Poetry in RAGs: Modern Greek interwar poetry generation using RAG and contrastive training

Stergios Chatzikyriakidis
Computational Linguistics
and Language Technology Lab (UCRC),
Department of Philology,
University of Crete
stergios.chatzikyriakidis@uoc.gr

Anastasia Natsina
Literary Genres and
Literary History Lab (UCRC),
Department of Philology,
University of Crete
natsina@uoc.gr

Abstract

In this paper, we discuss Modern Greek poetry generation in the style of lesser known Greek poets of the interwar period. The paper proposes the use of Retrieval-Augmented Generation (RAG) to automatically generate poetry using Large Language Models (LLMs). A corpus of Greek interwar poetry is used and prompts exemplifying the poet's style with respect to a theme are created. These are then fed to an LLM. The results are compared to pure LLM generation and expert evaluators score poems across a number of parameters. Objective metrics such as Vocabulary Density, Average words per Sentence and Readability Index are also used to assess the performance of the models. RAG-assisted models show potential in enhancing poetry generation across a number of parameters. Base LLM models appear quite consistent across a number of categories, while the RAG model that is furthermore contrastive shows the worst performance of the three.

1 Introduction

The advent of Large Language Models (LLMs) has greatly increased the capabilities of NLP systems to deal with generation issues. Poetry generation has been one of them, with LLMs having the ability to generate poetry that is sometimes indistinguishable from human-made poetry by non-experts (Porter and Machery, 2024). To some extent, this is to be expected. Developing an aesthetical taste for poetry requires expertise, and similarly to other art forms, like music, non-experts can find it hard to distinguish AI vs. human-made poetry. However, despite their achievements and quick pace of improvement, LLMs do not perform as well in languages and/or styles that are not well represented in terms of freely (and even non-freely) available data. Highly stylized poetry in a lower resourced language, like interwar poetry in Modern Greek, can provide a powerful case study. Such cases require a targeted use of limited resources to enhance the performance of LLMs. One method is Retrieval-Augmented Generation (RAG), while another is based on contrastive learning. RAG has been shown to provide very positive results in enhancing LLM performance across a number of NLP tasks like Information Extraction (Wang et al., 2021; Ren et al., 2023), Machine Translation (Wang et al., 2022; Zhong et al., 2022), Question Answering (Guu et al., 2020; Shi et al., 2024) and Dialogue Systems (King and Flanigan, 2023; Fan et al., 2021), among many other tasks.See (Wu et al., 2024) for a full survey on RAG methods in NLP. The idea in contrastive learning is to provide both positive (poems in the target style) and negative examples (similar content but different style), in order to help the model better understand and maintain the distinctive stylistic features of a particular poet or poetic school. Recent work in style representation learning (Wegmann et al., 2022) has shown that contrastive methods are able to disentangle content from style; the generation of a highly specific poetic style, such as interwar Greek poetry with its slight authorial variations, will provide a litmus test.

In this paper, we focus on Modern Greek poetry of the interwar years, and implement a system to compare the results between RAG and contrastive learning in generating poems of the distinctive style. We use a dual retrieval system that is able to not only find poems with similar themes by the target poet but also retrieve examples from other poets that are contrastive.

The results show that RAG-assisted models show potential in improving poetry generation across a number of parameters. The base LLM models are quite consistent across a number of categories, while the contrastive RAG model shows the worst performance of the three.[1]

2 Related Work

The issue of poetry generation is not new to NLP. It has a history that includes a variety of approaches to generate poetry: hand-crafted symbolic rules (Oliveira, 2012), using statistical rules based on statistical machine translation (Jiang and Zhou, 2008), vanilla neural network approaches (Wöckener et al., 2021; Lau et al., 2018), and transformer architectures. LLM architectures have shown impressive performance in a variety of tasks, poetry generation notwithstanding. Attempts to use these architectures for poetry generation include approaches that finetune GPT-2 for poetry generation (Zhang and Eger, 2024a), zero-shot approaches (Tian and Peng, 2022), fine-tuning of more advanced models like ByGPT5 (Belouadi and Eger, 2023). The main take away in all these approaches is that fine-tuning helps the models in the task of poetry generation and the absence of fine-tuning is detrimental to the models' performance on more specific tasks, e.g. generating poetry in a specific style (Sawicki et al., 2023). Zhang and Eger (2024b) introduce a multi-agent framework for poetry generation, using LLMs. The research suggests incorporating non-cooperative dynamics in AI systems for enhancing creative diversity in a way similar to how human artists often deliberately differentiate their work from others. However, (Chen et al., 2024) report that current AI poetry still lacks in diversity, rhyming and semantic complexity, noting however, that style conditioning and character level modelling can help remedy these deficiencies to some extent.

3 The dataset

This paper uses an open-access dataset created by the second author with the help of a group of undergraduate students at the Philology Department, University of Crete. The slightly modified and richer corpus used here comprises over 600 poems in txt. format by a group of interwar Greek poets, namely Tellos Agras, Fotos Giofyllis, Romos Filyras, Kostas Karyotakis, Napoleon Lapathiotis, Kostas Ouranis, Mitsos Papanikolaou, and Maria Polydouri. With the notable exception of Kostas Karyotakis, the most prominent figure of this group who is recognized as a major Greek poet, the interwar poets are often referred to collectively, with an emphasis on their shared features. Melancholy, pessimism, and existential anxiety, stemming among other sources from the frustration of national expansionist aspirations and the dire sociopolitical reality of Greek interwar, as well as an added emphasis on nostalgia and a sotto voce quality, all of which are ascribed to neoromanticism and/or neo/post-symbolism (Filokyprou, 2009), are the most frequently repeated features of these lyrical poets (Beaton, 1994).

4 The models

The first model we use is based on Retrieval-Augmented Generation. The main idea is to use external resources to augment the performance of LLM models. In our case, the system takes a theme and the name of the poet as input, and then tries to search through a collection of poems (in our case, using our dataset of interwar poetry) in order to use them as examples to prompt LLMs. Search is performed using a multilingual model (paraphrase-multilingual-MiniLM-L12-v2). Each poem is converted into vector embeddings that are then stored in a FAISS vector store. FAISS is an effective library for effective similarity search and clustering. When a query is received, it is converted into the same vector space as the input poems. Similarity is computed using cosine distance, with the the model trying to match poems that are thematically similar to the query. The poems are then filtered according to the poet, trying to ensure that the retrieved examples match both the theme and the poet's style. After this filtering, the retrieved poems are used to construct the prompt for the generation model. A prompt example can be seen at the appendix. The pipeline is shown in 1:

The second model we use combines this basic RAG system with a contrastive approach. While maintaining the same embedding and similarity search infrastructure, the system now retrieves two distinct sets of examples: poems by the target poet that match the theme, and poems about the same theme written by different poets. This dual retrieval

[1] Github of the paper material can be found here: https://github.com/StergiosCha/RAG-poetry

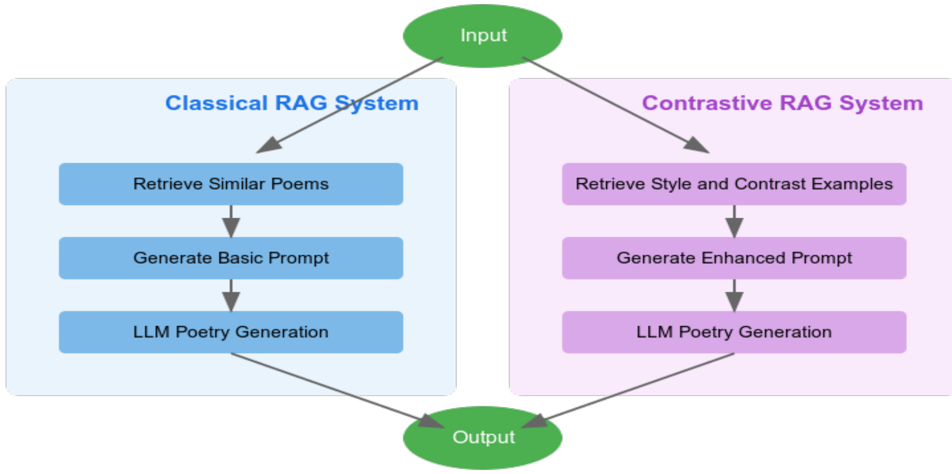

Figure 1: The Poetry Generation System Architecture. RAG retrieves similar poems from the target poet only, while Contrastive RAG additionally incorporates contrasting examples from other poets of the same school.

process uses the same multilingual embeddings and FAISS vector store, but applies different filtering criteria to create contrasting sets. The system first performs a broader similarity search to find thematically relevant poems, then splits these into positive examples (by the target poet) and contrastive examples (by other poets). These two sets are then incorporated into an enhanced prompt structure that explicitly guides the LLM to follow the stylistic patterns of the target poet while avoiding the stylistic features present in the contrasting examples.

In total we had 8 poet/theme pairs using GPT4-turbo with two poems each for base, RAG-assisted, and RAG-assisted contrastive generation (total of 48 poems) and 7 poet/theme pairs for GPT4o (total of 42).

5 Results and Discussion

Two expert evaluators were used for the GPT4o generated poems and three expert evaluators for GPT4-turbo. The evaluators were only shown the resulting poems without the corresponding prompts, and were asked to assess the closeness of the generated poem to the style and versification of the target poet, as well as evaluate the poem's relevance to its proclaimed theme and the level of creativity shown. The results of inter-annotator agreement show moderate agreement when using Spearman correlation (approx. 0.4). The agreement becomes moderate to strong when taking into account the relativity of judgments using nornalized z-scores (0.6). The results are shown in:

The table below shows the results of evaluation on several poems pairing themes and poets across a number of parameters as this was done by experts on Modern Greek poetry and tested on GPT-4-turbo and GPT-4o:

As we can see in figures 5 and 3 the RAG-model scores the highest for style and theme when using GPT4o and ties with base LLM in terms of theme in the GPT4-turbo case. Overall, the RAG system is marginally better w.r.t style and theme but the base model fares better w.r.t versification and creativity compared to the base models. The RAG plus contrastive model has the worst overall scores. This does not mean that the contrastive approach is not useful, but, probably, that the contrastive examples given to the system were not effective, because they were not distinctive enough, given that they were by poets of the same poetic school. The theme superiority of RAG is to be expected given that the retrieved poems are retrieved according to thematic fit. Versification is lacking in all approaches, however the base LLM outerperforms the enhanced approaches across the versification category.

Besides the expert evaluators, we also reverted to some metrics to assess the performance of the models vs. the original corpus, such as Vocabulary density (VD), Average words per sentence (AWpS)

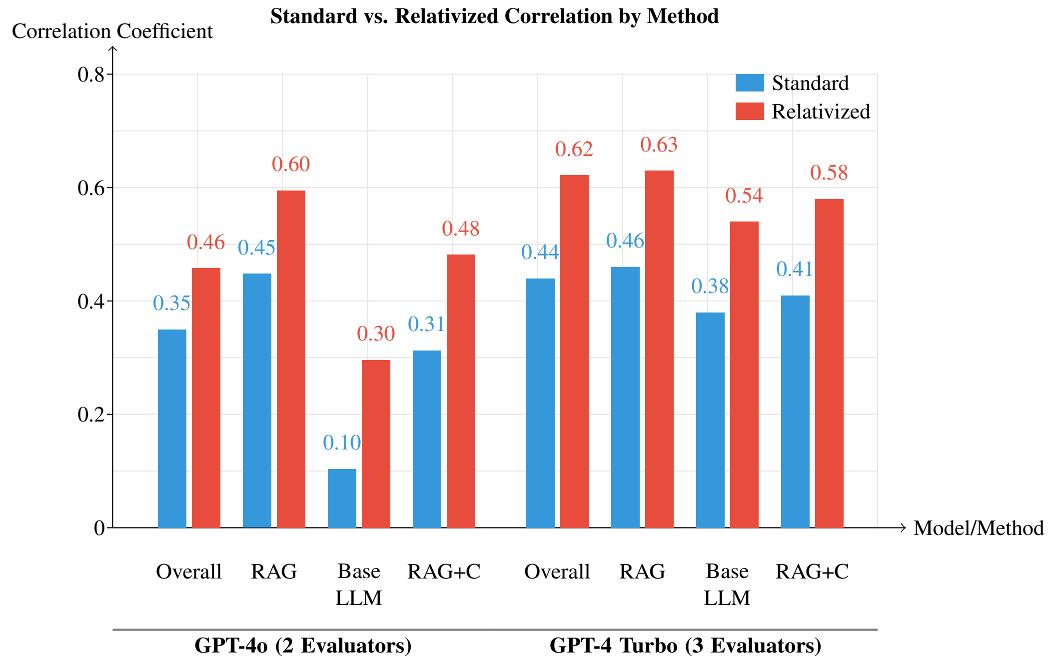

Figure 2: Comparison of standard (Spearman) correlation coefficients and relativized (Z-score normalized) correlation coefficients across different models and methods. Blue bars represent standard correlation values, while red bars show relativized correlation values after Z-score normalization to account for different scale usage patterns between evaluators. The left side shows results for GPT-4o with two evaluators, and the right side shows results for GPT-4 Turbo with three evaluators.

	CON			RAG			Base LLM			Original		
Poet	**Vocab**	**Avg**	**Read.**	**Vocab**	**Avg**	**Read.**	**Vocab**	**Avg**	**Read.**	**Vocab**	**Avg**	**Read.**
Papanikolaou	**0.558**	**26.1**	7.830	0.492	22.2	8.129	0.516	24.7	6.333	0.353	20.6	**9.229**
Δ *from orig.*	*+0.205*	*+5.5*	*-1.399*	*+0.139*	*+1.6*	*-1.100*	*+0.163*	*+4.1*	*-2.896*	–	–	–
Agras	**0.558**	20.7	8.114	0.537	19.3	**9.551**	0.545	**28.7**	8.302	0.199	18.8	9.229
Δ *from orig.*	*+0.359*	*+1.9*	*-1.115*	*+0.338*	*+0.5*	*+0.322*	*+0.346*	*+9.9*	*-0.927*	–	–	–
Lapathiotis	0.498	23.7	7.121	**0.527**	**28.1**	8.984	0.474	27.1	7.008	0.323	21.3	**9.033**
Δ *from orig.*	*+0.175*	*+2.4*	*-1.912*	*+0.204*	*+6.8*	*-0.049*	*+0.151*	*+5.8*	*-2.025*	–	–	–
Ouranis	0.474	24.0	8.691	**0.515**	32.3	10.246	0.495	27.6	8.819	0.273	**34.2**	**10.872**
Δ *from orig.*	*+0.201*	*-10.2*	*-2.181*	*+0.242*	*-1.9*	*-0.626*	*+0.222*	*-6.6*	*-2.053*	–	–	–
Karyotakis	0.422	**22.2**	9.412	**0.447**	18.1	**10.416**	0.437	19.5	7.663	0.322	14.2	10.249
Δ *from orig.*	*+0.100*	*+8.0*	*-0.837*	*+0.125*	*+3.9*	*+0.167*	*+0.115*	*+5.3*	*-2.586*	–	–	–
Polydouri	0.395	19.5	6.994	**0.414**	20.0	8.351	**0.397**	**22.8**	6.211	0.255	16.5	**8.631**
Δ *from orig.*	*+0.140*	*+3.0*	*-1.637*	*+0.159*	*+3.5*	*-0.280*	*+0.142*	*+6.3*	*-2.420*	–	–	–

Table 1: Combined metrics for common poets across all approaches, with deviations (Δ) from original poems shown below each row. For each metric, the highest score among CON, RAG, Base LLM, and Original is shown in **bold**.

and Readability Index (RI). We used Voyant Tools for this purpose. The results are shown in table1 The table does not give us a very clear picture, but some things do stand out: a) There is a clear tendency in all models to increase VD as well as AWpS. This is probably due to their base training in far more analytical discourse than in the elliptical poetic discourse exhibited in interwar Greek poetry; b) this is also related to the original poems exhibiting generally a greater RI, as Voyant Tools use the Coleman-Liau formula, based on number of letters, words and sentences; c) RAG, which seems to perform better according to the evaluators, has generally the most increased VD, but it stays closer to the original AWpS, a feature

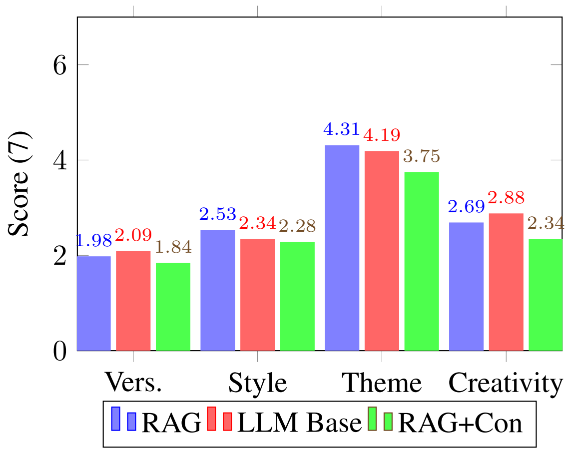

Figure 3: Comparison RAG, LLM Base, RAG+Con for the two annotators using GPT4o

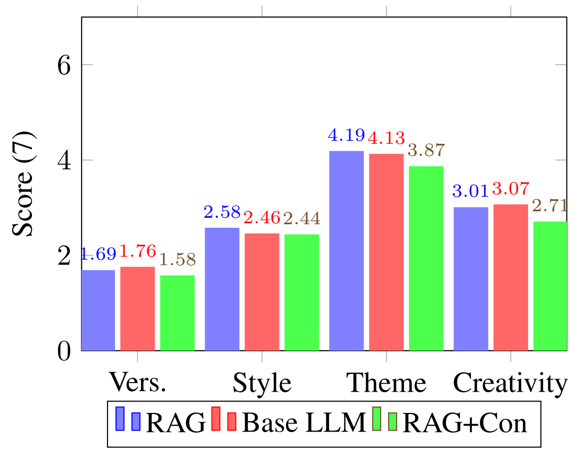

Figure 5: Average scores across all figures for all evaluators (RAG, Base LLM, RAG+Con).

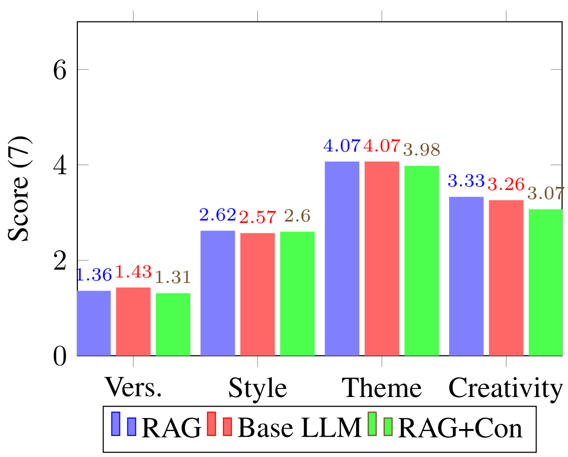

Figure 4: Comparison of approaches (RAG, Base, RAG+Con) for three-annotators using GPT-4-turbo.

that would be more readily recognized as distinctive of style, hence gaining more credibility with the evaluators.

6 Conclusions and future work

The paper has produced mixed results, showing some potential for the use of RAG in poetry generation, particularly as regards a recognizable style. RAG also seems slightly better at developing a theme consistently throughout a poem, as well as maintaining a style closer to the target poet, while it also has an edge in creativity in one of the two models. Still, base LLMs are quite consistent across a number of categories. The consistency in versification and the fact that they score higher in this dimension might have to do with their ability to maintain an internal rhythm and also being more successful at rhyming than the assisted models. This does not necessarily have to do with an understanding of the poetic style asked to generate. Most probably, this is the result of being trained on simple poems and/or song lyrics that have a sense of rhythm and rhyming. This is an interesting avenue to explore, using RAG models that are also improvements over this dimension. This might need a more nuanced approach where the retrieved poems are retrieved across a number of dimensions and not only thematic fit. For example, poems in this style rely heavily on rhyme and, as such, improving a model on this dimension needs a RAG system that is not only sensitive to meaning similarity but also to rhyme-sensitive meaning similarity. This is definitely one avenue that needs to be further explored. As far as contrastive training is concerned, future work might include working first with starkly contrastive poetic styles (eg. modernist, or surrealist) and then move on to train the model to the more nuanced differences within a poetic school.

Limitations

We acknowledge three main limitations to this work. The first one concerns exploring more variations of RAG and Contrastive RAG models to have a clearer picture of their effectiveness. The second one is about the effectiveness of these approaches as we move to other poetic styles and/or in other languages. The last one regards the limited pool of expert evaluators (experts in interwar Modern Greek poetry), should one wish to duplicate the results and broaden the research.

Ethics Statement

There are no considerable ethics considerations related to the work presented in this paper.

Acknowledgements

The authors gratefully acknowledge help from Dimitris Polychronakis and Elli Filokyprou for providing expert assessments for the generated poems presented in this paper. The first author is partially funded by the Special Account for Research Funding of the University of Crete (grant number: 11218).

References

Roderick Beaton. 1994. *An Introduction to Modern Greek Literature: Revised and Expanded*. Oxford University Press.

Jonas Belouadi and Steffen Eger. 2023. Bygpt5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating diversity in automatic poetry generation. *arXiv preprint arXiv:2406.15267*.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

E Filokyprou. 2009. I genia tou karyotaki fevgontas ti mastiga tou logou. *Athens: Nefeli*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Long Jiang and Ming Zhou. 2008. Generating chinese couplets using a statistical mt approach. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 377–384.

Brendan King and Jeffrey Flanigan. 2023. Diverse retrieval-augmented in-context learning for dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5570–5585, Toronto, Canada. Association for Computational Linguistics.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. *arXiv preprint arXiv:1807.03491*.

Hugo Gonçalo Oliveira. 2012. Poetryme: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence*, 1:21.

Brian Porter and Edouard Machery. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–306.

Piotr Sawicki, Marek Grzes, Fabricio Goes, Dan Brown, Max Peeperkorn, and Aisha Khatun. 2023. Bits of grass: Does gpt already know how to write like whitman? *arXiv preprint arXiv:2305.11064*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *NAACL-HLT*.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3587–3597, Seattle, United States. Association for Computational Linguistics.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of* *the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Jörg Wöckener, Thomas Haider, Tristan Miller, The-Khang Nguyen, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi, and Steffen Eger. 2021. End-to-end style-conditioned poetry generation: What does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 57–66, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

Ran Zhang and Steffen Eger. 2024a. Llm-based multi-agent poetry generation in non-cooperative environments. *arXiv preprint arXiv:2409.03659*.

Ran Zhang and Steffen Eger. 2024b. Llm-based multi-agent poetry generation in non-cooperative environments. *arXiv preprint arXiv:2409.03659*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673.

A RAG Prompt Example for poet Polydouri and the Theme *love* with k = 6

```
Δημιούργησε ένα νέο ελληνικό ποίημα στο ύφος
της Μαρίας Πολυδούρη.
Θέμα: αγάπη
Παραδείγματα παρόμοιων ποιημάτων για έμπνευση:

Ποίημα 1:
της ομορφιάς το πέρασμα,
τη νειότη που μ' αφήνει.
Έλα γλυκέ
Έλα γλυκέ, κι' αν φτάνη η νύχτα και το σκοτάδι
δε σ' αρέση, αστέρινο θαμπό στεφάνι
η αγάπη μου θα σου φορέση.
Στο ταραγμένο μέτωπό σου
αργά τα δάχτυλα θα σύρω
κι' ό,τι είνε πάθος στην καρδιά σου θ' ανθίση
δάκρια και μύρο.
---
Ποίημα 2:
μονάχα για τη διαλεχτήν αγάπη σου.
Μονάχα γιατί τόσο ωραία μ' αγάπησες
έζησα, να πληθαίνω
τα ονείρατά σου, ωραίε που βασίλεψες κ
έτσι γλυκά πεθαίνω
μονάχα γιατί τόσο ωραία μ αγάπησες.
Σεμνότης
Την ομορφιά που κλείνω μέσα μου
κανείς δεν θέλω να τη νοιώση.
Δε θα μπορούσε να τη σίμωνε
---
Ποίημα 3:
νάναι μονάχη του «χαίρε» η χορδή
στην καρδιά μου!
Πάνε τα ωραία, τ' αγνά, η ζωή.
Αδιαφορία στης αγάπης τα μάτια.
Κακίας μεθύσι στο χαλασμό
του ό,τι απομένει,
στο μαρασμό που έχει ανθίσει
μέσα μου κ' εξω - κισσού πλημμύρα,
σημαία αποκλεισμού!
Πάνε τα ωραία, τ' αγνά, η ζωή.
---
Ποίημα 4:
Αχ, με πονεί η καρδιά μου
Αχ, με πονεί η καρδιά μου. Ούτε η ματιά σου,
Φύση, που μου ήσουν μια παρηγοριά.
Μάταια το Δάσος μ' όλα τα κλαριά
νεύει και μου φωνάζει η ομορφιά σου.
Ούτε η ματιά σου, Αγάπη λυπημένη,
Αγάπη σιωπηλή, δε με πλανά.
Η σκέψη μου όχι πως σε λησμονά,
---
Ποίημα 5:
Μέσ' στην καρδιά μου
Μέσ' στην καρδιά μου τη βουβή,
καιρό πια ρημασμένη, επέρασεν η
αγάπη σου σαν άνοιξης
πνοούλα.
Και το αηδονάκι του καημού στάθη στην ανθισμένη
χαρά μου και τραγούδησε - λαχτάρα και τρεμούλα.
Γιατί θυμάσαι το βουβό, το ρημασμένο κάστρο
---
Ποίημα 6:
καμάρωσες στα χείλη μου απλωμένο
κ' έχεις μεσ' στων ματιών μου το ξαστέρωμα
τον πόθο σου τρελλά καθρεφτισμένο.
Με γνώρισες να γέρνω στην αγάπη σου
```

σαν πεταλούδα στο άλικο λουλούδι
και να σκορπίζω όσο η καρδιά μου εδύνοταν
μεθυστικό το ερωτικό τραγούδι.

Δημιούργησε ένα νέο πρωτότυπο ποίημα που να:
1. Διατηρεί το ύφος και την τεχνοτροπία των παραδειγμάτων
2. Χρησιμοποιεί παρόμοια δομή στίχων
3. Αξιοποιεί πλούσιες ποιητικές εικόνες
4. Είναι μοναδικό στην έκφραση

Το ποίημα:
RAG-Generation