ig \frac{\lambda^{a}}{2} A^{a}_{\mu}$$

$$F^{a}_{\mu\nu} = \partial_{\mu} A^{a}_{\nu} - \partial_{\nu} A^{a}_{\mu} + g f^{abc} A^{b}_{\mu} A^{a}_{\nu}$$

$$(2)$$

Notice the use of \nonumber in the align environment at the end of each line, except the last, so as not to produce equation numbers on lines where no equation numbers are required. The \label{} command should only be used at the last line of an align environment where \nonumber is not used.

$$Y_{\infty} = \left(\frac{m}{\text{GeV}}\right)^{-3} \left[1 + \frac{3\ln(m/\text{GeV})}{15} + \frac{\ln(c_2/5)}{15}\right]$$
 (3)

The class file also supports the use of \mathbf{R} , \mathbf{R} and \mathbf{R} produces \mathbf{R} , and \mathbf{R} respectively (refer Subsubsection 4.1.1).

6 Tables

Tables can be inserted via the normal table and tabular environment. To put footnotes inside tables you should use \footnotetext[]{...} tag. The footnote appears just below the table itself (refer Tables 1 and 2). For the corresponding footnotemark use \footnotemark[...]

Table 1 Caption text

Column 1	Column 2	Column 3	Column 4
row 1	data 1	$\begin{array}{c} \text{data 2} \\ \text{data 5}^1 \\ \text{data 8} \end{array}$	data 3
row 2	data 4		data 6
row 3	data 7		data 9 ²

Source: This is an example of table footnote. This is an example of table footnote.

The input format for the above table is as follows:

```
\begin{table}[<placement-specifier>]
\caption{<table-caption>}\label{<table-label>}%
\begin{tabular}{0{}11110{}}
\toprule
Column 1 & Column 2 & Column 3 & Column 4\\
\midrule
row 1 & data 1 & data 2 & data 3 \\
row 2 & data 4 & data 5\footnotemark[1] & data 6 \\
row 3 & data 7 & data 8 & data 9\footnotemark[2]\\
\botrule
\end{tabular}
\footnotetext{Source: This is an example of table footnote.
This is an example of table footnote.}
\footnotetext[1]{Example for a first table footnote.
This is an example of table footnote.}
\footnotetext[2]{Example for a second table footnote.
This is an example of table footnote.}
\end{table}
```

Table 2 Example of a lengthy table which is set to full textwidth

	Element 1 ¹			Element 2 ²		
Project	Energy	σ_{calc}	σ_{expt}	Energy	σ_{calc}	σ_{expt}
Element 3 Element 4	990 A 500 A	1168 961	1547 ± 12 922 ± 10	780 A 900 A	1166 1268	1239 ± 100 1092 ± 40

Note: This is an example of table footnote. This is an example of table footnote this is an example of table footnote this is an example of table footnote.

 $^{^1{\}rm Example}$ for a first table footnote. This is an example of table footnote.

 $^{^2{\}rm Example}$ for a second table footnote. This is an example of table footnote.

¹Example for a first table footnote.

 $^{^2}$ Example for a second table footnote.

In case of double column layout, tables which do not fit in single column width should be set to full text width. For this, you need to use \begin{table*} ... \end{table*} instead of \begin{table} ... \end{table} environment. Lengthy tables which do not fit in textwidth should be set as rotated table. For this, you need to use \begin{sidewaystable} ... \end{sidewaystable} instead of \begin{table*} ... \end{table*} environment. This environment puts tables rotated to single column width. For tables rotated to double column width, use \begin{sidewaystable*} ... \end{sidewaystable*}... \end{sidewaystable*}.

7 Figures

As per the LATEX standards you need to use eps images for LATEX compilation and pdf/jpg/png images for PDFLaTeX compilation. This is one of the major difference between LATEX and PDFLaTeX. Each image should be from a single input .eps/vector image file. Avoid using subfigures. The command for inserting images for LATEX and PDFLaTeX can be generalized. The package used to insert images in LaTeX/PDFLaTeX is the graphicx package. Figures can be inserted via the normal figure environment as shown in the below example:

```
\begin{figure}[<placement-specifier>]
\centering
\includegraphics{<eps-file>}
\caption{<figure-caption>}\label{<figure-label>}
\end{figure}
```

```
fig-eps-converted-to.pdf
```

Fig. 1 This is a widefig. This is an example of long caption this is an example of long caption this is an example of long caption

In case of double column layout, the above format puts figure captions/images to single column width. To get spanned images, we need to provide \begin{figure*}...\end{figure*}.

For sample purpose, we have included the width of images in the optional argument of \includegraphics tag. Please ignore this.

8 Algorithms, Program codes and Listings

Packages algorithm, algorithmicx and algorithms in LATEX using the format:

Table 3 Tables which are too long to fit, should be written using the "sidewaystable" environment as shown here

1168 1547 ± 12 780 A 1166 961 922 ± 10 900 A 1268

Note: This is an example of table footnote this is an example of table footnote this is an example of table footnote this is an example of table footnote.

 $^{^1\}mathrm{This}$ is an example of table footnote.

```
\begin{algorithm}
\caption{<alg-caption>}\label{<alg-label>}
\begin{algorithmic}[1]
. . .
\end{algorithmic}
\end{algorithm}
```

You may refer above listed package documentations for more details before setting algorithm environment. For program codes, the "verbatim" package is required and the command to be used is \begin{verbatim} . . . \end{verbatim}.

Similarly, for listings, use the listings package. \begin{lstlisting} ... \end{lstlisting} is used to set environments similar to verbatim environment. Refer to the lstlisting package documentation for more details.

A fast exponentiation procedure:

```
begin
   for i := 1 to 10 step 1 do
        expt(2,i);
                                              Comments will be set flush to the right margin
        newline() od
where
\operatorname{proc} \operatorname{expt}(x,n) \equiv
  z := 1;
  do if n=0 then exit fi;
      do if odd(n) then exit fi;
          comment: This is a comment statement;
          n := n/2; \quad x := x * x \text{ od};
       \{ n > 0 \};
      n:=n-1; z:=z*x od;
   print(z).
end
```

```
for i:=maxint to 0 do
begin
{ do nothing }
end;
Write('Case-insensitive-');
Write('Pascal-keywords.');
```

9 Cross referencing

Environments such as figure, table, equation and align can have a label declared via the \label{#label} command. For figures and table environments use the \label{} command inside or just below the \caption{} command. You can then use the \ref{#label} command to cross-reference them. As an example, consider the label

Algorithm 1 Calculate $y = x^n$

```
Require: n \ge 0 \lor x \ne 0
Ensure: y = x^n
 1: y \Leftarrow 1
 2: if n < 0 then
          X \Leftarrow 1/x
          N \Leftarrow -n
 5: else
          X \Leftarrow x
 6:
          N \Leftarrow n
 7:
 8: end if
     while N \neq 0 do
 9:
          if N is even then
10:
              X \Leftarrow X \times X
11:
              N \Leftarrow N/2
12:
          else[N \text{ is odd}]
13:
              y \Leftarrow y \times X
14:
              N \Leftarrow N - 1
15:
          end if
16:
17: end while
```

declared for Figure 1 which is \label{fig1}. To cross-reference it, use the command Figure \ref{fig1}, for which it comes up as "Figure 1".

To reference line numbers in an algorithm, consider the label declared for the line number 2 of Algorithm 1 is \label{algln2}. To cross-reference it, use the command \ref{algln2} for which it comes up as line 2 of Algorithm 1.

9.1 Details on reference citations

Standard LATEX permits only numerical citations. To support both numerical and author-year citations this template uses natbib LATEX package. For style guidance please refer to the template user manual.

Here is an example for \cite{...}: [1]. Another example for \citep{...}: [2]. For author-year citation mode, \cite{...} prints Jones et al. (1990) and \citep{...} prints (Jones et al., 1990).

All cited bib entries are printed at the end of this article: [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] and [13].

10 Examples for theorem like environments

For theorem like environments, we require amsthm package. There are three types of predefined theorem styles exists—thmstyleone, thmstyletwo and thmstylethree

thmstyleone	Numbered, theorem head in bold font and theorem	
	text in italic style	
thmstyletwo	Numbered, theorem head in roman font and theorem	
	text in italic style	
thmstylethree	Numbered, theorem head in bold font and theorem	
	text in roman style	

For mathematics journals, theorem styles can be included as shown in the following examples:

Theorem 1 (Theorem subhead). Example theorem text. Example theorem text.

Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text.

Proposition 2. Example proposition text. Example proposition text.

Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text.

Example 1. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem.

Sample body text. Sample body text.

Remark 1. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem.

Sample body text. Sample body text.

Definition 1 (Definition sub head). Example definition text. Example definition text.

Additionally a predefined "proof" environment is available: \begin{proof} ... \end{proof}. This prints a "Proof" head in italic font style and the "body text" in roman font style with an open square at the end of each proof environment.

Proof. Example for proof text. \Box

Sample body text. Sample body text.

Proof of Theorem 1. Example for proof text. \Box

For a quote environment, use \begin{quote}...\end{quote}

Quoted text example. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Sample body text. Sample body text. Sample body text. Sample body text. Sample body text (refer Figure 1). Sample body text. Sample body text. Sample body text (refer Table 3).

11 Methods

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work. Authors are encouraged to include RIIDs where appropriate.

Ethical approval declarations (only required where applicable) Any article reporting experiment/s carried out on (i) live vertebrate (or higher invertebrates), (ii) humans or (iii) human samples must include an unambiguous statement within the methods section that meets the following requirements:

- 1. Approval: a statement which confirms that all experimental protocols were approved by a named institutional and/or licensing committee. Please identify the approving body in the methods section
- 2. Accordance: a statement explicitly saying that the methods were carried out in accordance with the relevant guidelines and regulations
- 3. Informed consent (for experiments involving humans or human tissue samples): include a statement confirming that informed consent was obtained from all participants and/or their legal guardian/s

If your manuscript includes potentially identifying patient/participant information, or if it describes human transplantation research, or if it reports results of a clinical trial then additional information will be required. Please visit (https://www.nature.com/nature-research/editorial-policies) for Nature Portfolio journals, (https://www.springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/publishing-ethics/14214) for Springer Nature journals, or (https://www.biomedcentral.com/getpublished/editorial-policies#ethics+and+consent) for BMC.

12 Discussion

Discussions should be brief and focused. In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Some journals prefer a

section 'Results and Discussion' followed by a section 'Conclusion'. Please refer to Journal-level guidance for any specific requirements.

13 Conclusion

Conclusions may be used to restate your hypothesis or research question, restate your major findings, explain the relevance and the added value of your work, highlight any limitations of your study, describe future directions for research and recommendations.

In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Please refer to Journal-level guidance for any specific requirements.

Supplementary information. If your article has accompanying supplementary file/s please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
- Ethics approval and consent to participate
- Consent for publication
- Data availability
- Materials availability
- Code availability
- Author contribution

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

- [1] Campbell, S.L., Gear, C.W.: The index of general nonlinear DAES. Numer. Math. **72**(2), 173–196 (1995)
- [2] Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. 78, 74–80 (2000) https://doi.org/10.1007/s001090000086
- [3] Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**(2), 321–354 (1995)
- [4] Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)
- [5] Broy, M.: Software engineering—from auxiliary to key technologies. In: Broy, M., Denert, E. (eds.) Software Pioneers, pp. 10–13. Springer, New York (1992)
- [6] Seymour, R.S. (ed.): Conductive Polymers. Plenum, New York (1981)
- [7] Smith, S.E.: Neuromuscular blocking drugs in man. In: Zaimis, E. (ed.) Neuromuscular Junction. Handbook of Experimental Pharmacology, vol. 42, pp. 593–660. Springer, Heidelberg (1976)
- [8] Chung, S.T., Morris, R.L.: Isolation and characterization of plasmid deoxyribonucleic acid from Streptomyces fradiae. Paper presented at the 3rd international symposium on the genetics of industrial microorganisms, University of Wisconsin, Madison, 4–9 June 1978 (1978)
- [9] Hao, Z., AghaKouchak, A., Nakhjiri, N., Farahmand, A.: Global integrated drought monitoring and prediction system (GIDMaPS) data sets. figshare https: //doi.org/10.6084/m9.figshare.853801 (2014)
- [10] Babichev, S.A., Ries, J., Lvovsky, A.I.: Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon. Preprint at https://arxiv.org/abs/quant-ph/0208066v1 (2002)

- [11] Beneke, M., Buchalla, G., Dunietz, I.: Mixing induced CP asymmetries in inclusive B decays. Phys. Lett. **B393**, 132–142 (1997) arXiv:0707.3168 [gr-gc]
- [12] Stahl, B.: DeepSIP: Deep Learning of Supernova Ia Parameters, 0.42, Astrophysics Source Code Library (2020), ascl:2006.023
- [13] Abbott, T.M.C., et al.: Dark Energy Survey Year 1 Results: Constraints on Extended Cosmological Models from Galaxy Clustering and Weak Lensing. Phys. Rev. D 99(12), 123505 (2019) https://doi.org/10.1103/PhysRevD.99.123505 arXiv:1810.02499 [astro-ph.CO]