# FutureGen: LLM-RAG Approach to Generate the Future Work of Scientific Article

Ibrahim Al Azher†, Miftahul Jannat Mokarrama†, Zhishuai Guo†,
Sagnik Ray Choudhury‡, Hamed Alhoori†
†Northern Illinois University, Dekalb, IL, USA
‡University of North Texas, Denton, TX, USA
{iazher, mmokarrama1, zguo, alhoori}@niu.edu,
sagnik.raychoudhury@unt.edu

## Abstract

The future work section of a scientific article outlines potential research directions by identifying gaps and limitations of a current study. This section serves as a valuable resource for early-career researchers seeking unexplored areas and experienced researchers looking for new projects or collaborations. In this study, we *generate* future work suggestions from key sections of a scientific article alongside related papers and analyze how the trends have evolved. We experimented with various Large Language Models (LLMs) and integrated Retrieval-Augmented Generation (RAG) to enhance the generation process. We incorporate a LLM feedback mechanism to improve the quality of the generated content and propose an LLM-as-a-judge approach for evaluation. Our results demonstrated that the RAG-based approach with LLM feedback outperforms other methods evaluated through qualitative and quantitative metrics. Moreover, we conduct a human evaluation to assess the LLM as an extractor and judge. The code and dataset for this project are here: code : HuggingFace

## 1 Introduction

The future work section in a scientific article plays a crucial role in amplifying a study's impact by demonstrating foresight and highlighting the broader implications of research (Nicholas et al., 2015; Aguinis et al., 2018). It serves as a catalyst for further exploration, interdisciplinary collaboration, and new ideas, transforming a single study into a foundation for future advancements (Al Azhar et al., 2021; Hara et al., 2003). Beyond academia, future work insights benefit policymakers and funding agencies by identifying emerging research directions and prioritizing areas for strategic resource allocation (Hyder et al., 2011; Thelwall et al., 2023; Simsek et al., 2024). A well-constructed future work section serves both methodological and practical purposes: it encourages researchers to critically reflect on their study's limitations, fostering higher-quality subsequent research while also streamlining the peer review process by clarifying the authors' awareness of challenges and next steps (Kelly et al., 2014). Acknowledging a study's limitations provides clear signals for further exploration, bridging current findings with future advancements (Conaway et al., 2015). Additionally, analyzing the evolution of future work trends helps researchers align with current priorities, uncover unexplored gaps, and avoid redundancy. Furthermore, understanding long-term research trajectories allows early-career researchers to identify high-impact topics and strategically position their contributions in advancing scientific progress (Ortagus et al., 2020).

Author-written future work sections have many problems. Firstly, they are often unspecific, ambiguous, not always easy to find, and speculative in nature (Suray et al., 2024). Secondly, authors may struggle to articulate meaningful future directions, particularly when faced with space constraints or time limitations. Thirdly, they may hesitate to share research plans without a clear reward. This may partially explain why many future work proposals are overlooked post-publication (Teufel, 2017). To alleviate this problem, we generate future work from each paper and evaluate it against both the author's mentioned future work and long-term goals extracted from OpenReview peer reviews in our approach. By merging these peer-reviewed objectives with the author-mentioned future work, we build a more comprehensive and robust ground truth.

Advances in artificial intelligence (AI) offer transformative potential for addressing this gap. Unlike traditional methods, AI can systematically synthesize research trajectories, uncover latent connections, and propose novel directions that align with emerging trends (Wang et al., 2023). For example, Si et al. (2024) shows that LLM-generated ideas are more novel than human expert-generated ones. However, current applications of

AI like ChatGPT risk homogenizing outputs and reducing individual creativity (Ashkinaze et al., 2024; Anderson et al., 2024). Standard LLMs may generate overgeneralized, irrelevant, or fabricated future work directions. To address these challenges, this work utilizes an LLM to suggest future work, enhances its output using LLM-based feedback, and incorporates RAG using cross-domain insights.

Evaluation of AI-generated future work sentences is challenging. Traditional Natural Language Processing (NLP) evaluation metrics that rely on n-gram text overlaps or semantic similarity, such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019) to compare the generated text with references, fail to fully capture the nuances of this generation process. To address this issue, we incorporate LLM-based evaluation providing human-like assessments with explanations. If the generated future work does not meet quality thresholds, we refine it using an LLM-driven feedback loop, improving its alignment with the context. We mitigate issues like vagueness and redundancy observed in prior AI-driven ideation tools by integrating iterative LLM feedback. Beyond generation, our approach embeds temporal trend analysis to track evolving shifts over time. Our Contributions can be summarized as follows:

- **Dataset.** We created a dataset of future work sentences from nearly 8,000 papers collected from the ACL conference from 2012 to 2024 and 1000 papers from NeurIPS. Papers often do not have exclusive Future Work sections (they are combined with general conclusions or limitations), therefore, we use LLMs to extract relevant sentences. We also validated this extraction process on a random sample of the dataset with human annotators, ensuring that the LLM accurately identifies future-work content.

- **Trend Analysis.** We analyze how future work has evolved in the NLP domain over the past 13 years by generating topics and titles using LLM-based Topic modeling. Traditional approaches that rely only on topic words or clustering do not provide a nuanced, interpretable picture. By integrating LLM-based topic modeling with BERTopic (Grootendorst, 2022), our method generates not only topics but also descriptive titles, offering a more meaningful and comprehensive analysis of the trends in future work.

- **Future Work Generation.** We use LLMs to generate future work sentences for NLP papers. We overcome the context window length problem by selecting key sections. Our RAG system incorporates additional relevant information from other papers, enhancing the quality and depth of the generated future work

- **LLM as a Judge and Feedback Mechanism.** We incorporate an LLM as a judge to evaluate the quality of the generated text rather than only relying on NLP evaluation metrics and use the judge feedback to improve the generation process.

Table 1: Future-Work Suggestions from Different Sources

| **Source** | **Future-Work** |
|---|---|
| Ground truth | We will explore the application of SWAD to other robustness problems, including ImageNet classification and its robustness benchmarks. |
| LLM | Future research could explore the applicability of flatness-aware methods in other machine learning contexts, such as reinforcement learning or unsupervised learning. |
| RAG | Investigate the effectiveness of the SWAD method across a wider variety of datasets and domains, particularly those with significant distribution shifts. This would validate the generalization capabilities of SWAD beyond the current benchmarks. |

## 2 Related Work

Recent advancements in NLP and large language models (LLMs) have enabled the automatic extraction and generation of various sections of scientific articles, such as Abstracts (Gonçalves et al., 2018), Methodologies (Houngbo and Mercer, 2012), and Limitations (Al Azher et al., 2024; Azher, 2024; Al Azher and Alhoori, 2024). In the domain of future work, prior studies have focused on tasks such as extraction (Hu and Wan, 2015), classification (Zhang et al., 2023), identifying creative topics (Song et al., 2021), thematic categorization (Hao et al., 2020), and trend prediction (Qian et al., 2021). For example, a BERT-based model has been used to annotate “future research” sentences, enabling the extraction, clustering (Zhu et al., 2019), and impact analysis of these sentences (Suray et al., 2024). Other work has integrated RAG for idea generation (Radensky et al., 2024) and further trend prediction (Qian et al., 2021). However, most efforts have concentrated on identifying, collecting, or forecasting trends rather than on synthesizing suggestive future work that provides actionable insights for researchers. Our work addresses this gap by leveraging LLMs to generate suggestive future work from key sections of research papers. By incorporating RAG-augmented cross-paper insights, our approach enhances the relevance and coherence of the generated suggestions.

The application of LLMs in scientific discovery has gained significant attention, particularly in generating novel research ideas and hypotheses. For example, new scientific discoveries in the biomedical field (Qi et al., 2023), LLM-based agents are used to automatically generate and test social scientific hypotheses (Manning et al., 2024), and a probabilistic model is used for hypothesis generation and scientific exploration using reward functions (Jain et al., 2023). Moreover, LLMs are capable of generating novel research ideas (Lu et al., 2024). Statistical tests show that LLM-generated ideas exhibit greater novelty compared to human-generated ones after extracting research topics from recent conferences and prompting both LLMs and humans to generate ideas (Si et al., 2024). While prior studies have explored LLM-driven hypothesis generation in other fields, such as biomedical, automated suggestive scientific exploration and idea generation in the NLP domain, remain undiscovered.

A critical aspect of improving LLM-generated content is the use of human feedback mechanisms (Ouyang et al., 2022). In its absence, self-refinement techniques have been proposed, such as a self-debugging framework where LLMs identify and correct their mistakes without human intervention (Chen et al., 2023). Moreover, a model that generates instructions, inputs, outputs, and filters invalid results with fine-tuning itself iteratively (Wang et al., 2022). AI feedback has been used for in-context learning with prompt criticism (Fu et al., 2023), in a GPT-4-based three-feedback-mechanism system to generate novel hypotheses (Yang et al., 2023), and in LLM-generated responses with utility-based feedback for prompt revision (Peng et al., 2023).

Our work builds on these advancements by using LLMs not only to generate future work sections using LLM but also to act as a judge for scoring and justification of the output. Inspired by self-refinement techniques (Madaan et al., 2023), we iteratively incorporate LLM feedback to refine the input, eliminating the need for additional training or reinforcement learning and thereby improving the quality of the generated content.

Evaluating LLM-generated text is vital for scientific writing. Studies comparing human and LLM performance (Chiang and Lee, 2023; Nguyen et al., 2024) show that LLM-based evaluations (covering grammaticality, cohesiveness, accuracy, coverage, etc.) align well with human judgments. Building on these insights, we use LLM-based evaluation to assess the quality of generated future work. In summary, while some progress has been made in automating the generation of scientific article sections and leveraging LLMs for scientific discovery, the suggestive generation of future work sections remains an open challenge. Our work addresses this gap by proposing a novel approach that combines LLM generation with RAG, self-refinement, and robust evaluation metrics.

## 3 Dataset Collection and Extraction

We created our dataset by extracting Future Work sentences from approximately $8,000$ papers published in the ACL conference from 2012 to 2024. We also collected $1,000$ papers from NeurIPS, along with their open-access peer reviews from OpenReview [1] from 2021-22. The extraction process from the paper is as follows: (1) If a paper has a section explicitly labeled “Limitations and Future Work”, we extract the entire section using the Science Parse[2] tool (2) If “Future Work” is not in the section title, we extracted sentences with at least one of the phrases “future” or “Future Work" and collect all of the sentences up to the next section. We employed python regex for this purpose. Such regex-based string matching has a high recall, so we further filtered these sentences using LLMs to improve precision. Since most papers do not have a dedicated section, it is often scattered in any other section. Therefore, filtering out noisy sentences is crucial to ensure accurate extraction. Here we employed an **LLM as an extractor** role to isolate only Future Work sentences while removing irrelevant sentences. This produces Future Work paragraphs from 6,227 papers from ACL and 1000 papers fron NeurIPS, averaging five sentences per paper with an average word length of 65.

**OpenReview:** Since no ACL OpenReview were available during our data collection process, we collected OpenReview from the NeurIPS papers only. After parsing text from OpenReview, we

[1] https://openreview.net

[2] https://github.com/allenai/science-parse

gathered all peer feedback and used an LLM to extract potential future-work suggestions for the authors. We then applied a second LLM to validate each extracted sentence, discarding any that did not represent true long-term research goals.

### 3.1 Human Evaluation

Additionally, we conducted a human evaluation to assess whether LLM is good at extracting Future Work sentences or not. We conducted a user study with 200 randomly selected samples involving three annotators who rated below two questions from 1 to 3 (1 = worst, 3 = best) with the question "How well does LLM extract the Future Work (ground truth) without generating/hallucinating? (**Q1**)". The annotators are graduate students, knowledgeable about machine learning, and not related with this paper. We observed a mean user rating of 2.54 out of 3 (Table 11), indicating strong agreement; this procedure thus produced our silver-standard dataset.

## 4 Methodology

Our workflow consists of two main stages: 1. *Generating Topics and Titles of Future Work*: We applied Topic Modeling with LLM to generate topics and titles for trend analysis of future work (Al Azher et al., 2024).
2. *Generating Future Work Text*: We used an LLM-based RAG approach to generate future work from the key sections of research papers with LLM-based evaluation and feedback.

### 4.1 Generating Topics and Titles of future work

Our first work focuses on generating topics and titles from author-stated future work for trend prediction from ACL data. We applied Topic Modeling with LLM to generate topics and titles for trend analysis of Future Work (Figure 1). Here, the input consists of author-stated future work from various papers published in a single year, and the output is a set of generated titles.

Our dataset spans 13 years, containing approximately $n$ research papers in each year. We extracted the author's mentioned Future Work from $n$ research papers, denoted as $FW = fw_1, fw_2, ..., fw_n$, where $fw_i$ represents the 'Future Work' paragraph containing sentences from paper $p_i$. These input texts are sent into BERTopic, a topic modeling tool that utilizes a transformer-based embedding from BERT. BERTopic clustered them into $x$ topics (e.g., $t_1, t_2...t_x$). Each topic $t_i$ comprises a set of topic words $tw_1, tw_2...tw_z$ and a corresponding collection of relevant texts $R_i$. To enhance interpretability, we integrate an LLM with BERTopic to generate a descriptive title for each topic (Figure 1). We tuned the components of BERTopic with UMAP, Count Vectorizer, Sentence Transformer, HDBSCAN, and Zero-shot min similarity for guided topic modeling. This entire process is repeated for all articles in each year in our dataset and the number of topic titles collected $x$ for each year is presented in (Table 15).

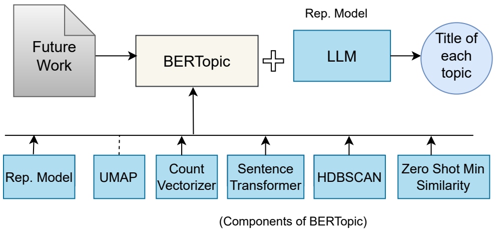

Figure 1: BERTopic + LLM pipeline for generating titles from Future Work.

### 4.2 Generating Future Work

The second work focuses on an LLM-based RAG pipeline that generates future-work sections directly from ACL and NeurIPS papers and refines them through LLM-driven evaluation and feedback. (Figure 2). In the ACL dataset, the ground truth consists of author-stated future work statements extracted using an LLM. For the NeurIPS dataset, the ground truth includes both the authors' future work statements and long-term suggestions from OpenReview peer reviewers. In both cases, we constructed input data by collecting all texts from the full paper after removing 'author-mentioned future work'. The output is the future work text generated by the model.

Figure 2 depicts our end-to-end pipeline: starting from a sample paper, we first extract candidate "future work" sentences via regex and refine them with an LLM, then pull peer-review comments from OpenReview and use the same LLM to isolate long-term goals—merging both into a robust ground truth. We remove the author's original future work, feed the paper to an LLM augmented by a vector-store retriever that supplies related documents, and generate new future work suggestions. Each paper is automatically scored on NLP met-

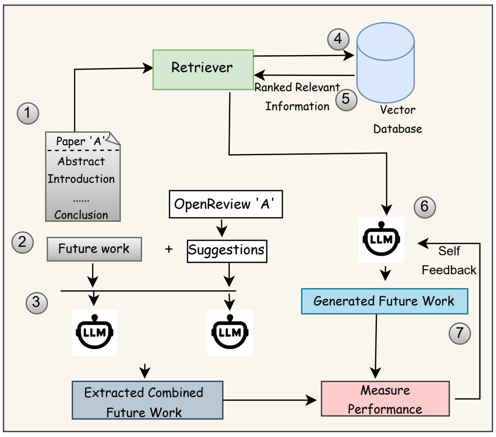

Figure 2: Overview of our LLM + RAG future-work pipeline: we extract author-written future work from the paper, extract long-term goals from peer reviews via a tool + LLM, generate new suggestions with a RAG-augmented LLM from the paper, evaluate their quality, and apply self-feedback for refinement.

rics and LLM-based criteria (coherence, relevance, novelty, grammar, overall impression), and any suggestion scoring below a threshold of 3 is critiqued and re-fed into the generator for a polished second iteration.

A more detailed overview of the tasks in this stage is given below.

**Task 1: Section Selection.** After extracting the texts (Section 3), we used a cosine similarity-based approach to identify the most relevant sections for generating Future Work. We calculated the cosine similarity between each section and the paper's Future Work text (Table 14, Appendix) across all papers, and we selected the top three sections with the highest average similarity. These sections formed the basis for generating Future Work content. For all experiments, we removed the author's mentioned future work from the paper and made the input text.

**Task 2: Generating *Future Work* Using LLMs and RAG.** Generating Future Work from a single paper may lead to narrow, redundant, or overly specific suggestions, missing broader research trends and interdisciplinary insights. For example, a zero-shot LLM often defaults to boilerplate suggestions like "explore more data" or "apply to new domains." RAG combats this by feeding the model real, paper-specific passages at generation time, anchoring its outputs in concrete evidence and substantially reducing hallucinations. Moreover, by incorporating content from related papers, RAG ensures that future-work sections remain contextually relevant, tailored to genuine research priorities. To generate the Future Work of a paper, we experimented with the top three sections selected in Task 1, and all sections (excluding future work) integrating a RAG system, which consists of 100 research papers (Details 5.1). The Retriever processes the input query, which includes the prompt and the content of the research paper. It then retrieves additional relevant information from the vector database using cosine similarity and applies a ranking mechanism to prioritize the extracted text (Figure 2). The final augmented input is fed into the LLM Generator, producing Future Work content based on the provided context (Figure 2, step 6).

**Task 3: Evaluation and Iterative Refinement.** We employed several NLP-based quantitative metrics and used an **LLM as a Judge** to assess LLM-generated text against the ground truth Future Work based on coherence, relevance, readability, grammar, overall quality, and novelty (Figure 2, steps 8 and 9). Each evaluation was rated on a discrete scale of 1 (worst) to 5 (best), with justifications provided except novelty which is 1-10. We set a *threshold* of **3** as an acceptable midpoint, 7 for novelty. If an LLM-generated Future Work received a score less than or equal to the midpoint in any metric, the justification was incorporated into the prompt, and the Future Work was regenerated accordingly. This iterative refinement process was repeated up to two times to assess whether performance improved (Figure 2, step 9, 10, and 11).

| Meth. | Abs. | Intro. | RW | Data | Meth. | Exp. | Con. | Lim. |
|---|---|---|---|---|---|---|---|---|
| CS | **25.08** | **24.52** | 21.09 | 8.96 | 14.82 | 21.79 | **22.0** | 21.38 |

Table 2: Average Cosine Similarity of each section with Future Work in ACL data. N.B: CS, Abs, Intro, RW, Data, Meth, Exp, Con, Lim means Cosine Similarity, Abstract, Introduction, Releated Work, Data, Methodology, Experiment, Conclusion, and Limitations, respectively.

## 5 Experimental Setup

### 5.1 Generating Topics and Titles of Future Work

Our first work, trend analysis, we leverage GPT-4o-mini's deep semantic understanding to produce concise, human-readable topic titles on top of BERTopic's clustering. We set the minimum similarity threshold to 0.75 to ensure that each

| Metrics | w/o LLM's feed | w LLM's feed |
|---|---|---|
| ROUGE-1 | 24.33 | **25.74**(+1.41) |
| ROUGE-2 | 5.27 | **7.85**(+1.05) |
| ROUGE-L | 17.24 | **20.25**(+2.58) |
| BScore(f1) | 87.23 | **87.50**(+0.27) |
| Jaccard S | 15.40 | **18.13**(+2.73) |
| Cosine S | 48.07 | **57.03**(+8.96) |
| BLEU | 2.38 | **5.74**(+3.36) |
| Coherence | 3.94 | **3.97**(+0.03) |
| Relevance | 4.07 | **4.62** (+0.55) |
| Readability | 3.19 | **3.49** (+0.30) |
| Grammar | 4.02 | **4.01** (-0.01) |
| Overall | 3.85 | **3.95** (+0.10) |

Table 3: Comparison of performance without and with LLM feedback using GPT-4o mini in ACL data. Metrics include Jaccard Similarity (Jaccard S) and Cosine Similarity (Cosine S).

topic cluster remains semantically tight (boosting coherence) without collapsing into overly narrow groups. After benchmarking several sentence-transformer embeddings, we selected all-MiniLM-L6-v2 for its superior coherence. Alternative setups—including KMeans clustering (15–30 clusters) and UMAP dimensionality reduction (5–20 neighbors/components)—consistently produced lower silhouette and topic coherence scores (Figure 1). This LLM-augmented, prompt-light setup gives us automatically generated, high-quality topic labels at scale, with no manual annotation or prompt-tuning overhead.

Our second work is future work generation, where we benchmarked a diverse mix of generative and retrieval-augmented methods to understand their relative strengths under realistic constraints. BART and T5 serve as strong, well-studied seq2seq baselines with fixed token-limit trade-offs, while GPT-3.5 and GPT-4o illustrate how off-the-shelf LLMs perform zero-shot under a controlled prompt budget. Fine-tuning LLaMA-3.1 with LoRA/QLoRA and FlashAttention demonstrates that even large open-source models can be adapted efficiently on modest hardware. Finally, integrating RAG grounds generation in concrete evidence, and comparing one-shot versus zero-shot LLM evaluators (GPT vs. Llama) lets us quantify both generation quality and evaluator bias. This multi-axis evaluation ensures our conclusions generalize beyond any single model or configuration.

**Generating Similarity between Other Sections and Future Works:** We used Sentence Transformers ('all-MiniLM-L6-v2') for embedding generation and scikit-learn's [3] cosine similarity function for similarity computation.

**Fine-Tuning Models:** First, we fine-tuned BART (1,024-token limit) and T5 (512-token limit) on a 70/30 train/test split, discarding any over-length inputs. For LLaMA 3.1 7B fine-tuning, it was trained using 'Abstract,' 'Introduction,' and 'Conclusion' as input, and the extracted ' Future Work' was used as output. Leveraging LLaMA's extensive pretraining, we fine-tuned it with QLoRA (4-bit, alpha=16) and FlashAttention on a 30/70 train/test split for 60 steps (learning rate 2e-4, 2048-token context). During testing, we provided a prompt with detailed instructions for generating Future Work, with a maximum output length of 128 tokens and a temperature setting of 1.

**Zero/Few Shot(s) LLM:** We ran GPT-3.5 and GPT-4o in zero-shot mode, leveraging their 16 K and 128 K context windows, respectively. For LLM as a judge and LLM as an extractor, we used a one-shot approach with GPT-4o mini and a zero-shot approach with GPT-4o mini, respectively.

**LLM with RAG Integration:** We integrated RAG with GPT-4o mini for Future work generation, leveraging OpenAI's 'text-embedding-3-small' model for relevant document retrieval. Our vector database comprises 100 randomly selected papers from the dataset, and these papers were removed from the dataset. For semantic search using vector embeddings, we employed LlamaIndex, utilizing its in-memory vector store rather than an external database. We used a hybrid retriever system to fetch the data from the vector database consisting of BME and FAISS with an equal weight of 50%. We segmented the data into chunks of up to 512 tokens to accommodate smaller context windows. The overall context window was set to 3,900 tokens—the maximum number of tokens the LLM can process at a time. Additionally, we set $K = 3$, meaning that the top three most relevant chunks are retrieved from the vector store to provide contextual support during the generation process.

## 6 Experiments and Results

### 6.1 Evaluation of Future Work Extraction

After extracting the author mentioned future work from a paper using a tool (Tool extracted), we send it to the LLM to re-extract the future work by removing noisy sentences (GPT Extracted). We measured NLP-based evaluation (Rouge, BERTScore, BLEU) and LLM-based evaluation (Coherence, Relevance, etc.) between LLM-generated text and

[3] https://scikit-learn.org/stable/

| GT | Iter. | R-1 | R-2 | R-L | BS | CS | JS | Bl | CS$_T$ | JS$_T$ | Coh | Rel | Read | Gram | Nov | Ov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GPT 4o mini** | | | | | | | | | | | | | | | | |
| FW | 1 | 26.03 | 10.28 | 15.54 | 85.02 | 68.23 | 11.21 | 2.29 | 33.02 | 74.11 | 4.31 | 4.63 | 3.82 | 4.81 | 7.28 | 4.29 |
| OR | 1 | 26.03 | 6.97 | 12.34 | 82.94 | 62.56 | 10.43 | 1.26 | 18.75 | 62.07 | 3.67 | 4.37 | 3.19 | 4.18 | 7.84 | 3.65 |
| FW+OR | 1 | **37.98** | **12.94** | **18.32** | **85.20** | **74.33** | **14.46** | **3.55** | **33.67** | **76.78** | 4.41 | **4.87** | 3.85 | 4.67 | 7.28 | 4.40 |
| FW+OR | 2 | 37.35 | 12.14 | 17.43 | 84.63 | 73.50 | 14.29 | 3.42 | 31.58 | 68.42 | 4.48 | 4.79 | 4.15 | 4.89 | **8.00** | 4.44 |
| **GPT 4o mini + RAG** | | | | | | | | | | | | | | | | |
| FW | 1 | 21.79 | 6.14 | 11.89 | 83.00 | 63.27 | 8.72 | 1.08 | 21.01 | 62.79 | 4.26 | 4.04 | 4.87 | **5.00** | 7.70 | 4.15 |
| OR | 1 | 26.12 | 6.72 | 12.43 | 82.81 | 62.91 | 10.15 | 1.01 | 18.54 | 51.67 | **4.56** | 4.56 | **4.89** | **5.00** | 7.80 | **4.52** |
| FW+OR | 1 | 34.28 | 9.57 | 15.90 | 83.47 | 72.57 | 12.29 | 2.02 | 24.16 | 62.97 | 4.17 | 4.36 | **4.89** | **5.00** | 7.11 | 4.38 |

Table 4: Performance comparison between different methods in NeurIPS data. Here, GT, FW, OR, Iter means Ground Truth, Author Mentioned Future Work, OpenReview suggestion, and Iterations, respectively. Metrics include ROUGE (R-1, R-2, R-L), BERTScore (BS), Cosine Similarity (CS), Jaccard Similarity (JS), BLEU (Bl), Coherence (Coh), Relevance (Rel), Readability (Read), Grammar (Gram), Novelty (Nov), Overall (Ov), and Similarity (Sim).

| Model | It. | Coh. | Rel. | Read. | Gr. |
|---|---|---|---|---|---|
| Llama 3 ZS | 1 | 3.83 | 4.47 | **4.04** | 3.44 |
| Llama 3 FT | 1 | 3.37 | 3.69 | 2.90 | 3.65 |
| GPT 3.5 | 1 | 3.89 | 4.47 | 3.38 | 4.04 |
| GPT 4om | 1 | 3.94 | 4.07 | 3.19 | 4.02 |
| GPT 4om + RAG | 1 | 3.93 | 3.96 | 3.18 | 4.01 |
| GPT 4om + RAG | **2** | **3.97** | **4.50** | 3.50 | **4.06** |
| GPT 4om + RAG | 3 | 3.96 | 4.34 | 3.36 | 4.05 |

Table 5: Performance of different models in generating Future Work, evaluated against paper ground truth extracted by LLM (GPT extracted Ground Truth) in ACL data. Note: GPT 4om refers to GPT-4o mini, ZS indicates Zero-Shot, FT represents Fine-Tuning, It. denotes Iteration, Coh. stands for Coherence and Logic, Rel. for Relevance and Accuracy, Read. for Readability and Style, and Gr. for Grammatical Correctness.

Tool extracted Ground Truth (GT). We also evaluated LLM-generated text with GPT extracted GT. As shown in (Table 7), GPT extracted GT outperformed across all metrics, demonstrating improved text quality and relevance when using GPT 4o mini. Additionally, Table 6 confirms that GPT extracted GT yields better performance than Tool extracted GT across different models when comparing the same model's Tool extracted GT. The reason is that Tool extracted GT often contains noisy sentences, and not all of the sentences are related to Future Work. Where GPT extracted GT conatins only sentences related to Future Work, achieves a high alignment with LLM generated text. We further conducted a human evaluation to evaluate the text extraction process by GPT in section 3.1.

## 6.2 Evaluation of Future Work Generation

We experimented with various LLMs (Table 6) to generate Future Work sections. Traditional NLP-based metrics primarily focused on lexical overlap and semantic quality, and LLM-based evaluation provides more contextual assessment and novelty. NLP-based metrics couldn't measure novelty and relied more on ground truth, missing the depth evaluation. Solely relying on an LLM-based evaluation system raises potential bias issues. To alleviate these problems, we incorporated both LLM-based and NLP-based evaluation systems to make a robust evaluation system for evaluation by focusing on coherence and logic, relevance and accuracy, readability and style, grammatical correctness, overall impression, and novelty. Here, novelty measures how novel the generated future work is in terms of the ground truth (higher values indicate more original ideas).

## 6.3 Future Work Generation

**Evaluation-considering all sections vs top-3 sections.** We used cosine similarity to select the three most relevant sections as input in ACL data. We compared the performance considering 3 sections and all sections (excluding the 'author's mentioned future work'). As shown in Table 7 and Table 6, using the top three sections led to a slight performance drop in every metric, showing that the top 3 sections contained sufficient information to generate meaningful future work.

**Evaluation-incorporating LLM feedback.** After generating the ' Future Work' text, a separate LLM served as a judge, scoring the output on a 1–5 scale with justification (feedback). This feedback was then integrated into the prompt to refine the generated text further. We evaluated the LLM-generated text after each iteration with GPT GT, and our results show a significant performance improvement across all metrics after incorporating

| Model | Sec. | Iter. | R-1 | R-2 | R-L | BS | JS | CS | Bl |
|---|---|---|---|---|---|---|---|---|---|
| **Tool extracted Ground Truth** | | | | | | | | | |
| BART | 3 | 1 | 15.29 | 26.3 | 4.83 | 84.53 | 11.33 | 35.39 | 0.6 |
| GPT 3.5 | All | 1 | 21.08 | 6.41 | 17.03 | 86.40 | 14.54 | 45.69 | 4.13 |
| GPT 3.5 | 3 | 1 | 20.03 | 5.36 | 16.65 | 86.28 | 13.70 | 44.84 | 3.18 |
| GPT4om | 3 | 1 | 17.50 | 3.11 | 14.22 | 85.91 | 11.55 | 40.53 | 1.29 |
| GPT4om+R | 3 | 1 | 21.20 | 5.09 | 16.41 | 86.40 | 13.27 | 40.36 | 2.84 |
| **GPT Extracted Ground Truth** | | | | | | | | | |
| GPT 3.5 | 3 | 1 | **28.17** | **8.68** | 20.64 | 87.88 | 18.49 | 55.36 | 5.61 |
| GPT 3.5 | 3 | 2 | 27.80 | 8.67 | 20.49 | 87.88 | 18.66 | 56.72 | 5.81 |
| Llama 3 Z | 3 | 1 | 14.15 | 3.23 | 10.38 | 85.61 | 14.52 | 52.27 | 2.27 |
| Llama 3 F | 3 | 1 | 26.75 | 5.61 | - | 85.54 | 12.28 | 45.37 | 1.14 |
| GPT 4om | 3 | 1 | 24.59 | 5.62 | 17.69 | 87.44 | 15.87 | 50.52 | 2.72 |
| GPT4om+R | 3 | 1 | 21.13 | 5.52 | 14.19 | 87.67 | 16.95 | 49.32 | 3.42 |
| GPT4om+R | 3 | 2 | 26.58 | 7.97 | **20.87** | **88.15** | **18.67** | **58.33** | **5.67** |
| GPT4om+R | 3 | 3 | 26.96 | 7.53 | 20.18 | 87.94 | 17.92 | 56.34 | 4.68 |

Table 6: Performance comparison of various models in generating Future Work, considering Extracted Future Work from papers using Tool extracted Future Work and GPT-extracted Future Work in ACL data. Note: GPT 4om, R, Llama 3 Z, F, Iter. denotes GPT 4o mini, RAG, Llama 3 Zero shot, Fine Tuning, Iteration, Sec. refers to Section, R-1, R-2, R-L, BS, JS, CS, Bl represent ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, Jaccard Similarity, Cosine Similarity, Bleu respectively.

| Metrics | All sec | 3 sec. | Tool extracted GT | GPT extracted GT |
|---|---|---|---|---|
| ROUGE-1 | **21.08** | 20.3(-0.78) | 17.50 | **24.59(+7.09)** |
| ROUGE-2 | **6.41** | 5.36(-1.05) | 3.11 | **5.62(+2.51)** |
| ROUGE-L | **17.03** | 16.65(-0.38) | 14.22 | **17.69(+3.47)** |
| BScore(f1) | **86.40** | 86.28(-0.12) | 85.91 | **87.44(+1.53)** |
| Jaccard S | **14.54** | 13.70(-0.84) | 11.55 | **15.87(+4.32)** |
| Cosine S | **45.69** | 44.84(-0.85) | 40.53 | **50.52(+9.99)** |
| BLEU | **4.13** | 3.18(-0.95) | 1.29 | **2.72(+1.43)** |

Table 7: Performance comparison of GPT-3.5 in generating Future Work using three selected sections versus full-text input, evaluated against both Tool extracted Ground Truth (GT) and GPT-extracted ground truth on ACL data.

LLM feedback once (Table 3). Moreover, from Table 6, we observed that after the first feedback loop (iteration 2), using GPT-4o mini + RAG, performance improved across nearly all metrics. However, in iteration 3, performance declined when the model received feedback a second time, indicating that excessive iterations introduced bias. These findings suggest that a single round of feedback integration is optimal for improving model performance without introducing unwanted biases. In Table 4, using the GPT 4o mini model, we added self-feedback in a second iteration in NeurIPS data, excluding the NLP metrics, other LLM-based metrics such as coherence, relevance, and novelty improved the results. In the first iteration, the novelty score was 7.28, indicating a strong similarity between the LLM-generated future work and the ground truth. In the second iteration, we incorporated self-feedback, which helped uncover more underexplored ideas across multiple directions and resulted in increased +0.72 (Table 4). In GPT 4o mini + RAG settings, all LLM-based metrics exceeded the threshold of 3 on the combined Future Work + OpenReview (FW + OR) ground-truth set, so we did not apply additional feedback in this case.

**Evaluation-among all models.** We applied various models, and among all of the models, GPT 4o mini with RAG performs best when LLM feedback goes to the model (iteration 2) when considering 'GPT extracted Ground Truth' in NLP-based metrics (Table 6) and LLM based metrics (Table 5). GPT 4o mini + RAG achieved better results than zero-shot GPT 4o mini, except n-gram based methods (Rouge-1,2, L) due to the introduction of new information from RAG (Table 6). In NeurIPS data (Table 4), incorporating RAG with GPT 4o mini didn't increase results in NLP metrics, but it boosts performance on some LLM-based metrics.

**Evaluation incorporating OpenReview** We evaluated our approach on NeurIPS data using three ground-truth sources: Author-Mentioned (FW), OpenReview (OR), and their combination (FW+OR), as shown in Table 4. Incorporating OpenReview feedback (FW + OR) in ground truth yielded improvements across all metrics compared to other ground truths (FW, OR) in both models GPT 4o mini and GPT 4o mini + RAG. Adding OpenReview (OR) with future work in ground truth achieves lower novelty, meaning most LLM-generated ideas are covered.

**LLM as a Judge** We assessed the generated suggestions against the ground truth using six criteria—coherence, relevance, readability, grammar, overall impression, and novelty—by treating an LLM as the evaluator. Our primary judge was GPT-4o-mini, and to mitigate any generator–evaluator bias, we also tried Llama 3 70B. However, we found that Llama 3 tended to assign overly high scores even to low-quality outputs.

**Hallucination Rate** We evaluated hallucinations by treating each LLM-generated text as the hypothesis and the concatenation of the original input text and ground-truth future work as the premise. For each pair, we asked GPT-4o-mini to classify the relationship as `entailment`, `neutral`, or `contradiction`. Using RAG, GPT-4o-mini's hallucination rate dropped to 19.52%, compared to the higher rate observed without RAG (Table 8).

| Model | Hallucincation Rate |
|---|---|
| GPT 4o mini | 26.26 % |
| GPT 4o mini+RAG | **6.74** % |

Table 8: Hallucination rate of each model in NeurIPS data (lower is better)

| Ques. | Rating (Avg.) | W. Kappa | Kend' Tau |
|---|---|---|---|
| Q1 | 2.54 | 0.24 | 0.29 |
| Q2 | 2.12 | **0.30** | **0.36** |
| Q3 | **2.75** | 0.28 | 0.29 |

Table 9: Average rating and annotators user agreement on user study.

## 6.4 Human Evaluation

Using the same LLM as extractor, generator, and evaluator may raise potential biases such as self-validation bias and confirmation bias. To overcome this problem, we compared LLM-based evaluation with human judgments and found a strong relationship. We conducted these questions with a human: **Q2:** How good is LLM's generated Future Work based on the ground truth? **Q3:** How does the LLM Feedback approach improve performance in terms of originality? In Table 11, the average rating across all questions is more than 2, reflecting an overall average performance. Comparing LLM-generated Future Work with ground truth (Q2) shows average performance (2.12) supported by strong kappa and Kendall's tau scores. Notably, in Q3, our proposed approach with one feedback iteration achieves an average human rating of 2.75, demonstrating that the LLM feedback loop significantly enhances result quality.

| Future Work |
|---|
| Future Research in Natural Language Processing and Model.. |
| Future Work in Natural Language Processing and Image Captioning |
| Future Directions in NLP Model Development and Sentiment Analysis |
| Sentiment Analysis and Dispute Detection Models in Conversational |
| Future Directions in Language Modeling and Machine Translation |
| Future Research Directions in NLP Models and Techniques |
| Future Directions in Model Development for Multilingual.. |
| Neural Machine Translation and Future Research Directions in NLP |
| Future Research Directions in NLP Models and Methods |
| Future Research Directions in Neural Machine Translation and .. |
| Evaluation Frameworks and Future Work in Generalization and .. |
| Future Directions in NLP Model Evaluation and Bias |
| Future Directions in Language Model Research and Performance .. |

Table 10: Prominent Future Work in each year (2012-2024) in ACL data.

## 6.5 Trends in NLP Future Work

We analyzed the evolution of future work directions in NLP from 2012 to 2024 (Table 15, 13, Appendix). The early years emphasized neural machine translation, model improvement, and basic future work identification. From 2017 to 2019, there was a strong focus on language modeling, machine translation, and multilingual text representation, reflecting the rise of deep learning in NLP. From 2020 to 2021, research has expanded towards multi-scale learning techniques and generalization. In recent years, evaluation frameworks, bias mitigation, and performance robustness have increasingly been prioritized, highlighting the community's shift toward more reliable, explainable, and fair NLP models. Early research focused on statistical methods, machine translation, and model improvement. Still, recent trends emphasize LLMs, prompt engineering, and data efficiency, also increasingly emphasize model performance robustness, privacy, adversarial attacks, generalization, and evaluation in NLP models (Figure 3, Table 20, 13 Appendix). From the topic analysis from the recent year (ACL 2024), we can see that the prominent topics are 'linguistic,' 'LLM,' 'privacy,' 'annotation,' etc. (Table 18, Appendix). Researchers also emphasize the need for better evaluation methods to assess and mitigate model biases, fairness, and societal implications of NLP models. Ensuring NLP models are trustworthy, explainable, and interpretable evaluation frameworks for document generation and annotation quality. Over time, the focus has shifted from a narrow concentration on model and translation improvements toward more nuanced themes that encompass ethical considerations, interdisciplinary approaches, and real-world applicability (Table 19, Appendix). In addition, Future research increasingly explore connections of interdisciplinary approaches between AI and human cognition.

| Ques. | Rating (Avg.) | W. Kappa | Kend' Tau |
|---|---|---|---|
| Q1 | 2.54 | 0.24 | 0.29 |
| Q2 | 2.12 | **0.30** | **0.36** |
| Q3 | **2.75** | 0.28 | 0.29 |

Table 11: Average rating and annotators user agreement on user study.

## 7 Discussions

Our framework is readily extensible to other fields, such as biomedicine or social sciences, by harnessing peer-review feedback to build a robust ground truth, constructing domain-specific retrieval corpora, applying LLM-based evaluation metrics (including novelty assessments), and iteratively refining outputs via LLM self-feedback. Moreover, our experiment shows that GPT-extracted Ground Truth aligns well with LLM-generated Future Work text, indicating that the Tool extracted Ground Truth contains more noise and irrelevant sentences. Our proposed approach can extract Future Work-related sentences from papers that lack explicit Future Work sections. Also, incorporating RAG with GPT-4o mini increased performance in most of the metrics, except the n-gram overlaps and reducing hallucinations. Notably, one-time LLM feedback improves the performance of LLM when considering the GPT extracted Ground Truth. To make generalizability, our model showed that incorporating a single round of feedback led to significant improvements. However, applying a second round of feedback negatively impacted performance, producing a more biased response. Also, considering the top 3 sections instead of all sections results in minor performance drops. Moreover, we incorporated long-term goal suggestions from peer review from OpenReview to make a robust ground truth, which increased the performance in every metric. LLM shows strong relations with human evaluators in the extractor, generator, and feedback provider tasks.

## 8 Conclusions

The Future Work section is a forward-looking guide, helping the research community explore new directions. We utilized an LLM to extract Future Work, producing a more coherent ground truth that enhances model performancea and incorporated a strong ground truth from OpenReview. Additionally, we integrated an external vector database to further improve LLM's performance to generate Future Work from input text. For evaluation, we applied NLP-based metrics alongside an LLM-as-a-Judge approach, using explainable LLM metrics to assess performance, provide feedback, and iteratively refine text generation. Furthermore, we conducted a 13-year trend analysis to examine how Future Work priorities have evolved over time, highlighting key shifts in research focus.

## Limitations and Future Work

Our analysis is confined to ACL papers (2012–2024) and NeurIPS (2021-22), which ensures domain relevance but limits cross-disciplinary generality. RAG retrieval was capped at 100 papers for cost reasons, and using the same LLM as both generator and evaluator may introduce bias. We treated LLM-extracted future work as ground truth and assumed that LLM-driven feedback and suggestions align with that standard, despite relying on a small pool of annotators and samples. Our evaluation criteria, coherence, relevance, readability, and grammar, do not fully capture scholarly attributes like originality or robustness, and iterative refinement with a single model risks stylistic convergence and potential hallucination. We limited T5 and BART inputs to 512 and 1,024 tokens and did not explore advanced feedback strategies (e.g., chain-of-thought, self-consistency). Taking 100 random papers from the RAG database can raise issues, such as generating future work of a paper where the data comes from after the publication date of the paper. To alleviate this problem, in the future we will take the related papers which is published before the actual paper. Also, in the future, we will extend our pipeline to additional research domains and improve methods for extracting implicit future-work mentions, evaluate open-source LLMs to reduce API costs, and build a large-scale domain-specific vector store to enhance RAG retrieval. We also plan to involve more human annotators to validate our extractor–generator–feedback loop and create a gold-standard dataset, mitigate evaluator bias through RLHF and RLAIF, and integrate advanced reasoning techniques while incorporating cited literature and reviewer perspectives to diversify and enrich future-work suggestions.

## Ethics Statement

Our work involves extracting and generating future work sections from scientific articles, raising important ethical considerations regarding intellectual property, authorship, and responsible AI use. Our approach does not entirely rely on the automated generation of research papers; our framework uses the process to identify potential directions or gaps in the research. We also incorporated external evaluation metrics and human feedback to assess the quality of the generated content. It helps mitigate risks like overfitting to internal metrics or biases inherent in the LLM, ensuring that the final content is both robust and original. To ensure compliance with ethical research practices and academic integrity, we adhere to the following principles: **1. Respect for Original Authorship.** We do not claim authorship of the generated future work. Instead, our method functions as an assistive tool for analyzing and refining research trends, complementing rather than replacing human intellectual contributions. **2. Fair Use and Transparency.** The Extracted text is used only for research and analysis purposes. It is neither directly republished nor misrepresented, and we acknowledge the source articles when applicable to maintain full transparency. **3. Responsible AI Use & Avoiding Misleading Content.** Our LLM-based generation process is designed to refine and organize existing content rather than fabricate entirely new research directions. We employ a RAG framework along with iterative refinement mechanisms to ensure that generated future work remains firmly anchored in the paper's actual contributions. This strategy prevents speculative or arbitrary directions and mitigates potential guideline violations. In Section 6.4, we detail a user study involving three annotators who assessed the effectiveness of our approach in generating future work based on the paper's mentioned content. Although the average rating of 1.92 on a scale from 1 (worst) to 3 (best) indicates moderate performance, it demonstrates that our method effectively grounds the output in validated, paper-specific information. **4. Non-Substitution of Human Contribution.** The generated future work is intended to assist researchers by organizing, summarizing, and clarifying potential research directions. Final decisions and substantive content remain the responsibility of human authors, thereby preserving the integrity of scholarly work. **5. Alignment with Ethical Standards** Our approach aligns with ACL's ethical research guidelines and maintains rigorous academic integrity standards. We do not automate the entire writing process; instead, we offer a tool to help researchers better structure and refine their future work suggestions.

## References

Herman Aguinis, Ravi S Ramani, and Nawaf Alabduljader. 2018. What you see is what you get? enhancing methodological transparency in management research. *Academy of Management Annals*, 12(1):83–110.

Ibrahim Al Azhar, Sohel Ahmed, Md Saiful Islam, and Aisha Khatun. 2021. Identifying author in bengali literature by bi-lstm with attention mechanism. In *2021 24th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Ibrahim Al Azher and Hamed Alhoori. 2024. Mitigating visual limitations of research papers. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8614–8616. IEEE.

Ibrahim Al Azher, Venkata Devesh Reddy, Hamed Alhoori, and Akhil Pandey Akella. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How ai ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. *arXiv preprint arXiv:2401.13481*.

Ibrahim Al Azher. 2024. Generating suggestive limitations from research articles using llm and graph-based approach. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–3.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Carrie Conaway, Venessa Keesler, and Nathaniel Schwartz. 2015. What research do state education agencies really need? the promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis*, 37(1_suppl):16S–28S.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2018. A deep learning approach for sentence classification of scientific abstracts. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 479–488. Springer.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Wenke Hao, Zhicheng Li, Yuchen Qian, Yuzhuo Wang, and Chengzhi Zhang. 2020. The acl fws-rc: A dataset for recognition and classification of sentence about future works. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 261–269.

Noriko Hara, Paul Solomon, Seung-Lye Kim, and Diane H Sonnenwald. 2003. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information science and Technology*, 54(10):952–965.

Hospice Houngbo and Robert E Mercer. 2012. Method mention extraction from scientific research papers. In *Proceedings of COLING 2012*, pages 1211–1222.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yue Hu and Xiaojun Wan. 2015. Mining and analyzing the future works in scientific articles. *ArXiv*, abs/1507.02140.

Adnan A Hyder, Adrijana Corluka, Peter J Winch, Azza El-Shinnawy, Harith Ghassany, Hossein Malekafzali, Meng-Kin Lim, Joseph Mfutso-Bengo, Elsa Segura, and Abdul Ghaffar. 2011. National policy-makers speak out: are researchers giving them what they need? *Health policy and planning*, 26(1):73–82.

Moksh Jain, Tristan Deleu, Jason Hartford, Cheng-Hao Liu, Alex Hernandez-Garcia, and Yoshua Bengio. 2023. Gflownets for ai-driven scientific discovery. *Digital Discovery*, 2(3):557–577.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *Ejifcc*, 25(3):227.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization. *arXiv preprint arXiv:2407.00747*.

David Nicholas, Hamid R Jamali, Anthony Watkinson, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. 2015. Do younger researchers assess trustworthiness differently when deciding what to read and cite and where to publish? *International Journal of Knowledge Content Development & Technology*, 5(2).

Justin C Ortagus, Robert Kelchen, Kelly Rosinger, and Nicholas Voorhees. 2020. Performance-based funding in american higher education: A systematic synthesis of the intended and unintended consequences. *Educational Evaluation and Policy Analysis*, 42(4):520–550.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

Yuchen Qian, Zhicheng Li, Wenke Hao, Yuzhuo Wang, and Chengzhi Zhang. 2021. Using future work sentences to explore research trends of different tasks in a special domain. *Proceedings of the Association for Information Science and Technology*, 58(1):532–536.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Müge Simsek, Mathijs de Vaan, and Arnout van de Rijt. 2024. Do grant proposal texts matter for funding decisions? a field experiment. *Scientometrics*, 129(5):2521–2532.

Ruoxuan Song, Li Qian, et al. 2021. Identifying academic creative concept topics based on future work of scientific papers. *Data Analysis and Knowledge Discovery*, 5(5):10–20.

Jacques Suray, Jan H. Klemmer, Juliane Schmüser, and Sascha Fahl. 2024. How the future works at soups: Analyzing future work statements and their impact on usable security and privacy research. *ArXiv*, abs/2405.20785.

Simone Teufel. 2017. Do" future work" sections have a purpose? citation links and entailment for global scientometric questions. In *BIRNDL@ SIGIR (1)*, pages 7–13.

Mike Thelwall, Subreena Simrick, Ian Viney, and Peter Van den Besselaar. 2023. What is research funding, how does it influence research, and how is it recorded? key dimensions of variation. *Scientometrics*, 128(11):6085–6106.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

Chengzhi Zhang, Yi Xiang, Wenke Hao, Zhicheng Li, Yuchen Qian, and Yuzhuo Wang. 2023. Automatic recognition and classification of future work sentences from academic articles in a specific domain. *Journal of Informetrics*, 17(1):101373.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zihe Zhu, Dongbo Wang, and Si Shen. 2019. Recognizing sentences concerning future research from the full text of jasist. *Proceedings of the Association for Information Science and Technology*, 56(1):858–859.

# A Appendix

| Metrics | RAW GT | GPT GT |
|---|---|---|
| ROUGE-1 | 17.50 | 24.59(+7.09) |
| ROUGE-2 | 3.11 | 5.62(+2.51) |
| ROUGE-L | 14.22 | 17.69(+3.47) |
| BScore(f1) | 85.91 | 87.44(+1.53) |
| Jaccard S | 11.55 | 15.87(+4.32) |
| Cosine S | 40.53 | 50.52(+9.99) |
| BLEU | 1.29 | 2.72(+1.43) |

Table 12: Comparison of performance between extracted ground truth from papers (RAW GT) and LLM-refined ground truth (GPT GT) using GPT-4o mini.

## A.1 Human Evaluation

We collected 200 random samples from the dataset and, for each sample, evaluated three key questions using Ground Truth, LLM-generated text, and LLM-regenerated text after feedback. Table 16 in the Appendix shows that each question is divided into two columns (A and B) representing different text types. Specifically, in Q1, human annotators compared the future work extracted directly from the paper (Ground Truth) using tools with the future work obtained via LLM extraction, assessing the LLM's ability to accurately extract content without generating new text. In Q2, the evaluation focused on how well the LLM generated future work

| **Topics** | **Year** |
|---|---|
| Model Improvement (31), Diffusion (13) | 2012 |
| Image Captioning (54), Machine Translation (syntax) (12) | 2013 |
| Sentiment Analysis (11) | 2014 |
| Machine Translation (20), Opinion Mining (10) | 2015 |
| NLP models (85) | 2016 |
| Multilingual Text (110) | 2017 |
| Neural Machine Translation (216) | 2018 |
| NLP Models (265) | 2019 |
| Neural Machine Translation (49), Bias in Language Models (LM) (21), Document Generation (21), Conversational QA (19), Dialogue Evaluation (16), Adversarial Attacks and (13), Named Entity Recognition (12) | 2020 |
| Generalization and Robustness (36), Multilingual Neural Machine Translation (33), Robustness (28), Conversational AI Evaluation (26) | 2021 |
| Machine Translation (63), Model Training and Evaluation (62), Multidisciplinary (30), NLP and Event Detection (28), Multilingual (27), Knowledge Augmentation in LM (24), Annotation Quality (19), Summarization in Educational Content (14), Funding (13) | 2023 |
| LM Performance Evaluation (147), Performance and Data Evaluation (61), Advancing LLMs (66), Logic and Relationships (41), Model Performance, Privacy and Model Defense (28), Common Sense Reasoning and Evaluation Methods (19), Funding (11) | 2024 |

Table 13: Prominent Topics of Each Year (2012-2024). Note: Number means how many time these topics occurred.

| **Model** | **Abs.** | **Intro.** | **RW** | **Data** | **Meth.** | **Exp.** | **Con.** | **Lim.** |
|---|---|---|---|---|---|---|---|---|
| Cos. Sim | **25.08** | **24.52** | 21.09 | 8.96 | 14.82 | 21.79 | **22.0** | 21.38 |

Table 14: Average Cosine Similarity of each section with future work. (Abs., Intro., RW, Data, Meth., Exp., Con., and Lim. indicate Abstract, Introduction, Related Work, Dataset, Methodology, Experiment and Results, Conclusion, and Limitation

from input texts (Abstract, Introduction, and Conclusion), with the paper mentioning future work (ground truth). Finally, Q3 examines the impact of incorporating an iterative feedback loop by comparing the output of the LLM + RAG method with and without the additional feedback iteration. This comparison highlights how the feedback mechanism contributes to refining and improving the final output.

| Dataset | future work |
|---|---|
| 2012 | Future Research in Natural Language Processing and Model Improvement Techniques |
| 2013 | future work in Natural Language Processing and Image Captioning |
| 2014 | Future Directions in NLP Model Development and Sentiment Analysis |
| 2015 | Sentiment Analysis and Dispute Detection Models in Conversational Text |
| 2016 | Future Directions in Language Modeling and Machine Translation |
| 2017 | Future Research Directions in NLP Models and Techniques |
| 2018 | Future Directions in Model Development for Multilingual Text Representation |
| 2019 | Neural Machine Translation and Future Research Directions in NLP Models |
| 2020 | Future Research Directions in NLP Models and Methods |
| 2021 | Future Research Directions in Neural Machine Translation and Multi-Scale Learning Techniques |
| 2022 | Evaluation Frameworks and future work in Generalization and Robustness |
| 2023 | Future Directions in NLP Model Evaluation and Bias |
| 2024 | Future Directions in Language Model Research and Performance Evaluation |

Table 15: Prominent future work in Each year.

| Quest | Column A | Column B |
|---|---|---|
| Q1 | Ground Truth (Extracted text from paper using parsing tool) | Ground Truth (Extracted text using LLM) |
| Q2 | Ground Truth future work | LLM Generated future work |
| Q3 | Generated future work by LLM + RAG | Generated future work by LLM + RAG with Feedback |

Table 16: Human Evaluation Details

| Model | Coherene | Relevance | Readability | Overall Impression |
|---|---|---|---|---|
| GPT 4o mini | 3.89 | 4.47 | 3.38 | 3.82 |
| Llama 3-80b | 4.62 | 4.37 | 4.81 | 4.49 |

Table 17: Performance between considering GPT 4o mini vs Llama-3 80b as a Judge

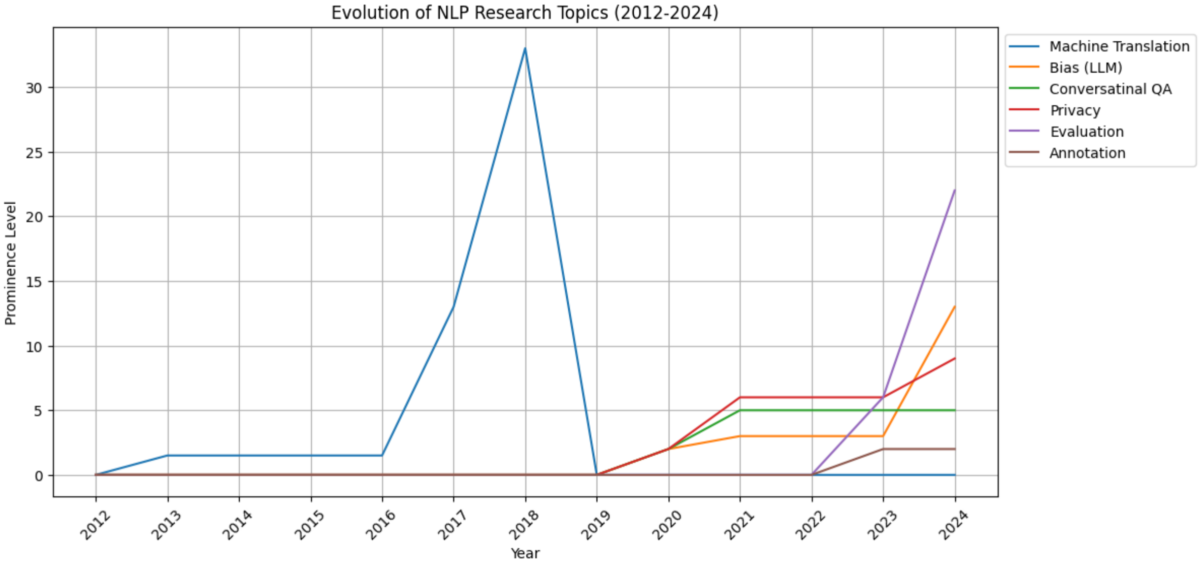

Figure 3: Evolving future work Over Years (2012-2024)

| **future work Title** |
|---|
| euphemisms, euphemism', turkish, linguistic, multilingual |
| aspectual, llm, feature, llms, hyperparameters |
| benchmarks, pipeline, hyperparameters, modeloutput, accuracy |
| future, goals, satisfaction, achievement, achieve |
| privacy, datasets, annotation, data, identifiable |
| semantic, commonsense, semantics, annotation, schema |
| ai, mobility, future, intelligence, nrrp |

Table 18: Prominent future work topics of ACL 2024 papers.

| **future work Title** | **Count** |
|---|---|
| **2024** | |
| Future Directions in Language Model Research and Performance Evaluation | 147 |
| Future Research Directions for Advancing LLMs and Prompt Engineering | 66 |
| Future Research Directions in Model Performance and Data Efficiency | 61 |
| Future Research on Logic and Relationships | 41 |
| Future Directions in Privacy and Model Defense Research | 29 |
| Future Directions in Commonsense Reasoning and Evaluation Methods | 19 |
| Project Funding and Support for AI Research | 11 |
| **2023** | |
| Simultaneous Machine Translation Research | 63 |
| Future Directions in Model Training and Evaluation | 62 |
| Future Research Directions in Multidisciplinary Approaches | 30 |
| Future Directions in NLP and Event Detection | 28 |
| Future Directions in Multilingual Research and Model Evaluation | 27 |
| Future Directions in Knowledge Augmentation for Language Models | 24 |
| Annotation Quality and Future Improvements in Predictive Tasks | 19 |
| Future Directions in Lay Summarization and Educational Content Analysis | 14 |
| AI Research Funding and Support Initiatives | 13 |
| **2021** | |
| Evaluation Frameworks and future work in Generalization and Robustness | 36 |
| Future Directions in NLP Research and Applications | 35 |
| Future Directions in Multilingual Neural Machine Translation Research | 33 |
| Robustness and Learnability in NLP Models | 28 |
| Future Directions in Conversational AI and Evaluation Systems | 26 |
| Future Multimodal Research in Visual Tasks and Event Predictions | 19 |
| **2020** | |
| Future Research Directions in Neural Machine Translation and Multi-Scale Learning Techniques | 49 |
| Lack of future work Mentioned in Provided Texts | 26 |
| Future Research on Language Models and Biases in Text Analysis | 21 |
| Future Research Directions in Document Generation and Understanding | 21 |
| Knowledge Base Question Answering and Future Directions in Conversational QA Systems | 19 |
| Future Directions in Open-Domain Dialogue Evaluation and Model Development | 16 |
| NLP Vulnerability and Robustness in Adversarial Attacks and Counterspeech Generation | 13 |
| Discontinuous Named Entity Recognition and Ontology Alignment in Information Extraction | 12 |
| **2019** | |
| Future Research Directions in NLP Models and Methods | 265 |
| future work Identification in Texts | 27 |
| **2018** | |
| Neural Machine Translation and Future Research Directions in NLP Models | 216 |
| future work Identification and Analysis | 29 |
| **2017** | |
| Future Directions in Model Development for Multilingual Text Representation | 110 |
| future work Identification and Analysis | 15 |
| **2016** | |
| Future Research Directions in NLP Models and Techniques | 85 |
| Lack of Identified future work in Provided Texts | 13 |
| **2015** | |
| Future Directions in Language Modeling and Machine Translation | 20 |
| Opinion Mining and Relation Modeling in NLP | 10 |
| **2014** | |
| Future Directions in NLP Model Development and Sentiment Analysis | 11 |
| **2013** | |
| future work in Natural Language Processing and Image Captioning | 54 |
| future work in Syntax-Based Statistical Machine Translation Improvements | 12 |
| **2012** | |
| Future Research in Natural Language Processing and Model Improvement | 31 |
| future work in Diffusion Prediction | 13 |

Table 19: Overview of recurring future work topics in ACL papers for each year, showing the number of times each topic appears. The count represents the frequency of mentions across various research papers.

| Hierarchical Future Directions | Category |
|---|---|
| **NLP Model Development** | **1** |
| Improve models and Techniques | |
| Sentiment Analysis | |
| Language Modeling and Machine Translation | |
| Model Development for Multilingual Text Representation | |
| Improve Neural Machine Translations (NMT) | |
| NMT with Multilingual Text Representation | |
| NMT with Multi-scale Learning Techniques | |
| Model Tuning with Multimodal Performance Analysis | |
| Language Model Research with Performance Evaluation | |
| Model Tuning with Multimodal Performance Analysis | |
| Language Model Research and Performance Evaluation | |
| Advancing LLMs and Prompt Engineering | |
| Model Performance and Data Efficiency | |
| **Machine Translation and Multilingual Models** | **2** |
| Syntax-Based Statistical Machine Translation Improvements | |
| **Sentiment Analysis and Relation Modeling** | **3** |
| Sentiment Analysis and Dispute Detection Models in Conversational Text | |
| Opinion Mining and Relation Modeling in NLP | |
| **Question Answering (QA) and Conversational Systems** | **4** |
| Knowledge Base Question Answering and Future Directions in Conversational QA Systems | |
| Future Directions in Open-Domain Dialogue Evaluation and Model Development | |
| Conversational AI and Evaluation Systems | |
| Retrieval-Based QA and Summarization | |
| **Robustness and Vulnerability in NLP** | **5** |
| NLP Vulnerability and Robustness in Adversarial Attacks and Counterspeech Generation | |
| Evaluation Frameworks and future work in Generalization and Robustness | |
| Robustness and Learnability in NLP Models | |
| Future Directions in Privacy and Model Defense Research | |
| **Information Extraction and Analysis** | **6** |
| Discontinuous Named Entity Recognition and Ontology Alignment in Information Extraction | |
| **Multimodal and Visual Tasks** | **7** |
| future work in Natural Language Processing and Image Captioning | |
| Future Multimodal Research in Visual Tasks and Event Predictions | |
| Future Research Directions in Multimodal Performance Analysis | |
| **Bias, Ethics, and Societal Impact** | **8** |
| Future Research on Language Models and Biases in Text Analysis | |
| Future Directions in NLP Model Evaluation and Bias | |
| **Document Generation and Annotation** | **9** |
| Future Research Directions in Document Generation and Understanding | |
| future work in Annotation Quality and Evaluation Techniques | |
| **Psychology and AI** | **10** |
| Future Research Directions and Methodological Extensions in AI and Psychology | |
| **Logic and Relationships in AI** | **11** |
| Future Research on Logic and Relationships | |
| **Evaluation Methods and Metrics** | **12** |
| Future Directions in Commonsense Reasoning and Evaluation Methods | |
| **Support and Funding for AI Research** | **13** |
| Project Funding and Support for AI Research | |

Table 20: Hierarchical future work.

Instructions:
You are provided with two texts for each pair: one is generated by a machine (Machine-Generated Text), and the other is the original or ground truth text (Ground Truth). Please read both texts carefully. After reviewing each text, assign a score from 1 to 5 based on the criteria outlined below. The score should reflect how well the machine-generated text compares to the ground truth, where 1 represents poor quality and 5 represents excellent quality that closely matches or even surpasses the ground truth in some aspects.

Scoring Criteria: Coherence and Logic:
5: The text is exceptionally coherent; the ideas flow logically and are well connected.
3: The text is coherent but may have occasional lapses in logic or flow.
1: The text is disjointed or frequently illogical.

Relevance and Accuracy:
5: The text is completely relevant to the topic and accurate in all presented facts.
3: The text is generally relevant with minor factual errors or slight deviations from the topic.
1: The text often strays off topic or includes multiple factual inaccuracies.

Readability and Style:
5: The text is engaging, well-written, and stylistically consistent with the ground truth.
3: The text is readable but may lack flair or have minor stylistic inconsistencies.
1: The text is difficult to read or stylistically poor.

Grammatical Correctness:
5: The text is free from grammatical errors.
3: The text has occasional grammatical errors that do not impede understanding.
1: The text has frequent grammatical errors that hinder comprehension.

Overall Impression:
5: The text is of a quality that you would expect from a professional writer.
3: The text is acceptable but would benefit from further editing.
1: The text is of a quality that needs significant revision to be usable.

Task:
For each text pair:
Rate the Machine-Generated Text on each criterion and provide a final overall score out of 5.
Provide a brief justification for your scores, highlighting strengths and weaknesses observed in the machine-generated text relative to the ground truth. ' ' '

---

Example of Usage: Text Pair 1:
Machine-Generated Text: "The quick brown fox jumps over the lazy dog repeatedly."
Ground Truth: "A quick brown fox consistently jumps over the lazy dog."

Evaluation:
Coherence and Logic: 5
Relevance and Accuracy: 4
Readability and Style: 5
Grammatical Correctness: 5
Overall Impression: 5
Justification: The machine-generated text maintains the core message and style of the ground truth, presenting it coherently and engagingly. Minor variations in wording do not impact the overall quality or relevance of the message. """

Figure 4: Prompt for evaluation (LLM as a Judge).

Your task is to generate a refined "future work" section for a scientific article.
Below, you will find sections from a scientific article including the 'Abstract', 'Introduction', 'Conclusion' of a scientific paper. The goal is to ensure that this section clearly outlines future research directions without the issues highlighted below.
Please focus on maintaining coherence, relevance, readability, grammatical correctness, and overall quality, ensuring the text flows logically and stays directly aligned with the main topic.

I am providing the texts and found these problems. At first read the feedback and try to improve them when you generte future work. [LLM Feedback]

Based on these details, please generate comprehensive and plausible future work suggestions that could extend the research findings, address limitations, and propose new avenues for exploration. future work should be within 100 words.

Figure 5: Incorporating LLM Feedback to the Model.

| Dataset | Coherence score | Silhouette score |
|---|---|---|
| ACL 12 | 0.63 | 0.66 |
| ACL 13 | 0.58 | 0.38 |
| ACL 14 | 0.61 | 0.46 |
| ACL 15 | 0.63 | 0.45 |
| ACL 16 | 0.55 | 0.47 |
| ACL 17 | 0.54 | 0.56 |
| ACL 18 | 0.53 | 0.49 |
| ACL 19 | 0.53 | 0.64 |
| ACL 20 | 0.52 | 0.73 |
| ACL 21 | 0.57 | 0.62 |
| ACL 22 | 0.60 | 0.53 |
| ACL 23 | 0.64 | 0.51 |
| ACL 24 | 0.59 | 0.38 |

Table 21: Performance of each year's future work.

| Model | Silhouette score |
|---|---|
| allenai-specter | 0.558131 |
| paraphrase-MiniLM-L6-v2 | 0.801949 |
| all-mpnet-base-v2 | 0.732606 |
| **all-MiniLM-L6-v2** | **0.819191** |
| all-MiniLM-L12-v2 | 0.479191 |
| paraphrase-multilingual-MiniLM-L12-v2 | 0.815246 |
| paraphrase-MiniLM-L6-v2 | 0.803524 |
| bert-base-nli-mean-tokens | 0.806526 |
| all-distilroberta-v1 | 0.488876 |
| msmarco-distilbert-dot-v5 | 0.436424 |
| multi-qa-MiniLM-L6-cos-v1 | 0.568171 |

Table 22: Performance of various Sentence Transformers

| **LLM extracted Ground Truth future work** | **LLM Generated future work.** | **BS (p)** | **BS (R)** | **BS (f1)** |
|---|---|---|---|---|
| future work should also include exploring the impact of different hyperparameters on model performance and investigating the potential benefits of incorporating external knowledge sources into the model. Furthermore, studying the generalizability of the model to different languages and domains would be an important direction for future research. Additionally, investigating the interpretability of the model's predictions and exploring ways to improve model explainability could also be valuable avenues for further study." | future work could focus on exploring novel architectures that combine the strengths of both encoder-only and decoder-only language models to improve semantic understanding in natural language processing tasks. Additionally, research could investigate alternative training methods or prompting techniques to enhance the capabilities of decoder-only models in comprehending word meaning. Furthermore, expanding the study to include languages other than English and evaluating a wider range of language models could provide a more comprehensive understanding of the performance differences between encoder-only and decoder-only architectures. Investigating the impact of model size and training data on semantic understanding could also be a valuable direction for future research. | 0.87 | 0.90 | 0.88 |

Table 23: LLM Extracted vs LLM Generated future work (Note: BS, p, R, f1 indicate BERTScore, Precision, Recall, and F1 score.