DOI: 10.26314/GARR-Conf25-proceedings-07

ARGOS: A Retrieval-augmented GeneratiOn approach for Scientific communication

Daniele Di Bella[1,2], Pietro Roversi[1,2]

[1]Consiglio Nazionale delle Ricerche–Istituto di Biologia e Biotecnologie Agrarie (CNR–IBBA), [2]Fondazione Telethon

Abstract. Effectively communicating biological research to non-specialist audiences remains a critical challenge. Within the Broad-Spectrum Rescue-of-Secretion project, we want to explore the potential of Retrieval-Augmented Generation (RAG) in life sciences communication. We hence developed ARGOS, a Python-based pipeline leveraging OpenAI's GPT-4.1 combined with bibliographic retrieval from Zotero libraries, to generate Wikipedia-style summaries tailored for diverse audiences. Expert evaluations of ARGOS-generated texts in English and Italian showed high scores for correctness and readability, though completeness was somewhat limited by dataset scope and prompt design. Overall, ARGOS proved to be a good instrument to conduct further studies

Keywords. large language models, retrieval-augmented generation, scientific communication, public outreach, rare diseases

Introduction

Effectively communicating biomedical research to non-specialist audiences—particularly patients affected by rare diseases—remains a critical yet insufficiently addressed challenge in contemporary science communication. One such effort was undertaken within the framework of the two-year project Broad-Spectrum Rescue-of-Secretion of Tdark Glycoprotein Mutants, funded by the Telethon Foundation1. The project investigates whether modulating the endoplasmic reticulum (ER) quality control enzyme UGGT can promote the secretion of misfolded yet functional ("responsive") glycoprotein mutants. These mutants are implicated in rare congenital diseases, and the approach aims to evaluate UGGT inhibition or deletion as a broad-spectrum therapeutic strategy, with particular focus on poorly characterized Tdark glycoproteins. The project also considers potential cellular risks associated with targeting such a central checkpoint in the ER quality control system. To support public engagement with the research, our team initiated a science communication effort focused on improving access to information about the ten Tdark glycoproteins studied in the project. Specifically, we are undertaking the creation of Wikipedia entries for each protein to provide accurate, accessible, and broadly disseminated summaries of existing knowledge for affected individuals and the wider public.

This initiative also presented an opportunity to investigate the potential of recent advances in natural language processing to support the dissemination of life sciences research. In fact, Large language models (LLMs), which are capable of generating fluent, gramma-

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

tically correct text in multiple languages, show promise in this domain, even though they are known to generate "hallucinations"—plausible but incorrect statements. Retrieval-Augmented Generation (RAG) addresses this limitation by combining LLMs with information retrieval systems that guide generation using external, user-specified sources, thus improving factual reliability (Lewis et al., 2021).

Wanting to assess whether RAG-based systems can meaningfully enhance life sciences communication, we developed ARGOS (A Retrieval-augmented GeneratiOn approach for Scientific communication), which we present in this contribution. ARGOS generates Wikipedia-style summaries using bibliographic sources retrieved from a user's Zotero library, and it is not an end in itself. It serves as an experimental tool to explore the broader question of how RAG architectures might improve the experience of public facing biomedical content.

In the following sections, we present the reasons that led us to think RAG tools can be beneficial for science communication, their inherent contradictions, ARGOS' workflow, and an initial assessment of its performance.

1. Intentions and contradictions

1.1 Intentions

Despite the rapid pace of discovery in fields such as biology, researchers often lack both the time and incentives to engage in public outreach (Nerlich 2017). Moreover, those who attempt to do so frequently encounter peer pressure and professional disincentives, discouraging sustained communication efforts (Rose et al. 2020). These systemic barriers contribute to a widening gap between the scientific community and the broader public.

Language further compounds this divide. English remains the dominant language of science communication, restricting access for non-English-speaking populations and reinforcing a singular cultural perspective in the interpretation and dissemination of knowledge (Márquez and Porras 2020). Consequently, scientific knowledge often remains both linguistically and culturally inaccessible to large segments of the global population and, together with the aforementioned issues and the intrinsic difficulty of scientific matters, exacerbates that separation between the scientific community and the general public that often leads them to perceive each other not as entities in continuity, but as different entities.

Providing tools that enable scientists to share their work more easily, accurately, and inclusively could foster the ongoing process of shaping the scientific community's social vocation, that sees it in dialogue with the rest of the global community, of which it is an integral part. RAG applications offer a promising solution for that. Such systems can assist researchers in creating accurate, accessible summaries of their work across multiple languages and cultural contexts.

1.2 Contradictions

Although the use of RAG has been proposed as a promising strategy for democratizing access to scientific knowledge, this approach presents inherent tensions, particularly con-

cerning linguistic equity and cultural representation. For instance, ARGOS relies on OpenAI's GPT-4.1 model, yet, the training data for GPT models are unevenly distributed in favour of English-language sources, limiting their performance in generating content in other languages and potentially introducing cultural bias into the output.
At present, ARGOS uses two standardized English-language prompts to generate both English and non-English outputs. The first prompt instructs the model on the desired output language and audience, and requests it to tailor the output text to the relevant cultural context, while the second orders the model to proofread what it has produced. This method highlights a paradox: a tool with linguistic and cultural biases is being used to address the very inequities it may perpetuate.
Unfortunately, due to the computational and financial costs associated with training LLMs from scratch, most developers—including our team—must rely on pre-trained models. Consequently, we are constrained by the design choices and implicit biases embedded by those who created such models.
Moreover, while AI technologies may appear low-cost, their affordability is relative, often excluding users in low- and middle-income countries (LMICs). This raises a concern: can reducing disparities be attempted through the use of technologies that are themselves products of those disparities? Encouragingly, a growing number of initiatives in LMICs are developing language models tailored to local non-European languages, social contexts, and computational constraints2. These efforts open new possibilities for the use of RAG systems in scientific communication.

2. ARGOS workflow and first validations

Currently ARGOS is built around a Python pipeline that can be called through a command line interface, and that follows the workflow depicted in Fig. 1.

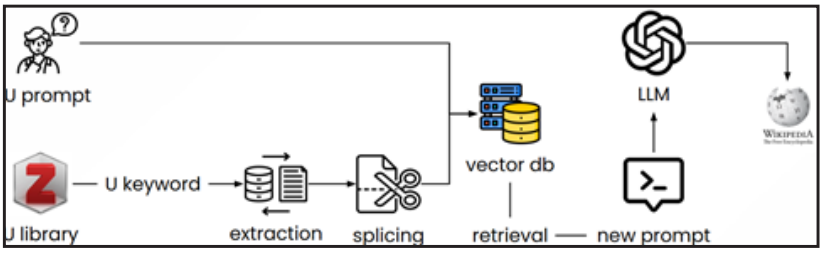

Fig. 1
ARGOS
workflow

Firstly, the user provides a prompt (U prompt) and some keywords (U keyword). The keywords are used to browse the user's Zotero library in search of items that are extracted and spliced in chunks. Each chunk is vectorized through OpenAI's text-embedding-3-large model and stored in a vector database. U prompt is vectorized as well and used to launch a similarity search that selects the chunks more likely to contain information related to the user's request. These chunks are then used to create a new prompt which is submitted to the LLM in charge to generate the output text (in our case, GPT-4.1). Before providing this output to the user, the same LLM is asked to correct any error according to the grammar and syntax of the user's desired language, and to adapt the tone and the style of the parts

of the text that aren't meeting the communication needs of the user's selected audience.
To validate the application, we decided to generate 10 texts, 5 in English and 5 in Italian, about 5 of the 10 glycoproteins of our interest, and submit them to three colleagues, experts in these three proteins. For each text, we asked the colleagues to rank its scientific accuracy (correctness), the presence of relevant information about the protein described (completeness) and their reading experience (readability) on a scale ranging from 1 to 10. Results are presented in Fig. 2.

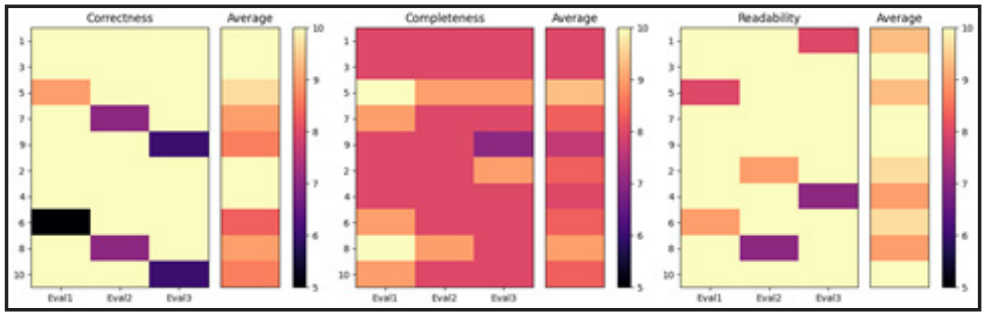

Fig. 2
First validation of ARGOS

The numbers labelling the rows of the matrices are the texts' identification codes: even numbers indicate Italian texts, while odd numbers are for English texts. In all three cases, 1 and 10 indicate very bad and very good ratings, respectively.
As can be seen from Fig.2, the correctness and readability of the texts are generally highly ranked. Both parameters present very small differences that identify Italian texts as slightly less correct and readable, which, for the readability, may be explained by the worse performances of GPT models in non-English languages. However, given the small size of the sample of experts, such differences are likely non-meaningful fluctuations.
On the other hand, while completeness as well is highly scored, it is visibly worse, and this could be due to our inexperience on the proteins ARGOS wrote about. In fact, during the creation of the Zotero datasets, we may have omitted some papers that the expert colleagues considered of primary importance. Moreover, our inexperience may have led us to formulate questions differently than the experts would have done. In fact, as underlined in the interesting contribution of Wong and colleagues (Wong et al. 2025), a key feature of RAG systems is the dependence on the user's prompt. This element leads such systems to select particular information from nonparametric memory (i.e., from the sources to which they are given access) and it is dependent on the user's starting beliefs, which is why Wong and colleagues warn against this feature of RAG systems. We probably approached ARGOS believing that the 5 selected proteins could be described by some features, while experts in those molecules would choose others, and this led the system to respond in a way that was assessed as not complete.
Overall, ARGOS produced texts that passed the experts' evaluation positively and even allowed us to find answers about a protein of our interest that had been sought for some time. We consider it sufficiently good to be used in the next steps of our project, which may include experiments with other RAG systems and broad groups of annotators.

References

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv, (arXiv:2005.11401). https://doi.org/10.48550/arXiv.2005.11401

Márquez, M. C., & Porras, A. M. (2020), Science Communication in Multiple Languages Is Critical to Its Effectiveness. Frontiers in Communication, (5). https://doi.org/10.3389/fcomm.2020.00031

Nerlich, B. (2017), Time and science communication. Making Science Public. https://blogs.nottingham.ac.uk/makingsciencepublic/2017/04/07/time-science-communication/

Rose, K. M., Markowitz, E. M., & Brossard, D. (2020), Scientists' incentives and attitudes toward public communication, Proceedings of the National Academy of Sciences, 117(3), pp 1274–1276. https://doi.org/10.1073/pnas.1916740117

Wong, L., Ali, A., Xiong, R., Shen, S. Z., Kim, Y., & Agrawal, M. (2025), Retrieval-augmented systems can be dangerous medical communicators, arXiv. https://doi.org/10.48550/ARXIV.2502.14898

Links

1https://www.fondazionetelethon.it/en/what-we-do/research/projects-funded/broad-spectrum-rescue-of-secretion-of-tdark-glycoprotein-mutants/
2https://www.nature.com/articles/d41586-025-01546-6

Autori

Daniele Di Bella daniele.dibella@ibba.cnr.it
Daniele Di Bella is a computational biologist at the IBBA-CNR Institute in Milan. From March 2023 to March 2024 he worked on his thesis at the Alfred Wegener Institute of Bremerhaven, Germany. After his graduation at the University of Milan, in 2024, he started working as a research fellow at IBBA-CNR, where he focuses on AI and bioinformatics.

Pietro Roversi pietro.roversi@cnr.it
Pietro Roversi is a structural biologist at the IBBA-CNR Institute in Milan. From 1996 to 2021, he worked in the UK at Cambridge, Oxford and Leicester. Since 2012 he leads a research project focusing on the potential of secretion-rescue strategies for the therapy of congenital rare disease due to a responsive missense mutation in a secreted glycoprotein gene.