GlossAPI and Apertus: Building Open Greek Language Infrastructure for Transparent and Reproducible AI

Athanasios Priftis* Foivos Karounos-Zamparloukos† Panagiotis Skarvelis†

Myrsini Ioannou† Katerina Spanou†

Abstract

Greek remains structurally underrepresented in the data and model infrastructures that support contemporary natural language processing and large language models. This underrepresentation is not only a technical limitation. It creates a broader dependency problem: Greek-speaking researchers, educators, public institutions and technology firms are forced to rely on opaque proprietary models whose training data, linguistic coverage and adaptation pathways cannot be inspected, reproduced or democratically governed. GlossAPI addresses this gap by developing an open, documented and reusable infrastructure for the creation, processing, publication and evaluation of Greek AI-ready datasets.

This paper presents GlossAPI as public linguistic infrastructure for open Greek AI. GlossAPI functions both as an open-source Python library and as a reproducible data production pipeline. It processes heterogeneous Greek sources such as PDF, DOCX and HTML files, performs cleaning, normalization and annotation, and exports data in AI-ready formats such as Parquet and Markdown. The project has already produced and published a growing set of Greek datasets, including public consultations, encyclopedic texts, academic dissertations, Greek literature, legislation, press material and educational resources, with project documentation reporting approximately 7.95 billion cumulative tokens under Creative Commons licences.

The paper further proposes a workflow for integrating GlossAPI with Apertus, the fully open, transparent and multilingual language model family developed by ETH Zurich, EPFL and CSCS. The approved collaboration with the Swiss AI Initiative, "Enhancing multilingual foundation models through lexicographic grounding: advancing GlossAPI for Apertus Greek language integration," provides 50,000 GPU hours on the Alps supercomputing infrastructure. The work aims to integrate lexicographic knowledge from GlossAPI into Apertus multilingual fine-tuning, develop open multilingual evaluation datasets for Greek, and release datasets, benchmarks and checkpoints under open licences.

We make four contributions. First, we characterize GlossAPI as an open Greek data infrastructure rather than a dataset collection. Second, we introduce lexicographic grounding as a mechanism for connecting corpus-scale Greek language data with dictionaries, morphology, semantic relations and cultural context. Third, we propose an open evaluation layer for Greek, including morphology-aware, lexicon-aware and domain-specific benchmarks. Fourth, we frame Greek language AI as public digital infrastructure: open source, legally reusable, scientifically inspectable and governed in the public interest.

1 Introduction

Large language models (LLMs) have turned language data into a strategic infrastructure. The quality, legality, documentation and linguistic balance of training corpora now shape how citizens,

*HES-SO, HEG-Genève, Business Informatics Department.

†GlossAPI / GFOSS — Open Technologies Alliance.

researchers, educators, firms and public administrations access knowledge, communicate with machines and build downstream services. In this environment, language inequality is not a secondary problem. Languages with limited open, documented and legally reusable data become dependent on systems trained elsewhere, governed by others, and optimized primarily for higher-resource languages.

Greek illustrates this problem sharply. Modern Greek has a rich written tradition, a morphologically complex structure, extensive public-sector and legal usage, a major educational and cultural role, and deep historical continuity. Yet Greek remains underrepresented in the data infrastructure that underpins contemporary multilingual NLP and LLMs. Existing Greek NLP work has produced important resources and models, including GREEK-BERT (Koutsikakis et al., 2020), GreekBART (Evdaimon et al., 2023), Meltemi (Voukoutis et al., 2024), GreekMMLU (Zhang et al., 2026), and task-oriented toolkits (Loukas et al., 2025). However, the broader problem remains structural: Greek needs not only models and benchmarks, but a durable, reusable, openly licensed and reproducible data infrastructure.

GlossAPI was initiated by GFOSS / Open Technologies Alliance to address this gap. It is an open-source Python library and technical infrastructure for creating, processing and publishing Greek AI-ready datasets (GlossAPI, 2024, 2026; GFOSS, 2026a). Its pipeline supports heterogeneous document ingestion, text extraction, OCR-assisted processing, cleaning, normalization, annotation, sectioning and export to formats suitable for machine learning and NLP, including Markdown, Parquet and JSONL (GFOSS, 2026a). Its significance is not limited to data processing. GlossAPI operationalizes a public-interest approach to language AI: legally reusable data, open-source tooling, transparent documentation and community reuse.

This paper develops the argument that GlossAPI should be understood as public linguistic infrastructure for Greek AI. We use "infrastructure" deliberately. A dataset is a static artifact. A model is a trained parameterization. An infrastructure is a reproducible system for producing, updating, documenting, auditing and reusing artifacts across institutional settings. GlossAPI aims to supply this infrastructure layer for Greek. The collaboration with the Swiss AI Initiative and Apertus expands this from a national data-production effort to a European open model integration pathway (GFOSS, 2026c; Swiss AI Initiative, 2025). Apertus is particularly relevant because it is released not only with model weights, but with scientific artifacts intended to support transparent audit, reproducibility and extension (Swiss AI Initiative, 2025, 2026).

The paper is primarily a systems and methodology paper. It does not claim completed benchmark gains from Apertus Greek integration. Instead, it defines an architecture, a data production method, a lexicographic grounding strategy, and an evaluation program that can support reproducible empirical work. This distinction matters: open language infrastructure must be evaluated over time, through continuously updated data, transparent benchmarks and inspectable training or fine-tuning runs.

2 Background and Motivation

2.1 Open language models and reproducibility

The current LLM ecosystem contains a wide spectrum of openness. Many models are described as open because model weights are downloadable, but their training data, filtering logic, intermediate checkpoints, training logs and evaluation pipelines are only partially disclosed. This limits scientific reproducibility and weakens accountability. Initiatives such as BLOOM (BigScience Workshop, 2022), OLMo (Groeneveld et al., 2024), LLM360 (Liu et al., 2023), and Apertus (Swiss AI Initiative, 2025) move toward a stronger conception of openness: open weights plus training data descriptions, code, evaluation infrastructure and documentation.

Apertus is especially relevant to this paper because it was designed as a fully open, multilingual and transparent model family. The Apertus technical report describes 8B and 70B models trained on approximately 15 trillion tokens across more than 1,800 languages, with a substantial

non-English share and with scientific artifacts released for audit and extension (Swiss AI Initiative, 2025). The engineering report describes the training campaign on the Alps supercomputer and positions the model as an academic-scale, fully open 70B effort (Swiss AI Initiative, 2026). This makes Apertus a suitable base for studying Greek adaptation under conditions that are more transparent than proprietary API-based systems.

2.2 Greek NLP and the infrastructure gap

Greek NLP has advanced significantly. GREEK-BERT demonstrated the value of monolingual pretraining for Modern Greek (Koutsikakis et al., 2020). GreekBART extended pretrained sequence-to-sequence modeling to Greek (Evdaimon et al., 2023). Meltemi adapted a 7B model through continued pretraining on a 40-billion-token Greek corpus and released Greek-oriented instruction-tuned variants (Voukoutis et al., 2024). GreekMMLU introduced a native-sourced benchmark of 21,805 multiple-choice questions across 45 subject areas, directly addressing the weakness of evaluation datasets translated from English (Zhang et al., 2026). The GR-NLP-TOOLKIT provides open-source components for core Greek NLP tasks such as part-of-speech tagging, morphological tagging, dependency parsing, named entity recognition and Greeklish-to-Greek transliteration (Loukas et al., 2025).

These contributions are essential, but they also clarify the remaining gap. The field needs a common layer for continuously producing and documenting Greek corpora across domains: public administration, law, education, literature, journalism, scientific dissertations, public consultations, cultural heritage and web-native material. Without this layer, Greek model development risks becoming a sequence of isolated projects with inconsistent licensing, inconsistent cleaning, limited provenance and weak reproducibility. GlossAPI addresses precisely this infrastructure layer.

2.3 Language infrastructure as a public good

GlossAPI's public-interest framing is aligned with the broader argument that open datasets, open models and open standards are necessary for democratic control over AI systems. GFOSS publications[1] argue that Greek should be treated as critical digital infrastructure in the age of foundation models and that open, documented language resources are preferable to waiting for proprietary platforms to improve Greek on their own timetable (GFOSS, 2026c). Related GlossAPI/GFOSS analyses emphasize that open-source adoption requires deployable infrastructure, documentation and support rather than mere catalogues (GFOSS, 2026b); that local and open-source AI can reduce dependency and improve institutional resilience (GFOSS, 2026f); and that RAG systems require traceability, source verification and post-generation inspection to limit hallucinations in public-interest applications (GFOSS, 2026d).

For Greek, the public-good argument has three dimensions. Scientifically, reproducible corpora enable comparable experiments. Economically, open resources reduce dependency on closed APIs and create local capacity for integration, hosting, fine-tuning and support. Democratically, inspectable data and models allow public institutions to justify, audit and contest AI-assisted outputs.

3 GlossAPI Architecture

GlossAPI is designed as a modular corpus production pipeline. Its architecture separates document acquisition, extraction, cleaning, annotation, export and documentation. This separation

[1] GFOSS / Open Technologies Alliance is the organization that develops and maintains GlossAPI; the GFOSS analyses cited in this paper therefore function as primary project and institutional documentation rather than third-party commentary. Each is a referenced, evidence-based analysis that draws on further primary and scholarly sources.

is important because Greek sources vary substantially in format, typography, orthography, encoding quality, OCR difficulty and licensing status.

3.1 Pipeline stages

Table 1 summarizes the pipeline.

Table 1: GlossAPI as a reproducible Greek corpus production pipeline.

Stage	Function	Typical outputs
Source registration	Identify source, licence, collection method and provenance	Source manifest, licence record, crawl or download plan
Acquisition	Download or collect PDF, DOCX, HTML or structured data	Raw files, metadata tables, download logs
Extraction	Convert documents to text or structured Markdown; apply OCR where needed	Markdown, JSON, extraction metrics
Cleaning	Remove boilerplate, corrupted text, duplicated segments and layout noise	Clean Markdown, quality flags, noise metrics
Sectioning	Segment documents into semantically useful units	Section tables, document structure records
Annotation	Add metadata, domain labels, language tags and optional linguistic annotations	Annotated Parquet, JSONL, metadata index
Export	Produce ML-ready formats for training, retrieval and evaluation	Parquet, Markdown, JSONL, dataset card
Documentation	Publish provenance, licence, processing steps and known limitations	Dataset card, datasheet, reproducibility notes

3.2 Corpus domains

GlossAPI's current and intended corpus coverage includes public consultations, encyclopedic content, academic dissertations, Greek literature, legal and administrative texts, press material, educational resources and cultural heritage collections (GlossAPI, 2024, 2026). Across these domains, the published collections currently total approximately 7.95 billion tokens released under Creative Commons licences, as reported in the project repository and the dataset documentation of the GlossAPI Hugging Face organization (GFOSS, 2026a; GlossAPI, 2026). This domain diversity matters for model development. A model trained only on general web text may underperform on legal Greek, administrative Greek, educational Greek, literary registers, historical language forms, dialectal forms or formal institutional writing. Conversely, a model trained only on institutional texts may fail at conversational and cultural usage.

3.3 Documentation and legal reuse

Data legality is not an afterthought. Opaque web crawling can produce corpora that are technically large but legally fragile. GlossAPI therefore treats licence status, provenance and dataset documentation as first-order metadata. This aligns with established documentation practices such as Data Statements (Bender and Friedman, 2018), Datasheets for Datasets (Gebru et al., 2021), and Model Cards (Mitchell et al., 2019). For Greek AI, the relevant question is not merely whether a corpus can be used internally, but whether it can be redistributed, inspected, cited, updated and reused by universities, public institutions, firms and civil society.

3.4 Relation to existing document-processing tools

GlossAPI's extraction stage builds on Docling (Auer et al., 2024), the open-source document-conversion toolkit that performs layout analysis and table-structure recognition and emits a unified, richly structured representation of PDFs, Word documents and other formats. In this respect GlossAPI shares an extraction engine with other Docling-based tools, most notably spaCy Layout (Explosion AI, 2024), which likewise converts documents into structured, AI-ready output with labelled sections and tables and is well suited to chunking for retrieval-augmented generation. The difference is one of scope rather than of competition. Docling and spaCy Layout solve the extraction and structuring problem: they turn an individual document into clean, labelled text, tables and sections. GlossAPI treats extraction as a single stage inside a larger corpus-production process and adds the layers that an open language infrastructure requires but that general-purpose extractors do not aim to provide: source registration and provenance tracking, licence metadata, Greek-specific cleaning and normalization, sectioning and annotation tuned to Greek registers, reproducible dataset cards, and published, openly licensed corpora. GlossAPI is therefore best understood not as an alternative to Docling-based extractors but as a Greek-focused data infrastructure built on the same extraction foundation, extending it from "parse this document" to "produce, document and publish reusable Greek datasets."

4 From Corpora to Lexicographic Grounding

4.1 Why lexicographic grounding?

LLMs learn statistical regularities from corpora, but corpus-scale learning alone does not guarantee robust handling of morphology, etymology, polysemy, register, derivational families, semantic relations or historical variation. This is particularly important for Greek. A Greek model must manage inflectional morphology, compound formation, katharevousa forms, ancient and medieval inheritance, dialectal forms, domain-specific terminology and orthographic variation. Lexicographic grounding aims to connect corpus-scale learning with structured linguistic knowledge.

We define lexicographic grounding as the integration of dictionary-like, morphology-aware and semantic resources into model adaptation and evaluation. It does not replace pretraining or fine-tuning. It adds a structured layer that can be used for:

- constructing training examples that expose models to lemma–form relations, definitions, semantic contrasts and derivational families;
- evaluating whether models preserve meaning across inflected forms and paraphrases;
- improving retrieval and entity normalization in RAG systems;
- aligning generated explanations with verified lexical and semantic resources;
- measuring domain transfer across legal, educational, literary and administrative Greek.

4.2 Grounding workflow for Apertus Greek integration

The approved Swiss AI Initiative collaboration provides a practical setting for this workflow. The proposed process contains six steps:

- **Corpus selection.** Select legally reusable GlossAPI datasets across registers and domains.
- **Lexicographic resource construction.** Build or integrate lemma, inflection, definition, synonymy, antonymy, etymology and semantic relation resources where licences permit reuse.
- **Example generation.** Convert structured lexical knowledge into training examples, contrastive pairs, retrieval entries and evaluation items.

- **Apertus adaptation.** Use continued pretraining, supervised fine-tuning or parameter-efficient adaptation to improve Greek handling while preserving multilingual competence.
- **Evaluation.** Evaluate on GreekMMLU, existing Greek NLP tasks, morphology-aware probes, lexicon-aware tasks and domain-specific benchmarks.
- **Release.** Publish datasets, benchmarks, processing scripts, model checkpoints or adapters, model cards and evaluation reports under open licences.

4.3 Training data mixtures

A central design decision is mixture composition. We propose that Greek integration experiments maintain explicit source proportions rather than collapse all data into a single undifferentiated corpus. This enables ablation studies. For example, one can compare:

- general Greek corpora only;
- general corpora plus legal and administrative corpora;
- general corpora plus lexicographic examples;
- domain-balanced corpora;
- temporally stratified corpora;
- mixtures with and without dialectal or historical material.

Such mixture transparency is necessary for scientific interpretation. If performance improves on legal tasks but decreases on educational or conversational tasks, the source mixture should make that trade-off diagnosable.

5 Open Evaluation Layer for Greek

5.1 Evaluation principles

Open Greek AI requires evaluation beyond generic multilingual benchmarks. Machine-translated benchmarks are useful but insufficient: they often fail to capture Greek-specific morphology, cultural references, administrative forms, educational content and idiomatic usage. GreekMMLU directly responds to this problem by using native-sourced Greek questions (Zhang et al., 2026). GlossAPI can complement this direction by contributing open evaluation datasets tied to documented corpora and reproducible generation procedures.

We propose five evaluation families.

5.2 General language understanding

This layer includes GreekMMLU, Greek reading comprehension, classification, inference and question-answering tasks. It should report performance by domain and difficulty level, not only aggregate accuracy.

5.3 Morphology-aware evaluation

Greek morphology requires targeted evaluation. Tasks should test:

- lemma recognition across inflected forms;
- agreement and case-sensitive generation;
- handling of clitics and compound forms;
- transformation between active and passive forms;
- consistency across orthographic variants;
- preservation of meaning across inflectional paraphrases.

5.4 Lexicon-aware semantic evaluation

This layer evaluates whether a model can use definitions, synonymy, antonymy, derivational relations, semantic fields and domain-specific terminology. It is especially useful for measuring lexicographic grounding. Possible tasks include definition selection, lemma–definition matching, semantic relation classification and contrastive word sense disambiguation.

5.5 Domain and register evaluation

Greek public-interest AI must work in domains where precision matters: law, public administration, health, education, cultural heritage and journalism. Domain evaluation should include:

- legal and administrative question answering with source citations;
- public-service procedure retrieval;
- educational item answering;
- cultural heritage description;
- scientific dissertation summarization;
- press and public consultation classification.

5.6 Faithfulness and RAG evaluation

RAG is central for public-sector and knowledge-intensive use. However, retrieval alone does not guarantee correctness. A system may retrieve the right evidence and still generate an answer that contradicts it. GlossAPI/GFOSS analyses identify failure modes including ungrounded confidence, factual contradiction, hallucinated entities, negation reversal and answer drift (GFOSS, 2026d). We therefore propose a faithfulness evaluation layer:

- answer claims must be traceable to retrieved passages;
- numerical and date claims must match evidence;
- named entities and citations must appear in retrieved sources;
- negation must be preserved;
- index and prompt updates must be monitored for answer drift.

6 Public-Interest Use Cases

GlossAPI and Apertus Greek integration can support both research and practical deployments.

6.1 Education

Open Greek language models can support feedback, summarization, reading assistance, accessible educational materials and teacher tools. Because education is a high-trust setting, open models and documented datasets are preferable to opaque systems that cannot be audited for linguistic coverage, bias or data provenance.

6.2 Public administration

Public administration uses formal Greek, legal references and procedural language. AI assistance should not replace accountable decision-making, but it can support search, summarization, drafting, translation, classification and citizen guidance. In this context, systems should operate with verified knowledge bases, source citations, logging, versioning, audit trails and domain-expert review (GFOSS, 2026e). GlossAPI can provide the documented corpora and RAG indexes needed for such systems.

6.3 Cultural heritage and digital humanities

Greek spans ancient, medieval, katharevousa, modern and dialectal varieties. A fully open Greek language infrastructure can serve philology, history, literature, theatre, archives, museums and public cultural institutions. This does not imply collapsing all Greek varieties into one undifferentiated model. It requires metadata-rich corpora and evaluation by period, register and domain.

6.4 Research and SMEs

Open infrastructure allows researchers and small firms to build on shared resources rather than pay repeatedly for opaque proprietary services. It also enables local hosting, domain adaptation, reproducible benchmarking and professional support ecosystems. This aligns with the broader open-source adoption argument that Europe needs practical deployment pathways, documentation and support services, not merely policy declarations (GFOSS, 2026b).

7 Governance, Licensing and Reproducibility

7.1 Open releases

A complete release should include:

- dataset files in standard formats;
- source manifests and licence metadata;
- processing scripts and versioned pipeline configuration;
- dataset cards and datasheets;
- benchmark construction scripts;
- evaluation reports;
- model cards for fine-tuned models or adapters;
- known limitations and recommended-use statements.

7.2 Human oversight

A public language infrastructure does not remove the need for human responsibility. In administrative, legal, medical or educational settings, model outputs must remain subordinate to accountable human judgment. The appropriate role for LLMs is assistance under traceability and review, not autonomous decision-making.

7.3 Versioning and contamination control

Open evaluation requires contamination control. Benchmarks should be versioned, split into public and private portions where needed, and accompanied by source provenance. GreekMMLU's public/private split for leaderboard robustness offers one useful precedent (Zhang et al., 2026). GlossAPI-based benchmarks should follow similar principles, especially where datasets are also used for training.

8 Limitations and Future Roadmap

This paper defines an architecture and research program rather than reporting final training results. This is an intentional limitation: the contribution is to specify the infrastructure, data-production method, grounding strategy and evaluation agenda that can support reproducible empirical work in later phases.

First, open licensing does not automatically imply high linguistic quality. Future work must continue to address corpus cleaning, OCR correction, deduplication, register balancing and metadata validation as core engineering tasks rather than as secondary preprocessing steps.

Second, lexicographic grounding depends on the availability and legal usability of lexical and semantic resources. Where existing resources cannot be reused under open licences, the project should prioritize the construction of new open lexicographic assets, including lemma inventories, inflectional resources, definitions, semantic relations and domain terminology.

Third, Greek adaptation must be evaluated without losing sight of multilingual competence. Continued pretraining, supervised fine-tuning or parameter-efficient adaptation should therefore report both Greek-specific benchmarks and broader multilingual metrics, so that improvements and trade-offs remain visible.

Fourth, public-sector deployment requires more than a model. Practical use in administration, education or legal settings depends on procurement rules, secure hosting, access control, monitoring, logging, human review and institutional accountability.

Fifth, openness does not eliminate risk. Open models and datasets can be used irresponsibly, but the appropriate response is not proprietary closure. It is documentation, risk classification, red-teaming, monitoring, clear recommended-use statements and operational conditions for public-interest deployment.

9 Conclusion

Greek AI should not be treated as a downstream localization problem to be solved by proprietary systems when market incentives allow. It should be treated as public linguistic infrastructure. GlossAPI provides the data production layer needed for this shift: open-source tooling, documented corpora, reproducible processing and legally reusable outputs. Apertus provides a complementary open-model foundation: multilingual, transparent and designed for audit and adaptation.

The integration of GlossAPI and Apertus offers a practical path toward open Greek AI: corpus-scale data for continued pretraining, lexicographic grounding for morphology and semantics, and open evaluation for reproducible measurement. The broader contribution is institutional as much as technical. When language infrastructure is open, documented and reusable, universities, public institutions, firms and citizens can participate in shaping how Greek is represented in AI systems. That is the difference between consuming foreign black boxes and co-designing public digital infrastructure.

Data and Code Availability

GlossAPI code is available through the GFOSS GitHub repository. GlossAPI datasets are published through the GlossAPI Hugging Face organization and related project pages. Future Apertus Greek integration artifacts should include datasets, processing scripts, evaluation suites and model checkpoints or adapters under open licences whenever legally possible.

Acknowledgements

The authors acknowledge GFOSS / Open Technologies Alliance, the GlossAPI contributors, and the Swiss AI Initiative. The approved project "Enhancing multilingual foundation models through lexicographic grounding: advancing GlossAPI for Apertus Greek language integration" provides compute support on the Alps supercomputing infrastructure for Greek language integration experiments.

References

Auer, Christoph, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. "Docling Technical Report." arXiv preprint arXiv:2408.09869.

Bender, Emily M., and Batya Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6: 587–604.

BigScience Workshop. 2022. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model." arXiv preprint arXiv:2211.05100.

Evdaimon, Iakovos, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou. 2023. "GreekBART: The First Pretrained Greek Sequence-to-Sequence Model." arXiv preprint arXiv:2304.00869.

Explosion AI. 2024. "spaCy Layout: Process PDFs, Word Documents and More with spaCy." Software, integrating Docling for structured document processing. https://github.com/explosion/spacy-layout.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

GFOSS / Open Technologies Alliance. 2026a. "Eellak/glossAPI: Greek Dataset Production from PDF+." https://github.com/eellak/glossAPI.

GFOSS / Open Technologies Alliance. 2026b. "From Catalogues to Real Adoption: Why Europe Must Back Open Source That People Can Actually Use." https://gfoss.eu/from-catalogues-to-real-adoption-why-europe-must-back-open-source-that-people-can-actually-use/.

GFOSS / Open Technologies Alliance. 2026c. "GlossAPI and the Swiss AI Initiative: A Strategic Investment in Greek as Public Digital Infrastructure." https://gfoss.eu/glossapi-and-the-swiss-ai-initiative-a-strategic-investment-in-greek-as-public-digital-infrastructure/.

GFOSS / Open Technologies Alliance. 2026d. "How to Reduce Hallucinations in RAG Systems." https://gfoss.eu/how-to-reduce-hallucinations-in-rag-systems/.

GFOSS / Open Technologies Alliance. 2026e. "Large AI Models Should Not Be Used in Administration Without Guardrails, Verification, and Experienced Human Review." https://gfoss.eu/large-ai-models-should-not-be-used-in-administration-without-guardrails-verification-and-experienced-human-review/.

GFOSS / Open Technologies Alliance. 2026f. "Low-Cost, Open-Source Artificial Intelligence Models." https://gfoss.eu/low-cost-open-source-artificial-intelligence-models/.

GlossAPI. 2024. "GlossAPI: Open Infrastructure for Greek AI-Ready Data." https://glossapi.gr/aboutus.html.

GlossAPI. 2026. "glossAPI Hugging Face Organization." https://huggingface.co/glossAPI.

Groeneveld, Dirk, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. "OLMo: Accelerating the Science of Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15789–15809. Bangkok, Thailand: Association for Computational Linguistics.

Koutsikakis, John, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. "GREEK-BERT: The Greeks Visiting Sesame Street." arXiv preprint arXiv:2008.12014.

Liu, Zhengzhong, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023. "LLM360: Towards Fully Transparent Open-Source LLMs." arXiv preprint arXiv:2312.06550.

Loukas, Lefteris, Nikolaos Smyrnioudis, Chrysa Dikonomaki, Spiros Barbakos, Anastasios Toumazatos, John Koutsikakis, Manolis Kyriakakis, Mary Georgiou, Stavros Vassos, John Pavlopoulos, and Ion Androutsopoulos. 2025. "GR-NLP-TOOLKIT: An Open-Source NLP Toolkit for Modern Greek." In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, 174–182. Abu Dhabi, UAE: Association for Computational Linguistics.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29.

Swiss AI Initiative. 2025. "Apertus: Democratizing Open and Compliant LLMs for Global Language Environments." arXiv preprint arXiv:2509.14233.

Swiss AI Initiative. 2026. "An Engineering Journey Training Large Language Models at Scale on Alps: The Apertus Experience." arXiv preprint arXiv:2604.12973.

Voukoutis, Leon, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, and Vassilis Katsouros. 2024. "Meltemi: The First Open Large Language Model for Greek." arXiv preprint arXiv:2407.20743.

Zhang, Yang, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang, and Michalis Vazirgiannis. 2026. "GreekMMLU: A Native-Sourced Multitask Benchmark for Evaluating Language Models in Greek." arXiv preprint arXiv:2602.05150.