REVERINO: REgesta generation VERsus latIN summarizatiOn

Giovanni Puccetti[1,*], Laura Righi[2], Ilaria Sabbatini[3] and Andrea Esuli[1]

[1]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", via G. Moruzzi 1, 56124, Pisa PI, Italy

[2]Università degli Studi di Modena e Reggio Emilia - Dipartimento di Educazione e Scienze Umane, viale Timavo, 93 – 42121, Reggio Emilia RE, Italy

[3]Università degli Studi di Palermo - Dipartimento Culture e Società, via delle Scienze, 15 – 90128, Parlermo Pa, Italy

Abstract

In this work we introduce the REVERINO dataset, a collection of 4533 pairs of Latin *regesta* with their respective full text medieval pontifical document extracted from two collections, *Epistolae saeculi XIII e regestis pontificum Romanorum selectae. (1216-1268)* and *Les Registres de Gregoire IX (1227/41).* We describe the pipeline used to extract the text from the images of the printed pages and we make high level analysis of the corpus.

After developing REVERINO we use it as a benchmark to test the ability of Large Language Models (LLMs) to generate the *regestum* of a given Latin text. We test 3 LLMs among the best performing ones, GPT-4o, Llama 3.1 70b and Llama 3.1 405b and find that GPT-4o is the best at generating text in Latin. Interestingly, we also find that for Llama models it can be beneficial to first generate a text in English and then translate it in Latin to write better *regesta.*

Keywords

Regesta, Latin Text Summarization, Large Language Models, Digital Humanities

1. Introduction

ITSERR [1] (Italian Strengthening of the ESFRI RI RESILIENCE) is a interdisciplinary and distributed Research Infrastructure for Religious Studies. In the context of this project, REVERINO is a novel dataset of *regesta* with the medieval Latin texts they summarize and their apparatus. A dataset designed to recreate the methodology of *regesta* generation, and specifically designed for the creation of an Artificial Intelligence-based tool for summarizing medieval documents, with a particular focus on pontifical documents. The decision to employ the system of *regesta* for organizing, indexing, and summarizing medieval texts through generative AI stems from the integration of various scholarly needs, which we explore and test in depth. To create a new automated organizational process tailored to historical documents, we have chosen to focus on automatic summarization, drawing on an established and scientifically validated methodology – namely the creation of *regesta*, a practice improved by humanists and scholars since the 19th century. Scholars studying medieval charters often need to explore specific topics, historical figures, or places within vast corpora of sources, sometimes employing a comparative or *longue durée* approach. These corpora remain widely dispersed and are preserved across various libraries and archives that are geographically distant and differently organized, making them difficult to access. This is particularly true for the extensive documentation produced by royal and papal chanceries. Starting from this observation, we decided to work on the creation of a specifically designed dataset for

IRCDL 2025: 21st Conference on Information and Research science Connecting to Digital and Library science, February 20-21 2025, Udine, Italy

[*] Corresponding author.

[†] Contributions: Puccetti and Esuli set up eScriptorium, wrote the code for automatic *regesta* generation and wrote Section 3.2. Righi conducted the qualitative analysis of the results and wrote Section 1 and Section 3.3. Sabbatini took care of the full data pipeline and wrote Section 2. All the authors wrote abstract, conclusions, and revised the final version of the paper.

✉ giovanni.puccetti@isti.cnr.it (G. Puccetti); laura.righi@unimore.it (L. Righi); sabbatini@fscire.it (I. Sabbatini); andrea.esuli@isti.cnr.it (A. Esuli)

🌐 https://gpucce.github.io (G. Puccetti); https://esuli.it/ (A. Esuli)

iD 0000-0003-1866-5951 (G. Puccetti); 0000-0002-5770-0978 (L. Righi); 0009-0001-2936-6562 (I. Sabbatini); 0000-0002-5725-4322 (A. Esuli)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

the development of a tool for the summarization of documents produced by medieval pontiffs (c. 1200 to 1350).

1.1. Regesta

A *regestum*, the Latin word for *list, enumeration, specification*[1], is a summary of a document made for the use of stakeholders and scholars, making the document content readily available without the need to consult it in its entirety. Each *regestum* contains some essential information, namely, 1) the name of the author (i.e. the Pope); 2) the name of the recipient; 3) an abstract of the content (with the object and the operative verb); 4) the date (calculated from the year of pontificate) and 5) the place of production of the document. A *regestum* always has a reference full text document, of which it is the "summary" and both come together with an *apparatus*, a formal text indicating the collection and the manuscript where the *regestum* is found.

While the three components, *regestum*, full text and *apparatus*, are conceptually close, collecting them together can be challenging. Indeed, there are three main issues when trying to retrieve both a *regestum* and its full text from a collection or creating a new one: a) the *regestum* and the corresponding document are often not collected in the same volume (as in the Potthast collection [2]); b) these modern printed collections are not easily accessible and readable; c) the publication of new editions or the update of existing *regesta* collections is extremely time consuming. For these reasons, many *regesta* collections have been created in the past, especially of medieval documents produced by royal and papal chanceries, but few of these have been updated or created from scratch in recent years.

1.2. Text summarization

In the Natural Language Processing (NLP) literature, automatic text summarization is the task of rewriting the content of a text passage into a shorter form while retaining the relevant information without involving a writer in the process. *Regesta* fit well in this framework since they are summaries of longer texts meant to expose specific information in an easy to consult form. Therefore, **REVERINO** is well suited to be both **an easy to inspect dataset of Latin *regesta***, as well as a training and testing **benchmark for Latin text summarization**.

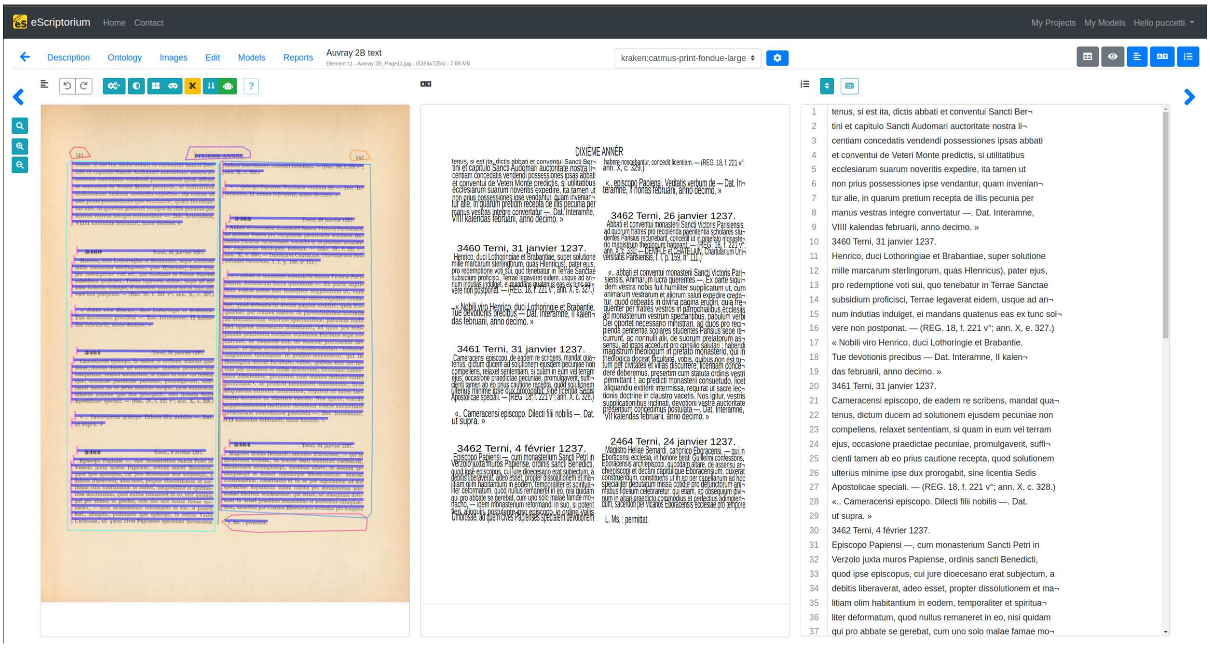
2. The REVERINO Dataset

There is an extensive number of printed collections of *regesta* in Latin, edited from several scholars during the 19th and 20th century, however only a few examples that are digitally available, generally as sets of high quality digital images, and only very rarely with a full text, machine readable version (and often performed with old or bad quality OCRs). One of the most relevant fields in which *regesta* have been produced is the corpus of the pontifical acts and letters issued by the popes and the papal chancery during the Middle Ages. This corpus guarantees the presence of large collections of *regesta* and extended texts that were already edited and published in printed versions during the 19th and 20th centuries, such as Jaffé and Potthast's *Regesta Pontificum Romanorum*; the collection published by the Bibliothèque des Écoles françaises d'Athènes et de Rome (BEF); the editorial series of the *Monumenta Germaniae Historica* (MGH).

2.1. Data Selection

Regesta collections are only available as images of full pages, this poses a first obstacle to the creation of a large scale corpus of these documents. Each manuscript has different pagination, layout, writing font, image quality, format, etc., and thus requires a custom pipeline for the extraction of the text into a machine readable format. Nevertheless, manuscripts from one collection undergo a similar digitization

[1] https://glosbe.com/la/en/regesta



tenus, si est ita, dictis abbati et conventui Sancti Ber¬
tini et capitulo Sancti Audomari auctoritate nostra li¬
centiam concedatis vendendi possessiones ipsas abbati
et conventui de Veteri Monte predictis, si utilitatibus
ecclesiarum suarum noveritis expedire, ita tamen ut
non prius possessiones ipse vendantur, quam invenian¬
tur alie, in quarum pretium recepta de illis pecunia per
manus vestras integre convertatur —. Dat. Interamne,
VIIII kalendas februarii, anno decimo. »
3460 Terni, 31 janvier 1237.
Henrico, duci Lothoringiae et Brabantiae, super solutione
mille marcarum sterlingorum, quas Hlenricus), pater ejus,
pro redemptione voti sui, quo tenebatur in Terrae Sanctae
subsidium proficisci, Terrae legaverat eidem, usque ad an¬
num indutias indulget, ei mandans quatenus eas ex tunc sol¬
vere non postponat. — (REG. 18, f. 221 v°; ann. X, e. 327.)
« Nobili viro Henrico, duci Lothoringie et Brabantie.
Tue devotionis precibus — Dat. Interamne, II kalen¬
das februarii, anno decimo. »
3461 Terni, 31 janvier 1237.
Cameracensi episcopo, de eadem re scribens, mandat qua¬
tenus, dictum ducem ad solutionem ejusdem pecuniae non
compellens, relaxet sententiam, si quam in eum vel terram
ejus, occasione praedictae pecuniae, promulgaverit, suffi¬
cienti tamen ab eo prius cautione recepta, quod solutionem
ulterius minime ipse dux prorogabit, sine licentia Sedis
Apostolicae speciali. — (REG. 18, f. 221 v°; ann. X. c. 328.)
«.. Cameracensi episcopo. Dilecti filii nobilis —. Dat.
ut supra. »
3462 Terni, 4 février 1237.
Episcopo Papiensi —, cum monasterium Sancti Petri in
Verzolo juxta muros Papiense, ordinis sancti Benedicti,
quod ipse episcopus, cui jure dioecesano erat subjectum, a
debitis liberaverat, adeo esset, propter dissolutionem et ma¬
litiam olim habitantium in eodem, temporaliter et spiritua¬
liter deformatum, quod nullus remaneret in eo, nisi quidam
qui pro abbate se gerebat, cum uno solo malae famae mo¬

Figure 1: Example of the escriptorium Interface.

procedure and therefore can be processed together through a single pipeline, to extract the content of all the documents.

Given the need for a custom approach for each corpus, we choose to limit our collection to two sets of printed collections, specifically, we identify 2 main sources:

1. **MGH**: *Epistolae saeculi XIII e regestis pontificum Romanorum selectae. (1216-1268)* [3]
2. **Auvray**: *Les Registres de Gregoire IX (1227/41)* [4]

While conceptually similar these two collections are formally different: MGH is written in a single column format while Auvray in two columns, the first has *apparati* visually separated from the original document, while for the second they are part of the *regestum*. The different collections have thus several smaller visual differences linked to the layout and the font used.

Finally, from a qualitative perspective they collect and summarize the documents related to two different medieval popes: Gregory IX and Honorius III. In particular, Auvray collects only the documents related to pope Gregory IX, and MGH collects the documents issued by both Gregory IX and Honorius III. These collections were chosen as starting corpus because they allow for the collection of different types of *regesta*. Indeed, although the creation of a *regestum* is based on specific rules shared in the research domain, different scholars inevitably produce different *regesta* from the same document. It is therefore important to consider *regesta* produced by different scholars in different periods.

2.2. Data Curation

The pipeline leading from a collection of images of printed pages to the REVERINO corpus is composed of 4 steps: *Annotation*, *Training*, *Extraction* and *Post-processing*.

Annotation We manually annotate a selected set of pages from each collection of *regesta* to use as a training dataset, this is done on a local instance of the eScriptorium platform [5], an example of the interface can be seen in Figure 1. Our pipeline involves both segmentation of the written parts of each image as well as OCR. Annotating data for the latter is too time demanding and existing models are effective enough, therefore we limit ourselves to annotating pages to train a segmentation model and rely on available OCR ones.

The models in eScriptorium ingest annotations with two kinds of information: 1. *areas* isolating the parts of a page that contain text, and 2. *lines* identifying the text of a line and its position in the page. Thus, each page is annotated in two steps, first the relevant areas are circled and then each line is colored.

Training To adapt models to the outline of a manuscript, we start from a working segmentation model provided by eScriptorium, *catmus print large* [6]. This model works sufficiently well on MGH and we can use it as is. Differently, the Auvray collection has a two columns format and we need to train the model on a dataset collected in the Annotation step. We go back and forth between Training and Annotation to fix the limitations of each trained model, reaching a total of 91 annotated pages. This process led to high quality results in segmenting the outline of the Auvray manuscript and extracting text.

Extraction Once the model has been trained, we use it to process all the pages in each collection, obtaining text lines that the model is able to identify along with their position on the page, and thus giving us a clean continuous stream of text spanning the full document.

Post-Processing The last step consists in using a series of heuristics based on the content and the position information in the page of each text line extracted, to separate each *regestum* from the longer text it summarizes and from the apparatus.

2.3. Data Statistics

From a quantitative perspective, MGH is composed of a total of 2283 *regesta* and full text pairs and Auvray of 3983. However, for Auvray, several of the full texts extracted were only short passages, often quotations or incipits that don't contain the information needed to generate a *regestum*, therefore we drop them. After this cleaning there are 2250 *regesta* left in Auvray.

While MGH and Auvray are similar, – they are both collections of *regesta* –, they show several differences: they collect documents written by different popes and they are edited by different scholars, and on top of this they are different as printed publications. Indeed, due to the layout of the two collections, as MGH is composed of single column pages while Auvray is composed of two columns pages. Also, the quality of the second dataset is generally lower, due to minor errors in OCR quality, few characters and numbers are wrongly transcribed by our custom model. Therefore we keep two separate splits of the dataset.

To provide a qualitative understanding of the difference, we use t-SNE [7], after encoding the *regesta* using LaBERTA a latin adaptation of BERT [8].

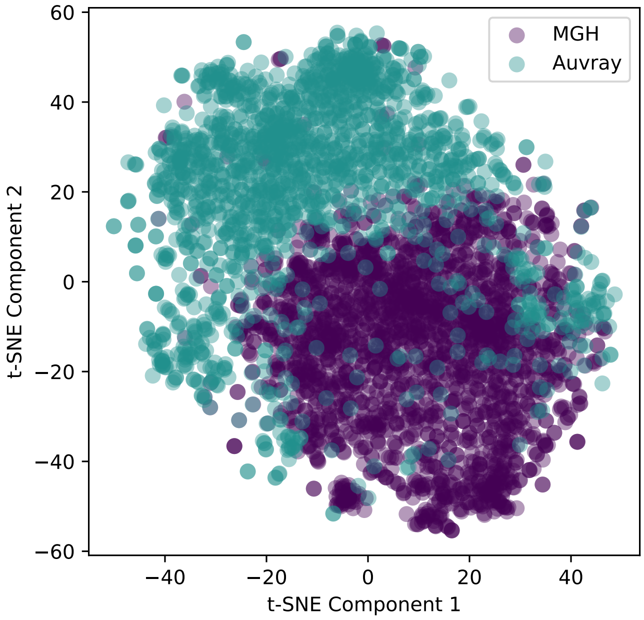


Figure 2: T-SNE plot showing samples from the two manuscripts MGH and Auvray.

Table 1
Prompts used to make the LLMs generate a regestum given a full text.

Format	*Backtranslate*
Given the following text in Latin please write in Latin a «regesto» for it, containing:	Given the following text in Latin please first translate it to English and then write in Latin a «regesto» for it, containing:
Shared Prompt	
1. The name of the author (i.e. the Pope); 2. The name of the recipient; 3. An abstract of the content (with the object and the operative verb); 3. The date (calculated from the year of pontificate); 4.The place. TEXT: ...	

Figure 2 shows the t-SNE plot of the two datasets, while they are well separated there is non-negligible overlapping between the two, hinting that in the future they can be used together to train a language model able to automatically write the *regestum* of a Latin document.

In the future, to prevent this "bipartite" distribution of the samples in REVERINO, we will add *regesta* from different manuscripts to contribute a more broadly distributed dataset.

3. Text Summarization in Latin

Text summarization is a long standing task in Natural Language processing [9], which in the past was tackled through dedicated approaches often based on the retrieval of similar passages. Since LLMs have shown the ability to generate free-form text [10] they are currently the best performing systems for summarizing texts. An example of a widely used benchmark is the XSUM dataset [11], which is a dataset composed of CNN articles from 2021 along with their summary and the task consists in generating a summary given the full article.

To evaluate text summarization the most used metrics are based on text overlap, the most spread one is Rouge [12], given an integer n Rouge measures the number of overlapping n-grams between the generated and the reference text. We focus on Rouge-1, Rouge-2 and Rouge-L, the first two measure respectively the number of overlapping words, 1-grams, and the number of overlapping word pairs, 2-grams, between the reference and the generated text. The third, Rouge-L, measures the longest overlapping n-gram between the reference and the generated text.

An alternative metric, Bleu [13], is also based on quantifying text overlap, but it measures overlapping sub-strings instead of words.

3.1. Experimental Setup

To understand how well LLMs can summarize text in Latin, we measure the performance of three powerful LLMs, Llama 3.1 70b, Llama 3.1 405b and GPT-4o. The first two are openly available language models released from Meta [14] while the third is a closed source model from OpenAI [15]. We test these models in two settings, in the first, *format*, the model is asked to generate the *regestum* directly based on the full text it refers to in the second, *backtranslate*, when presented with the full text, the model is asked to initially write a "regestum" in English and then to translate it in Latin.

Each setting, format and backtranslate, is identified by the prompt we provide the LLM to make it generate the *regesta*, Table 1 shows the prompt used for each setting as well as a *Shared Prompt*, added next to the setting specific one, where we request the model to at least add the key elements of a *regestum*, as mentioned in Section 1.1: the author, the recipient, the summary, the date and the place. Finally, to facilitate the model during generation we add two full texts with their respective *regesta*.

We let the models generate up to 8048 tokens and we use greedy decoding, i.e. we pick the most likely word and avoid any form of sampling during inference since the *regestum* is meant to be a short and detailed summary, we will ablate different sampling techniques in future works.

Table 2
Summarization scores for gpt4o, Llama-70b-instruct-hf and Llama-405-instruct-hf, in **bold** the highest result for each metric.

model	dataset	experiment	n samples	rouge1	rouge2	rougeL	bleu
llama-3.1-70b-instruct-hf	mgh	format	2213	0.13	0.05	0.11	0.03
		backtranslate	2213	0.23	0.09	0.20	0.06
	auvray	format	2054	0.12	0.05	0.10	0.03
		backtranslate	2054	0.17	0.07	0.14	0.04
llama-3.1-405b-instruct-hf	mgh	format	2213	0.21	0.09	0.19	0.06
		backtranslate	2213	0.21	0.09	0.19	0.06
	auvray	format	2054	0.13	0.06	0.11	0.03
		backtranslate	2054	0.15	0.07	0.13	0.04
gpt-4o	mgh	format	2213	**0.39**	**0.18**	**0.34**	**0.16**
		backtranslate	2207	0.34	0.16	0.30	0.05
	auvray	format	2052	0.28	0.14	0.24	0.12
		backtranslate	2051	0.25	0.12	0.21	0.06

To evaluate model performance we use both a quantitative and a qualitative analysis: first, the quantitative analysis is based on Rouge and Bleu measuring the similarity between synthetic *regesta* generated by an LLM and the original ones from the REVERINO dataset summarizing the same text, second, the qualitative analysis is based on inspecting in detail a subset of the machine generated *regesta* to understand which of the 5 key properties they lack.

3.2. Quantitative Results

Table 2 shows Rouge and Bleu achieved by the three models we test: Llama 3.1 70b, Llama 3.1 405b and GPT-4o. The first finding is that no model can generate *regesta* proficiently, this appears from the fact that none of those we tested can achieve a Rouge higher than 0.40 and a Bleu above 0.15. We can also see that GPT-4o strongly outperforms both Llama models. The highest Rouge-1 is 0.38, achieved by GPT-4o on MGH, which also has the higher Rouge-1 on Auvray, although at a significantly lower value, 0.28, which we attribute to the lower quality of the Auvray dataset.

The wide gap between Rouge-1 and Rouge-2 shows how generally LLMs output texts that share the general context, higher value of 1-grams overlap, but they find it harder to generate actually similar texts, lower value for 2-grams overlap.

The two versions of Llama, Llama 3.1 70b and Llama 3.1 405b, show a small performance gap, indicating that it is not useful to use the larger and more costly Llama 3.1 405b, comparing *format* and *backtranslate* the first is the best setting for the GPT-4o model, while the opposite is true for Llama models, which show higher performance when asked to translate in English before writing in Latin. We have performed a limited prompt tuning that resulted in the choice of the format and backtranslate settings and we will further explore this aspect in future works.

Finally, we notice how in rare cases, between 2 and 6 for MGH and between 2 and 3 for Auvray, the guardrails preventing GPT-4o to answer questions involving violence make it refuse to generate a *regestum*, thus the lower values in the *N. Samples* column, while Llama models do not incur in this issue.

3.3. Qualitative Results

For a more in-depth analysis of the model abilities as seen in the quantitative results, we identify a corpus of 20 pairs of extended *regesta* (10 from MGH and 10 from Auvray) to conduct a qualitative analysis of the results. The *regesta* are humanly checked by a domain expert in their different versions produced by GPT-4o, Llama 3.1 70b and Llama 3.1 405b, for a total of 120 artificially created *regesta* reviewed and compared with their original versions. Through a manual inspection it is possible to identify the reasons for the results presented in Table 2 and possibly take corrective actions in the future.

In agreement with Table 2, also from a qualitative analysis GPT-4o performs better than both Llama models. This mainly concerns the generation of Latin text and thus the summarization of the document content in the *regestum* form. Indeed, Llama models encounter more problems in text generation, as shown by the fact that the best results are obtained when the summarization is created in English and then translated into Latin (i.e., backtranslate). More broadly, it can also be observed that the systems perform better in the case of MGH, but as already mentioned, this can be traced back to how the dataset is constructed.

Finally, the qualitative analysis reveals the most critical failures in automatic summarization (i.e., automatic *regesta* creation). One of the problems identified concerns the recognition of documents' author, namely the Pope. Indeed, in the case of MGH, which collects documents from multiple popes, both systems show difficulties in recognizing the correct author. In fact, GPT-4o correctly recognizes the Pope in 11 cases out on 20 taken into account. Another critical element concerns dating, which in these medieval texts is based on the year of pontificate and not on the modern dating system. Although Llama 3.1 70b, Llama 3.1 405b and GPT-4o recognize and identify the dating system used in the extended text of the medieval document and show that they have the tools to accomplish the conversion, both systems show difficulty in providing correct dating (either because they do not offer it or because they miscalculate it). Out of 20 manually inspected records generated by GPT-4o, only 3 cases correctly report the date of the document. The result is improved in the case of the recognition of the document recipient (often reported in the first line of the extended text), which in the same sample is recognized correctly by GPT-4o in 15 out of 20 cases.

Finally, it should be noted that in a few cases, since our prompt requests the *data topica* (the place) to be extracted, the systems correctly extract it even when the original *regestum* does not report this information. Thus lead to a lower score in the table, but to a qualitatively better result in *regesta* generation.

4. Conclusions

In this work we have developed the REVERINO dataset, a dataset of 4533 pairs of *regesta* with their respective full text (and apparatus). The texts in this dataset come from two collections of *regesta*, Epistolae saeculi XIII e regestis pontificum Romanorum selectae. (1216-1268) (MGH) and Les Registres de Gregoire IX (1227/41) (Auvray), to collect the dataset we have followed a pipeline composed of 4 steps: annotation, training, extraction and post-processing.

Despite containing more than 4000 samples, REVERINO is too small to be used as a training set for a language model that automatically generates *regesta*, however it can be used as a benchmark to test the ability of existing LLMs to do summarization in Latin and thus to develop better tools and methodologies in the future.

We have tested 3 LLMs among the best performing ones, our general finding is that these models can't be used as-is to summarize texts in Latin. More precisely, we find that GPT-4o is the best and that models from the Llama family are less able to generate text in Latin. Interestingly, for both Llama 3.1 70b and Llama 3.1 405b we find that initially translating to English is an effective technique to generate better *regesta*.

We also want to underline the limitations of our work, the samples in our dataset are automatically extracted, and therefore a share of them contain transcription errors and imperfections. However, we use the dataset only as a benchmark and it is still too small to serve as a training dataset for a text-summarization model.

Despite these limitations, we hope that REVERINO will foster future works on the development of Language Models proficient in Latin and we will continue improving on by extending it to grow larger than 10k samples, and by using it to train a custom Language Model specifically tailored to the generation of *regesta* in Latin.

Acknowledgments

This work was supported by project "Italian Strengthening of ESFRI RI RESILIENCE" (ITSERR) funded by the European Union under the NextGenerationEU funding scheme (CUP:B53C22001770006).

References

[1] ITSERR (Italian Strengthening of the ESFRI RI RESILIENCE), 2024. URL: https://itserr.it.

[2] A. Potthast, Regesta Pontificum Romanorum, Rudolf de Decker, 1874.

[3] G. H. Pertz, K. Rodenberg, Epistolae saeculi XIII e regestis pontificum Romanorum selectae. (1216-1268), volume 1-3, Weidmann, 1894.

[4] L. Auvray, Les Registres de Gregoire IX (1227/41), volume 1-3, Bibliothèque des Écoles françaises d'Athènes et de Rome, 1890 - 1918.

[5] B. Kiessling, R. Tissot, P. Stokes, D. Stoekl Ben Ezra, escriptorium: An open source platform for historical document analysis, 2019, pp. 19–19. doi:10.1109/ICDARW.2019.10032.

[6] S. Gabay, T. Clérice, Catmus-print [large], 2024. URL: https://doi.org/10.5281/zenodo.10592716. doi:10.5281/zenodo.10592716.

[7] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[8] F. Riemenschneider, A. Frank, Exploring large language models for classical philology, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23), Association for Computational Linguistics, Toronto, Canada, 2023. URL: https://arxiv.org/abs/2305.13698, to appear.

[9] M. Gambhir, V. Gupta, Recent automatic text summarization techniques: a survey, Artificial Intelligence Review 47 (2017) 1–66.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] S. Narayan, S. B. Cohen, M. Lapata, Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1797–1807. URL: https://aclanthology.org/D18-1206. doi:10.18653/v1/D18-1206.

[12] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[14] M. L. . Team, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[15] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023). URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/ARXIV.2303.08774. arXiv:2303.08774.