

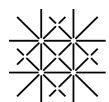
# Large Language Models and their role in Latin American Cultural Heritage

A proposal for the development of a Fine-Tuned LLM

Paola Melissa Lechuga Santin  
pao.lechugasantin@unibas.ch, 22-068-167  
 0009-0001-3417-2906

August 15, 2025

University of Basel  
Digital Humanities Lab  
Switzerland



University  
of Basel

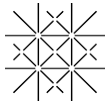


Digital  
Humanities  
Lab

# Table of Contents

- 1 Introduction** **3**
  
- 2 Contextual Framework** **4**
  - 2.1 Historical Framework . . . . . 4
    - 2.1.1 Immigrant Patterns . . . . . 4
  - 2.2 Cultural Contributions . . . . . 4
  - 2.3 Theoretical Framework . . . . . 6
    - 2.3.1 Integration of Models in the Swiss Context . . . . . 6
  - 2.4 Macro Context . . . . . 6
    - 2.4.1 Basel as Switzerland’s Cultural Capital . . . . . 6
    - 2.4.2 Relevance to Latino Communities . . . . . 7
  - 2.5 Research Gaps . . . . . 7
    - 2.5.1 Implications of Becoming an Immigrant . . . . . 7
    - 2.5.2 Limited Historical Documentation . . . . . 8
    - 2.5.3 Community Organizations . . . . . 9
  
- 3 Conceptual Framework** **10**
  - 3.1 Operational Definitions . . . . . 10
  
- 4 Technological Frameworks in Cultural Preservation** **13**
  - 4.1 Digital Documentation Methodologies . . . . . 13
  - 4.2 AI-Driven Preservation Strategies . . . . . 14
    - 4.2.1 Digitization and Reconstruction of Physical Artifacts . . . . . 14
    - 4.2.2 Natural Language Processing for Textual Heritage . . . . . 15
    - 4.2.3 Predictive Conservation and Risk Mitigation . . . . . 15
    - 4.2.4 Virtual and Augmented Reality Experiences . . . . . 16
  - 4.3 GPT Models for Informal Cultural Heritage Preservation . . . . . 16
    - 4.3.1 Oral History Processing and Analysis . . . . . 16
    - 4.3.2 Community-Driven Archiving . . . . . 17
    - 4.3.3 Ethical and Technical Challenges . . . . . 17
  
- 5 Source Description** **18**
  - 5.1 Source Guidelines . . . . . 18
    - 5.1.1 Consent and Age Restrictions . . . . . 18
    - 5.1.2 Use of Aliases or Anonymization . . . . . 19
    - 5.1.3 Data Sensitivity and Accessing . . . . . 19
    - 5.1.4 Cultural Sensitivity and Community Engagement . . . . . 19
    - 5.1.5 Data Quality and Attribution . . . . . 20
    - 5.1.6 Handling of Sensitive and Historical Content . . . . . 20
    - 5.1.7 Compliance with Legal and Ethical Frameworks . . . . . 20
    - 5.1.8 Continuous Consent and Review . . . . . 20
  
- 6 Methodologies** **21**
  - 6.1 Archival Research Methods . . . . . 21
  - 6.2 Workflow Establishment . . . . . 22

<b>7</b>	<b>Domain-Specific GPT Proposal</b>	<b>23</b>
7.1	Theoretical Proposal: GPT-Based Preservation Tool . . . . .	23
7.1.1	Objectives . . . . .	24
7.2	Rationale for GPT Approach . . . . .	24
7.3	Data Sources and Preparation . . . . .	25
7.4	Fine-Tuning an AI Model . . . . .	29
7.4.1	Fine-Tuning Techniques and Tools . . . . .	31
7.4.2	Practical Considerations for Fine-Tuning . . . . .	31
7.4.3	Code Example to Fine-Tune with LoRA and Axolotl . . . . .	32
7.5	Deployment and Integration . . . . .	38
7.5.1	Cloud Hosting and API Wrapping . . . . .	39
7.5.2	HTML Interface and Langflow . . . . .	40
7.5.3	Archival Database Integration . . . . .	48
7.5.4	Large-Scale deployments . . . . .	51
7.5.5	Security Considerations . . . . .	54
7.6	Documentation and Maintenance . . . . .	54
7.6.1	Documentation of the Codes and Workflows . . . . .	54
7.6.2	Maintenance Strategy . . . . .	55
7.6.3	User Documentation . . . . .	57
7.6.4	Data Governance and Provenance . . . . .	58
7.7	Model Capabilities and Limitations . . . . .	61
7.8	Ethical and Practical Considerations . . . . .	66
<b>8</b>	<b>Conclusions</b>	<b>69</b>
	<b>References</b>	<b>71</b>
<b>A</b>	<b>Appendix</b>	<b>81</b>
A.1	Example of Consent Form: . . . . .	81



## Erklärung zur wissenschaftlichen Redlichkeit

Ich bestätige hiermit, dass ich vertraut bin mit den Regelungen zum Plagiat der «Ordnung der Philosophisch-Historischen Fakultät der Universität Basel für das Bachelorstudium vom 25. Oktober 2018» (§21) bzw. der «Ordnung für das Masterstudium vom 25. Oktober 2018» (§25) und die Regeln der wissenschaftlichen Integrität gewissenhaft befolgt habe. Die vorliegende Arbeit ist ausserdem weder ganz noch teilweise an einer anderen Fakultät oder Universität zur Begutachtung eingereicht und/oder als Studienleistung, z.B. in Form von Kreditpunkten verbucht worden.

Des Weiteren versichere ich, sämtliche Textpassagen, die unter Zuhilfenahme KI-gestützter Programme verfasst wurden, entsprechend gekennzeichnet sowie mit einem Hinweis auf das verwendete KI-gestützte Programm versehen zu haben.

Eine Überprüfung der Arbeit auf Plagiate und KI-gestützte Programme – unter Einsatz entsprechender Software – darf vorgenommen werden.

Bei begründetem Verdacht auf eine unerlaubte oder nicht gekennzeichnete Anwendung von KI verpflichte ich mich, an der Klärung des Sachverhaltes mitzuwirken, z.B. durch Teilnahme an einem Gespräch.

Ich habe zur Kenntnis genommen, dass unlauteres Verhalten zu einer Bewertung der betroffenen Arbeit mit einer Note 1 oder mit «nicht bestanden» bzw. «fail» oder zum Ausschluss vom Studium führen kann.

Name, Vorname: Lechuga, Paola

---

Titel der schriftlichen Arbeit:

**Large Language Models and their role in Latin American Cultural Heritage:  
A proposal for the development of a Fine-Tuned LLM**

Datum: 14/8/25

Unterschrift:  
1

## On the Use of AI

This paper, including the packages presented and their documentation, were written by me. No text or software snippets were automatically generated by AI; nevertheless, AI tools were an integral part of the development of the proposal and were also used, for example, for grammatical and syntax verification of the text and codes. Blackbox AI was used in the development of the fine-tuning and deployment codes (See Chapter 8). Blackbox AI is a tool that helps to complete, generate and verify coding languages. This is useful for making corrections automatically. Blackbox and Overleaf's AI assistants were used to spellcheck and suggest grammar corrections. This document, the Python codes, bash codes and HTML codes and all its related components have been written in English. The tools mentioned were only used to point out obvious formulation and spelling mistakes. It should be noted that, strictly speaking, not even the use of a traditional search engine can take place without technologies that fall within the AI spectrum. Working without the use of modern AI-supported infrastructure is practically impossible and not every application is as transparent as the generation of AI citations for a dedication. That being said - this is the work has been written by a human.

# 1 Introduction

One of the most common difficulties immigrants face is the expectation to conform. Many arrive in a country where the social norms, customs, and even daily interactions are unfamiliar. At work, in schools, and in public spaces, there's often an unspoken pressure to fit in—to dress a certain way, adjust speech patterns, or behave in ways that align with the dominant culture. (Wilde, 2025)

Moving to another country often implies adapting to new cultural practices. Immigrant communities frequently employ different strategies to preserve parts of their culture despite the need of involving themselves in the cultures and traditions from their new country of residence. Some of these often include the retention of language, in which they normally speak their native language at home or with family members, continuing to celebrate traditional festivities, rituals, and customs from the homeland, create support groups with people coming from the same country, teach new generations about traditions and their history, and sometimes buy, travel with or create small physical memories from their country as sentimental value and a way to remember their country.

In 2016, the Swiss Federal Statistical Office released statistics on two million foreigners living in the country. Among the most common foreign nationalities in Switzerland, people from Italy, Germany, Portugal, France, and Spain were found. In 2023, 9,315 immigrants from the Americas were registered as permanent residents in the country. Although it is unknown the specific amount of Latin American immigrants currently living in Switzerland, particularly in Basel. The previously mentioned statistics infer that a small part of the population coming from countries like Brazil, Mexico, and Peru, etc, reside permanently or non-permanently in the country.

This project aims to determine how Latin American migrant communities can preserve their culture, identifying their advantages, disadvantages, maintenance challenges, and explore possible digital alternatives for cultural preservation. Through the use of large language models (LLMs) and generative pre-trained transformers (GPTs)<sup>1</sup> this project aims to identify how these communities can preserve, share and educate others about their traditions, customs and history.

---

<sup>1</sup>Refer to Chapter: Operational Definitions

## 2 Contextual Framework

### 2.1 Historical Framework

#### 2.1.1 Immigrant Patterns

Early migration records in the Basel State Archives contain a large dossier of documents from the early 20th century that prove existence of the lives of various immigrants; from handwritten letters to autobiographies, these documents state how immigration has played an important role in Basel's society.

In these archives, examples like the life of Carlos Reyes, a Spanish immigrant who attempted to settle in Basel around 1932 can be found, as well as other interesting cases. Although his records in the country are not vast and could be considered negative for the cultural enrichment of Basel, the life of Carlos Reyes in Basel proves to be significant in terms of migratory patterns and possible settlements of Spanish-speaking communities.

However, during 1960 to 1970, the presence of Spanish-speaking communities in Basel increased. According to Claudio Bolzman, with dictatorships gaining power in South America, the growth of Latin American immigrants in Switzerland started increasing exponentially, to the point that these immigrants were now considered an exception to the common amount of Latin Americans residing temporarily in the country.

Bolzman also describes four types of Latin Americans living in Switzerland, as quoted by him these four types were categorized as "the Europeanized, the smugglers, the exiles and the relocated".

From wealthy families sending their children to highly prestigious boarding schools, to art merchants and writers like the famous author Julio Cortázar and political asylum refugees, these people became an important part of the social, economic and cultural structure of Switzerland. (Bolzman, [2004](#))

### 2.2 Cultural Contributions

Although there is no registry of large cultural contributions from Spanish-speaking or Latin American communities in Switzerland, a few examples of community events, museum exhibitions, convenience stores, and publications make the list.

Examples like the Spanish Catholic Mission in Basel<sup>2</sup> offer a space for those with

---

<sup>2</sup>In comparison with the Basler Protestant Mission, which focuses primarily in providing help for Asian communities, but doesn't mention special efforts to join migrant communities in Basel

religious beliefs to congregate and participate not only in catholic rituals but also form a closed relationship with those who share the same history, country of origin, religious approaches and traditions. The Mission until today has invited a now larger amount of Latin American and Spanish immigrants to form part of their community. (“Nuestra Identidad | Misión católica de lengua española de basilea”, [n.d.](#))

It is important to also mention the other contributions to immigrant community integration done, such as "Active Asyl", a non-profit organization which intended to support immigrant communities teaching them useful skills, engage in sport events and be part of casual gatherings and game nights. Since 2023, the organization seems to not be active anymore.

"Nosotras", founded in 1995 by immigrant Delia Krieg-Trujillo, is a non-profit organization, independent of political or religious affiliation, whose main objective is addressing social and cultural issues related to migration in Basel. Over the years, this organization has been able to integrate new immigrants into Basel's society without losing sight of the importance of preserving the Latino culture and traditions. (“Nosotras Basel”, [n.d.](#))

Besides offered language courses, this organization provides official translations, computer skills and legal and medical assistance, they also organize a range of events. "Encuentro Culinario"<sup>3</sup> is a small event where people living in Basel, immigrant or not, can taste different dishes from Latin American countries. This event has opened the possibility not only for other immigrants but for locals too, to learn about traditional dishes, food techniques, and flavors traditionally found in northern, central, and South America.

A similar type of event can be found by La Tienda Latina, where every weekend they offer traditional Latin American dishes throughout the day, inviting immigrants and their families to engage into social and cultural activities found in their countries. (“La Tienda Latina Basel”, [n.d.](#))

Nosotras also offers the annual event "Gala Latina" in which immigrants are invited to experience a night of eating, listening and dancing to Latin American food and music. This event offers the chance to meet other people that similarly have relocated to Switzerland, whether it was 50 years ago or recently.

---

<sup>3</sup>English: Culinary Meeting

## 2.3 Theoretical Framework

### 2.3.1 Integration of Models in the Swiss Context

Switzerland is a country often known for having different communities co-existing side by side instead of together as one. (Mathari, 2024) However, Basel offers a wide range of programs and integration policies to promote multiculturalism, such as:

1. Recognizing cultural and religious minorities, for example the Alevi Community. (“Alevi Communities In Western Europe”, 2011)
2. Actively supporting immigrant associations, for example GGG Migration. (“Geschichte | GGG Migration”, n.d.)
3. Contrasting integration strategies with Zurich’s stricter assimilationist policies. For example, free german lessons for those new to the Canton. <sup>4</sup> (D’Amato, n.d.)

## 2.4 Macro Context

### 2.4.1 Basel as Switzerland’s Cultural Capital

Basel has up to almost 40 museums, including institutions like the Kunstmuseum (the oldest public art collection in the world), Fondation Beyeler, and Museum Tinguely. These museums attract audiences from around the world and contribute to Basel’s reputation as a cultural hub. The city is also filled with libraries and numerous theaters, including Theater Basel, Switzerland’s largest multipurpose theater, which has won multiple awards and hosts opera, ballet, and drama performances.

Art Basel, as another example, is one of the biggest modern and contemporary art fairs, which also originated in Basel in 1970. It has now become a a very known event with editions in Miami Beach, Hong Kong, and Paris. The event invites galleries, collectors, artists, and enthusiasts from around the world to showcase art, proving that is full of multiculturalism. (“Basel as Cultural Capital”, n.d.)

Basel also has centers for classical music with venues like the Stadtcasino Basel. Basel often hosts international orchestras such as the Sinfonieorchester Basel. The city is also full of diverse musical tastes with jazz clubs (e.g., Bird’s Eye or the Em Bebbi Sy Jazz Festival) and festivals like the Baloise Session and Groove Now Blues Weeks. (“Cultural Capital of Switzerland”, 2024)

---

<sup>4</sup>Assimilationist policies refer to those in which immigrants should completely adapt to the new culture, disregarding their own culture and traditions.

With residents from about 160 countries, Basel's cultural life reflects its diversity and how immigrants have in a way or another, made an impact. Immigrant communities contribute to the city's cultural landscape through associations, music and food festivals, and artistic activities.

On the other hand, Basel is also geographically situated at the tri-border region of Switzerland, Germany, and France (Dreiländereck), this is often described as a historical landmark for trade and culture. Its location brings out cross-border cultural exchange and makes it a major transportation hub for Europe.

Basel also provides a series of large cultural festivals such CULTURESCAPES. This festival highlights the cultural landscapes of different regions worldwide. By opening dialogue and mutual learning about culture, it reinforces the concept of Basel as a meeting point for global cultures. ("About CULTURESCAPES", n.d.)

With this understanding of Basel's cultural life, it is possible to consider the existence and relevance of Latin American communities.

#### **2.4.2 Relevance to Latino Communities**

Latin American communities can take advantage of Basel's multicultural openness to showcase their heritage through festivals, art exhibitions, music performances (e.g., Latin jazz), or culinary events. By residing in a multi-culturally open society, these communities have the opportunity to share and enrich themselves from other similar communities.

Organizations like Nosotras, Olla Común from K5 Basler Kurszentrum or other cultural initiatives provide platforms for Latino residents to integrate into the cultural landscape while maintaining their unique identities.

### **2.5 Research Gaps**

Research Gaps go from understanding how throughout history, Latin American migrants have experienced the influence of living a new country forcing themselves to switch or sometimes hide their identity, up to comprehending other reasons as for potential lack of detailed information of their lives in Switzerland.

#### **2.5.1 Implications of Becoming an Immigrant**

Hugo, 2025, describes how the concept of "diaspora" now involves a series of contextual definitions. The word "diaspora" origins from the Greek word for "colonize", however

now it is considered a word for referencing a large group of people linked by common ethno-linguistic and/or religious bonds who have left their homeland. However this word is highly related to those usually leave their origin place because of political or economic reasons.

These communities will sometimes be excluded or intentionally exclude themselves from the new society they live in. Rationale for this, often includes illegal immigration, in which these groups voluntarily do not participate in data recompilation techniques such as national censuses.

Hugo also explains the lack of detailed documentation regarding the amount of immigrants found around the world using census data as an example:

- Countries that conduct censuses may intentionally exclude people not considered citizens or permanent residents.
- Some expatriates have not got full working rights and avoid being present in an official census.
- Some immigrants can perceive being part of the censuses as unnecessary.
- Censuses may not be able to identify all immigrants if their questions are not specific enough. For example, only asking about birthplace or nationality.
- Undocumented migrants often avoid inclusion of the census because they consider it a risk for their stay in the country.
- Censuses often exclude second and later generations of migrants living in the country.

With this, we can understand how immigrants perceive and are influenced to become or not part of official records, increasing the potential for research gaps.

### **2.5.2 Limited Historical Documentation**

The lack of archival records specifically focused on Latin American communities in Basel is a significant challenge for the project. While Basel's State Archive hold a large collection of historical documents, these mostly focus on other demographic and administrative records without explicitly highlighting the existence of Latin American communities.

Potential gaps can include the Immigrant Dossiers, which even if they contain archives up to 500,00 immigration documents from the 20th century, may lack detailed information or analysis of Latino migrants as a distinct group. (swissinfo.ch, 2023)

Other potential gap is community representation, in which often larger groups are easily represented, leaving other communities like Latinos underrepresented in official documentations. For example, the Alevi community. (“Alevi Communities In Western Europe”, 2011) It is important to state that this should not be considered a negative thing; however, under representation of communities can increase cultural bias. (zhmurko\_native\_2023)

Finally, the third identified gap is the lack of official documentation of cultural contributions like music and food festivals, businesses, and artistic activities. Up to the time this proposal was written, no official state archive has been found.

This project focuses primarily on working with local Latin American organizations to access personal archives, such as photographs, letters, official registries, or diaries, extract data from the State Archive, and finally social media.

### 2.5.3 Community Organizations

As mentioned above, there is limited research on the formation, evolution, and impact of Latin American community organizations in Basel. Existing insights suggest irregular efforts to unite Spanish-speaking individuals, but lack comprehensive historical analysis such as early initiatives. The Basel Spanish Catholic Mission Church<sup>5</sup>, represents one of the first institutional efforts to bring together Spanish-speaking residents. However, their history and impact on society remain unexplored and inaccessible.

Groups and small businesses such as Andina Latin Store and La Tienda Latina reflect ongoing efforts to unite immigrants among their communities by organizing small food festivals on weekends. Although these initiatives focus on socializing and preserve local, traditional recipes, they are not officially documented as part of Basel’s multicultural history.

This leads us to the understanding that unlike other immigrant communities with well established organizations such as the Turkish Alevi Association, Latin American groups lack centralized documentation or archives detailing their activities.

To address this gap, this project proposes conducting interviews with community participants to trace organizational histories, as well as partnering with cultural centers

---

<sup>5</sup>Misión Católica de Lengua Española de Basilea

such as Nosotras or La Tienda Latina to gather information about events and initiatives led by Latin American groups in Basel.

## 3 Conceptual Framework

### 3.1 Operational Definitions

To continue with this project, it is important to define several definitions that have been and will be used throughout the proposal for better understanding.

1. **Latino Communities:** Groups of people in Basel that identify with or trace their heritage to Latin American countries included those that are Spanish and Non-Spanish-speaking nations in Northern, Central and South America as well as the Caribbean. This definition includes first-generation immigrants and their descendants. (Jacquez et al., 2024)
2. **Archival Documentation:** The process of identifying, collecting, preserving, and providing access to primary source materials. In this research, materials can be described as personal papers, organizational records, photographs, and oral histories. These files document the history, activities, and contributions of the Latino community in Basel.
3. **Database:** Organized collection of information, usually with one central topic. A database is composed of records in tables. A record contains information that has been collected by one individual or entity in the database. A table holds the records that created, and the database encompasses the tables. (Derclaye, 2002)
4. **Corpus:** A large collection of written or spoken texts that can be developed manually or electronically and be representative of the authentic language data. The corpus has numerous applications in the field of applied linguistics such as development of dictionaries, construction of grammar books, enhancement of language learning and teaching, study of dialects, machine translation, etc. (Al-Gamal and Mohammed Ali, n.d.)
5. **Metadata:** A formal data documentation, including databases. The metadata record involves a set of information fields, which capture the characteristics of data as information resources. This answers the question of who, what, when, why, and about a resource. (“Metadata”, n.d.)

6. **Community Organization:** Classified as a formal or informal group established by members with Latino heritage in Basel with the aim of promoting cultural, social, economic, political and migratory advice and interests. These groups can include religious affiliations, cultural associations, social clubs, small businesses, and advocacy groups.
7. **Migrant Integration:** A two-way process involving the adaptation of migrants and the host society, usually based on principles like respect, tolerance, and non-discrimination. Integration includes the participation of migrants into the public life, acquisition of new knowledge and skills, and access to rights and services within the new community in which they reside. (“Migrant integration”, 2020)
8. **Cultural Pluralism:** A society framework in which individual ethnic groups have the right to maintain their unique cultural identities while coexisting within a larger society. Compared to other theories like assimilation, cultural pluralism values the preservation of cultural heritage alongside the participation in a new community. (Haas, n.d.)
9. **Social Connection:** The distinguished ways in which Latino immigrants experience and establish relationships, belonging, and support within both their own community and Basel’s society. This connection includes interpersonal support, community belonging, and social acceptance. (Jacquez et al., 2024)
10. **Digitization:** The process of turning physical archival materials such as documents, photographs, audio recordings, organizational registries, etc., into digital formats to ensure their preservation, easy access, and enable online publication and research use. (Roche, n.d.)
11. **Vector Embedding:** Numerical representations of complex data such as text, images, and audio, have become foundational in machine learning by encoding semantic relationships in high-dimensional spaces. (Pajo, 2025)
12. **Oral History:** A qualitative research method that involves recorded interviews with individuals from the Latino community to capture personal stories, experiences, and perspectives that might not be documented in written form. (Roche, n.d.)
13. **Primary Source:** Original material created or experienced by the individuals or organizations contacted during the period being studied, such as letters, photographs, records, and official government documents, which serve as direct evidence of historical events or community life. (Roche, n.d.)
14. **Workflow:** A set of tasks grouped chronologically into processes that are necessary

to accomplish a given goal. An organizational workflow is the set of processes needed to accomplish, the set of people or other resources available to perform those processes, and the interactions among them. (Cain and Haque, [n.d.](#))

15. **Artificial Intelligence:** AI is the branch of computer science, which makes the computers mimic human behavior to assist humans for better performance in the fields of science and technology. (Ghosh and Arunachalam, [n.d.](#))
16. **Machine Learning:** Machine learning describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks. (Janiesch et al., [2021](#))
17. **Generative Pre-trained Transformer or GPT:** Generative Pre-Trained transformers are a type of Large Language Models that use deep learning to produce natural language texts based on a given input. (Kivindy, [2023](#))
18. **Fine-Tuning:** The process of adapting a pre-trained model for specific tasks or use cases. It is subset of the another technique called transfer learning: the practice of leveraging knowledge an existing model has already learned as the starting point for learning new tasks. (Bergman, [2024](#))
19. **Retrieval Augmented Generation or RAG:** An architecture for optimizing the performance of an artificial intelligence model by connecting it with external knowledge databases. RAG helps large language models deliver more relevant responses at a higher quality. (Belcic, [2024](#))
20. **Container:** A standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. (“What is a Container?”, [n.d.](#))
21. **Cluster:** A cluster in docker refers to multiple nodes joined using swarm mode. Containers are scheduled across the various nodes, and networking is configured with overlay networking to look similar to bridge networks to the containers, but across multiple nodes. (“Swarm mode”, [800](#))
22. **Deployment:** Model deployment involves placing a machine learning model into a production environment. Moving a model from development into production makes it available to end users, software developers and other software applications and artificial intelligence systems. (Caballar, [2025](#))

## 4 Technological Frameworks in Cultural Preservation

The preservation of cultural heritage has been modified by the integration of various technological innovations, creating new ways to document, protect, analyze, and share history. From basic digitization to applications of Artificial Intelligence, also known as AI, technology has become a crucial component of modern preservation efforts, offering solutions to long-lasting challenges while also introducing new methodologies for preserving cultural legacy. (Skublewska-Paszkowska et al., 2022)

Understanding the technological advancements in cultural preservation will provide a broader idea on how these new techniques can be applied to preserve not only traditional culture as artwork, but to also find possibilities for preservation of informal or niche Cultural Heritage such as diaries, letters and personal photography.

### 4.1 Digital Documentation Methodologies

Digital cultural heritage preservation incorporates a systematic approach divided in three different stages:

1. Internal storage
2. Network sharing
3. Content interaction

The methodologies of this project are primarily based on three components:

1. Digital documentation
2. Research management
3. Presentation, visualization and/or interpretation of Data

Cultural Heritage comes in a widespread of forms, for example, digital documentation relies on several specialized recording techniques that have transformed how cultural artifacts and sites are preserved. These include examples like remote sensing technologies that enable data collection without having direct contact, photogrammetry and image modeling that create detailed visual records, 3D laser scanning for precise dimensional

documentation, infrared and multi-spectral imaging for revealing hidden details, and underwater exploration technologies for documenting submerged heritage sites. These technologies allow people who advocate preservation to capture comprehensive data without physical contact with artifacts or structures. (Baraka, [n.d.](#))

The management aspect of digital cultural heritage involves technology for disaster prevention, monitoring, and protection systems, computer simulation capabilities, and sophisticated database analysis through information management platforms. These systems enable cultural heritage documentation but also support protection and preservation. (Vaštakas et al., [n.d.](#))

Digital preservation technologies transform raw data into accessible formats through web and interactive content creation, 3D modeling with processing and animation capabilities, light projection and holographic displays, immersive experiences using virtual or augmented reality, and text or image generation. These preservation technologies democratize access to cultural heritage that might otherwise be inaccessible or unknown due to geographical limitations, fragility concerns, or damage caused over time. (Sound, [2018](#))

## 4.2 AI-Driven Preservation Strategies

With the integration of Artificial Intelligence, Cultural Heritage preservation is evolving, offering innovative solutions for tangible artifacts and intangible traditions. This project examines AI's diverse applications in heritage conservation and evaluates the potential of Large Language Models (LLMs) as tools for preserving informal cultural heritage, such as oral histories, community narratives, and non-institutional historical records. (Chakrabarti, [2024](#))

### 4.2.1 Digitization and Reconstruction of Physical Artifacts

AI's computer vision systems allow documentation and restoration of cultural artifacts. For example, algorithms trained on datasets of historical texts can reconstruct damaged sections of manuscripts by predicting missing characters or phrases based on contextual patterns. Similarly, 3D modeling techniques combined with photogrammetry create digital replicas of artifacts, allowing detailed analysis without physical handling. (Chakrabarti, [2024](#))

Machine learning models, for example, enhance image resolution and color accuracy in digitized artworks, revealing faded details that perhaps are invisible to the human eye.

Examples like the Smithsonian Institution use AI to transcribe handwritten documents, demonstrating how old damaged texts can become legible for researchers. (Richardson, 2024)

For fragile objects like ancient pottery or textiles, robots powered by AI systems can replicate traditional craftsmanship techniques, preserving knowledge that might otherwise be lost. (SWR Landesschau Rheinland-Pfalz, 2023) For example, the artisanal work learned from older generations that have a high potential of being lost because older artisans retire or newer generations that loose interest in learning.

#### 4.2.2 Natural Language Processing for Textual Heritage

Natural Language Processing or NLP algorithms serve as tools useful for analyzing and preserving linguistic heritage. By processing and cleaning<sup>6</sup> ancient scripts, dialects, and oral traditions, NLP can:

- Translate old dialects into modern spoken languages
- Identify semantic patterns in folklore or religious texts. (Maaithili and Phil, n.d.)
- Generate metadata from unstructured archival materials. (“Metadata”, n.d.)

For example, GPT-5 has been used to improve searchability in cultural databases by interpreting ambiguous terms in historical contexts. A 2024 study showed that LLM-enhanced search systems increased the discoverability of niche artifacts by 40 percent compared to traditional keyword-based methods. (“Metadata”, n.d.)

#### 4.2.3 Predictive Conservation and Risk Mitigation

AI’s predictive abilities help keep heritage sites safer from environmental and human threats. For example, machine learning models analyze satellite imagery to prevent risks such as:

- Coastal erosion threatening archaeological sites
- Urban encroachment on historical districts
- Climate-induced deterioration or organic materials

---

<sup>6</sup>Cleaning data refers to the process of eliminating human readable exclamation or question marks, unifying texts by keeping all characters in lowercase form, and eliminating verb tenses that are only human understandable.

UNESCO uses these systems to monitor World Heritage Sites, enabling preventive conservation measures. AI-driven surveillance also detects looting activities by analyzing patterns in satellite data and social media feeds. (Chakrabarti, 2024)

#### 4.2.4 Virtual and Augmented Reality Experiences

In addition, computer vision and generative AI combine as a tool to create immersive heritage experiences. For example, the REALM project uses Augmented Reality to overlay historical context onto physical ruins, while Virtual Reality reconstructions of destroyed sites such as Palmyra allow global audiences to explore lost treasures. (Baraka, n.d.) Another example found in Switzerland, is the AR Augusta Raurica Experience where Augmented Reality is combined with storytelling, playing a fictional story about a family living in the currents ruins found in Pratteln. (“Augusta Raurica”, n.d.)

With the emergence of OpenAI and ChatGPT, to this day, it is now possible to find any style, type, themed or custom GPT model. Custom GPT models can potentially create interactive guides that adapt narratives based on user requests or queries, making cultural education more engaging for the everyday user. For example, asking ChatGPT to explain to you the history of Pinochet’s dictatorship in Chile as a historian or someone who lived during those years may output different perspectives, making it not only easier to understand, but also more attractive leaving the possibility to learn about it in different ways.

### 4.3 GPT Models for Informal Cultural Heritage Preservation

Informal cultural heritage like oral histories, family archives, community rituals, etc. create unique preservation challenges due to its decentralized nature and big reliance on human memory. GPT architectures, for example, can offer solutions to capture and sustain these momentary traditions.

#### 4.3.1 Oral History Processing and Analysis

GPT models are able to transcribe, translate, and contextualize spoken and written narratives. Some significant applications include:

- **Automatic Interview Summarization:** Reducing hours of recordings to structured transcripts while preserving linguistic variations.

- **Cross-Generational Knowledge Transfer:** Converting everyday speech into standardized formats without losing cultural specificity.
- **Sentiment Analysis:** Identifying emotional undertones in personal narratives that reveal community values.

For example, the UCLA Center for Oral History Research experiments with GPT-4 to analyze interviews about marginalized communities, revealing hidden patterns in migration stories and labor practices. This project provides evidence how Large Language Models are valuable tools in the Humanities and Social Sciences research, setting a precedent for this project. (“AI on How New and Evolving Technologies Will Impact Professions”, 2024)

### 4.3.2 Community-Driven Archiving

GPT-powered platforms enable decentralized contributions to heritage databases, such as:

- Collaborative Story Archival: Community members share voice memos or texts about local traditions, whether it is blogs or social media, which AI interprets into searchable archives.
- Contextual Enrichment: GPTs add metadata to content submitted by users.
- Bias Mitigation: Algorithms can flag culturally insensitive interpretations during public contributions.

For example, Japan’s AI-assisted preservation of textile weaving techniques demonstrates how GPT models can distill artisan interviews into instructional guides, making it possible to preserve the crafts survival despite its declining amount of practitioners. (Yuri, n.d.)<sup>7</sup>

### 4.3.3 Ethical and Technical Challenges

While still promising, GPT implementations require the careful consideration of:

- Consent and ownership: Ensuring communities still have control over their narratives used to train models.

---

<sup>7</sup>Patterns and weaving techniques reproduced but this system are currently used by the textile industry to create Kimono designs.

- Cultural authenticity: Preventing AI from homogenizing regional dialects, traditions or country stereotypes.
- Data security: Protecting sensitive oral histories from unauthorized commercial use.

As a case study, during the 2024 AI in Oral History Symposium the risks of using AI-generated "totalized narratives" overwriting individual perspectives provided accurate information on how when working with AI and GPTs, raw and unprocessed interviews should be kept as "vaulted" archives. (“AI on How New and Evolving Technologies Will Impact Professions”, 2024)

## 5 Source Description

This chapter aims to establish the credibility and provenance of the data used for the proposal. by providing context for interpretation, this chapter hopes to help clarify the scope, strenghts, and limitations of the project. This chapter also aims to explain how sources will be selected, processed, transformed, and protected, making technical steps more understandable and replicable.

The research will draw its data from the following sources:

- Open-Source datasets, archives, blog posts, and social media posts.
- In-person interviews with local communities around Switzerland, prioritizing communities found in Basel.
- Personal documents, letters, diaries, and government records.

### 5.1 Source Guidelines

To gather information from the sources mentioned above, it will be necessary to follow a series of ethical and operational guidelines to ensure the sources are appropriate for the project. This will involve:

#### 5.1.1 Consent and Age Restrictions

All participants should be provided a detail consent form at the time of interviewing, or before providing any original documentation. In this consent form is also important to establish how the data will be used, stored, and shared. following FAIR DATA principles

(Findable, Accessible, Interoperable, and Reusable Data). The forms provided should also be available in the different languages considered for the web application, for example, English, Spanish, German, French, and Portuguese, which are the most common languages used in Latin American immigrant communities in Basel.

The minimum age limit for participation will be 18 years old. In case some of the data includes information regarding minors, consent for legal guardians or representatives should be authorized. For technologies and cases where this is not consent, data coming from minors should be deleted or discarded correctly.

Each participant has the right to choose whether or not they would like their personal information mentioned or stored in the official corpus. Withdrawal of consent can be also performed at any time of the data recompilation.

### **5.1.2 Use of Aliases or Anonymization**

As described above, participants will have the right to use aliases or pseudonyms during the process of interviews, transcripts, or published databases to protect their privacy.

To anonymize and identify the provenance of each data source, an additional ID number or personal identification number (PIN) and a geographic tag can be assigned. For data including, for example, other family members, community members or additional mentions during the interviews or found in the written histories, anonymization would be considered prior to requesting consent.

### **5.1.3 Data Sensitivity and Accessing**

All official records like dossiers or census data, should be treated in accordance of the governmental restrictions and institutional policies.

Access to sensitive materials should also be limited to those in charge of processing the metadata for the creation of the database.

### **5.1.4 Cultural Sensitivity and Community Engagement**

All communities and families involved in the contribution for the databases should be aware and acknowledge their rights and ownership of their data.

The processes of metadata extraction and tagging, and natural language preprocessing for the final corpus should follow a rigorous technical process to avoid misinterpretation or loss of narratives or cultural heritage.

Communities have the right to engage during data repossession to ensure that their content is treated with their values and needs.

### **5.1.5 Data Quality and Attribution**

Provenance of the official records, community documents and open source social media data should be verified or authenticated before being added to the corpus. This will prevent biases and misinformation found in the database and fine tuning process.

Oral histories and personal archives should be attributed properly by human or machine tagging the correct authors and provenance of the data. Summaries can also include specific data regarding its authenticity.

If the data already includes metadata, this should be kept intact and just be addition to the official JSON<sup>8</sup> files for the final corpus. These includes keeping records like dates, conditions of use, and ethical restrictions if applicable.

### **5.1.6 Handling of Sensitive and Historical Content**

Data regarding historical events, personal documents, and those that might contain politically or socially sensitive information should be handled with caution, following guidelines established by universities or governments.

### **5.1.7 Compliance with Legal and Ethical Frameworks**

All data should be following national and international data protection laws, such as the GDPR, specifically data collected in Switzerland and the European Union. (European Parliament. Directorate General for Internal Policies of the Union., [2017](#) )

Proper licensing and acknowledgments for academic and media sources should also be included in the official documentation of the project, as well as written in the consent forms for interviews.

Detailed and organized records of consent forms, data licenses, and ethical reviews should be written and saved in the official deployment documentation.

### **5.1.8 Continuous Consent and Review**

Policies for periodic data review and re-consent should be written and sent during and after the deployment of the web application. These policies should include key points

---

<sup>8</sup>JSON files are text-based open standard formats used for representing structured data based on JavaScript syntax

regarding long-term usage and data storage.

Participants and communities will be allowed to review how their data is being used and provide feedback or request changes or removal even after deployment.

## 6 Methodologies

This project will employ a qualitative and theoretical methodology, emphasizing in critical synthesis, conceptual argumentation, and the articulation of a practical, and innovative solution grounded in the gaps and challenges identified through theoretical exploration.

### 6.1 Archival Research Methods

As a collection of unique, single documents, created contemporaneously with the events they discuss, the materials lodged in archival repositories provide a particular window on the geography of earlier times. (Roche, [n.d.](#))

To develop a domain-specific GPT, this project requires a methodological approach that ensures accurate, representative, and well-structured training data. The proposed archival research method is a hybrid approach that combines traditional archival practices with digital processing and metadata enrichment, tailored for AI model training and fine-tuning.

The initial phase involves the identification, collection, and digitization of relevant archival sources that come in multiple forms: written letters, transcript interviews, diaries, etc. This project will employ rigorous selection criteria, prioritizing materials such as the sources mentioned above. These materials should be digitized at high resolution to preserve fidelity. Afterwards, Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) technologies are employed to convert physical documents into machine-readable text, therefore facilitating processing and analysis. (Patton, [2024](#))

Following the digitization process, the organization of metadata constitutes the next step. Applying metadata schemas will enable consistent cataloging of key attributes, that can include creators, dates, subjects, and contextual annotations. This metadata is meant to enhance easy discovery and retrieval of archival items but to also provide semantic context for the training corpus. Furthermore, domain-specific annotations such as thematic content, linguistic features, and entity relationships should be also incorporated. (“Data Archive for AI | Significance, Benefits, and Use cases”, [2024](#))

Data processing should also maintain the integrity and utility of the dataset. This process will incorporate the consolidation, correction of OCR errors, and resolution of inconsistencies. Once the corpus is cleaned, should be formatted as prompt completion pairs or other structured inputs to suit the requirements of GPT fine-tuning. For fine-tuning a model, several prompt completion templates exist; for this project the most suitable prompt completion template is **alpaca**.

Example of the alpaca prompt completion:

```
{
  "description": "Template used by Alpaca-LoRA.",
  "prompt_input": "Below is an instruction that describes a task, paired
  → with an input that provides further context. Write a response that
  → appropriately completes the request.\n\n###
  → Instruction:\n{instruction}\n\n### Input:\n{input}\n\n###
  → Response:\n",
  "prompt_no_input": "Below is an instruction that describes a task. Write
  → a response that appropriately completes the request.\n\n###
  → Instruction:\n{instruction}\n\n### Response:\n",
  "response_split": "### Response:"
}
```

More advanced strategies will involve the construction of knowledge graphs and semantic networks. With the use of Natural Language Processing (NLP) techniques to extract entities and their relationships from the corpus, this will enable the creation of structured knowledge representations. These semantic frameworks are intended to enrich the contextual understanding of the model and facilitate more exact information retrieval and generation. (Z. Zhang et al., 2024)

Finally, this proposal aims to synthesize traditional archival practices with digital innovations to provide a most robust foundation for developing a domain-specific GPT that effectively serves both academic research and public engagement with the history of Latin American communities.

## 6.2 Workflow Establishment

The following workflow outlines a rigorous, methodological approach for conducting archival research on Latin American communities and developing a domain-specific GPT model. This workflow is designed to ensure academic guidelines, transparency, and reproducibility throughout all stages of the project.

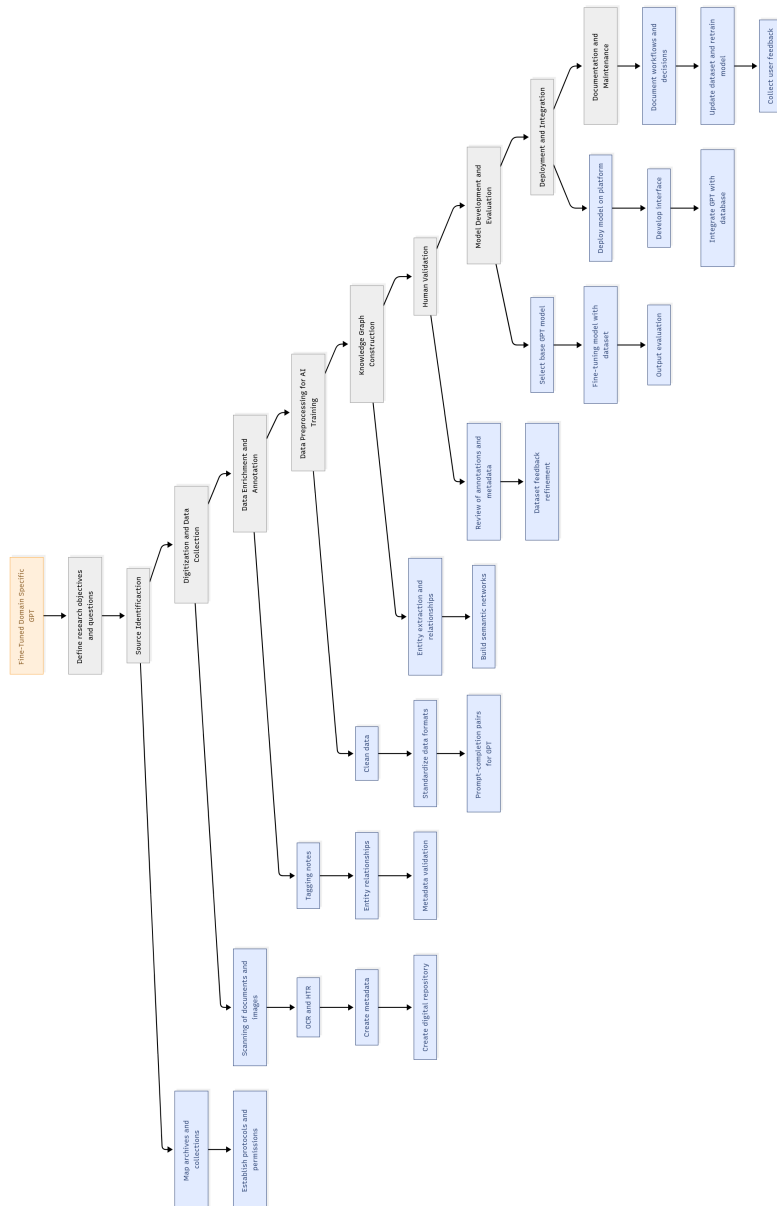


Figure 1: Mermaid Flowchart

## 7 Domain-Specific GPT Proposal

### 7.1 Theoretical Proposal: GPT-Based Preservation Tool

This section aims to outline the conceptual framework for manipulating a GPT model as a digital preservation and access tool for the archival record of Latin American communities in Switzerland. The proposal imagines a system that not only stores and organizes digitized archival materials but also enables advanced search, contextualization, and interpretation through natural language processing.

### 7.1.1 Objectives

1. Facilitate accessible, rich exploration of archival content for researchers, community members, and the public.
2. Support preservation by enabling the use of digital archival tools and metadata tagging enrichment.
3. Enhance discovery, education and interpretation through AI-driven summarization, question answering, and content relationships.

## 7.2 Rationale for GPT Approach

Sohail et al., 2023 claim at the moment of their study, that technologies like GPTs have not only captured the attention of the research community, but also claimed the attention of daily life search engine users. These models have demonstrated a large capability of natural language understatement and content generation. With the development of GPT chatbots, people can now engage in human style conversations and extract and/or access information considered accurate and fast.

With new natural language interactions, GPT models allow users to query data conversationally, lowering the barriers for unexperienced and multilingual users. Comparison of query methods between GPTs and traditional search engines demonstrates how natural language and semantics are applied differently. Traditional search engines are based on keyword and semantic understanding based search, as where GPT models use advanced natural language and user intention. (Kerner, n.d.)

Another key advantage of using GPT-based archival tools is their capacity for contextual understanding. Models can synthesize information from a wide range of sources. This ability surpasses what traditional search engines often do, which typically involves retrieving information in isolated pieces rather than integrating narratives or analytical contexts.

Other significant benefit is scalability. Once the model has been trained or fine-tuned with the desired corpus, it can process and analyze large volumes of text. This makes it achievable to manage and browse digital archives, accommodating a continuous growth of digitized materials without a large increase in manual labor or use of computational resources.

Furthermore, Large Language Models are considered to know how to do full knowledge

integration. By linking different formats of data, such as oral histories, photographs, and archival documents, GPTs can construct a richer, more interconnected narrative. This not only uncovers previously hidden connections between the data found in the corpus but also provides users with a deeper and more exact understanding of historical and cultural records.

GPTs have the quality to tailor responses based on the conversation history, as for traditional search engines which only can only personalize responses based on user data and search history. (Kerner, [n.d.](#))

<b>Aspect</b>	<b>GPT models</b>	<b>Traditional Search Engines</b>
Responses	Direct with conversational responses	List of links
Content Generation	Original content	Only retrieves existing information
Context Awareness	Maintain context throughout conversation	Limited context awareness. Each query treated separately
Synthesizing Capabilities	Combine multiple sources	Separate results from different sources
Update frequency	Incorporates new information	Depends on web crawling and indexing cycles
Personalization	Tailored responses based on conversation	Personalize based on user data and search history

### 7.3 Data Sources and Preparation

The data collection proposed for the official corpus can be constructed from diverse sources, offering a wider range of perspectives and forms of documentation.

Official records can provide a good foundation for historical research. Even if official records do not contain specific data on Latin American communities, they provide historical context in migration patterns. These include immigration dossiers, municipal documents, and census data housed in state archives. Such materials besides offering good insights into migrational patterns, also provide demographic changes, and administrative recognition of their existence over time.

Community materials add another dimension by capturing information about collec-

tive activities and organizational lives of Latin American groups. Newsletters, event fliers, meeting minutes, and other similar sources produced by these organizations and/or families document cultural events, advocacy efforts, and the daily life of community networks. These sources are valuable for understanding how these communities have organized, celebrated, and advocated for their culture and traditions outside of their native country.

Oral histories, also known as micro-history, are also essential for preserving the voices and experiences of individuals who might be absent from official documentation or personally decide to exclude themselves from them for security reasons. Transcribed interviews with community members, leaders, and their descendants provide personal narratives that improve personal migration journeys, adaptation and integration processes, and the transmission of cultural identity with younger generations.

Personal archives, such as letters, diaries, photographs, and memorabilia contributed by families, offer a more complex insight into the personal lives and relationships of migrants. These artifacts help reconstruct the realities of Latin American residents, highlighting their resilience, adaptation, integration to Swiss lifestyles and cultural preservation.

Finally, academic and social media sources, such as articles, local news reports, blogs, and exhibition catalogs can help contextualize the community's experiences making social, historical, and cultural frameworks wider and more distinct. These materials not only interpret primary sources but also intend to help position the presence of Latin American groups within national and transnational records.

Together, these sources allow for a more comprehensive and precise reconstruction of the history and contributions of Latin American communities in Switzerland.

The data preparation process for this project begins with obtaining the data and then digitizing it. Physical materials such as the documents, photographs, and handwritten records described on the sources, will be scanned at high resolution to preserve details and fidelity. For text-based sources, Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) technologies will be applied, converting printed and handwritten content into machine-readable text. This step will make the archival materials accessible for computational processing and analysis.

Following the digitization process, metadata extraction and enhancement will be performed. Each item will be tagged to capture relevant information, including the creator, dates, subjects, and contextual details. This process will be developed by humans, using machine assistance for correction and error identifying. This way of structuring the meta-

data will make it more efficient for retrieval and organization for the semantic context necessary for further search and interpretation inside the digital archive.

Annotation also improves the dataset by adding layers of meaning and connection. This process involves identifying themes, recognizing people, places, and organizations mentioned in the materials, and mapping relationships among them. This annotation increases the contextual value of the dataset and supports a more sophisticated AI analysis.

Data cleaning will involve correcting errors that occur during OCR processing, removing duplicate entries, and standardizing the formats across the dataset. This process aims to increase quality and reliability, preventing potential hallucination of model, bias and noise.

Finally, structuring the data for AI training also involves formatting the material into prompt completion pairs or other structured entries suitable for fine-tuning the model. The formats follow the previously mentioned alpaca template and will be exported in JSON and CSV formats. These formats are widely used and supported for FAIR data approaches and will guarantee the models can easily interpret the data used for fine-tuning.

Example for data preparation using Python libraries with outputs in JSON and CSV files:

#### 1. Sample Archival Data Structure: Metadata records

- Title
- Creator
- Date
- Subject
- Text (main content)
- Language
- Tags (thematic keywords)

#### 2. Data cleaning and preprocessing using Python:

```
import pandas as pd
import json

# Sample records
```

```

records = [{ "id": "001",
"title": "Immigration Dossier of Carlos Reyes","creator": "Basel
↳ State Archive","date": "1965-04-15","subject":
↳ "Immigration","text": "Carlos Reyes arrived in Basel in
↳ 1965...","language": "Spanish","tags": ["immigration", "Latino
↳ community", "Basel"]}]

# Convert to DataFrame
df = pd.DataFrame(records)

# Standardize date format
df['date'] = pd.to_datetime(df['date'])

# Add year column
df['year'] = df['date'].dt.year

# Filter by languages
df_languages = df[df['languages'] == 'es','en','de']

# Prepare prompt-completion pairs for GPT fine-tuning
fine_tuning_data = []
for _, row in df_languages.iterrows():
    fine_tuning_data.append({
        "prompt": f>Title: {row['title']}\nSubject:{row['subject']}_
↳ \nDate:{row['date'].strftime('%Y-%m-%d')}\nTags:
↳ {'.'.join(row['tags'])}\nText:", "completion": f"
↳ {row['text']}"
    })

# Convert to JSONL (one JSON object per line)
jsonl_str = '\n'.join(json.dumps(item) for item in fine_tuning_data)

# Save to CSV
csv_str = df_languages.to_csv(index=False)

```

### 3. Example python output:

```

{"prompt": "Title: Immigration Dossier of Carlos Reyes\nSubject:
↳ Immigration\nDate: 15-04-1965\nTags: immigration, Latino
↳ community, Basel\nText:", "completion": " Carlos Reyes arrived
↳ in Basel in 1965..."}
{"prompt": "Title: Community Newsletter April 1970\nSubject:
↳ Community Events\nDate: 01-04-1970\nTags: community, festival,
↳ culture\nText:", "completion": " The Latino Cultural Association
↳ held a festival..."}
{"prompt": "Title: Oral History Interview with Maria Lopez\nSubject:
↳ Oral History\nDate: 20-06-2010\nTags: oral history, migration,
↳ personal narrative\nText:", "completion": " Maria Lopez recounts
↳ her migration journey..."}

```

#### 4. Example CSV output:

id	title	creator	date	subject	text	language	tags
001	Immigration Dossier of Carlos Reyes	Basel State Archive	15.04.1965	Immigration	Carlos Reyes arrived in Basel in 1965...	es, en, de	['immigration', 'Latino community', 'Basel']
002	Community Newsletter April 1970	Latino Cultural Association	01.04.1970	Community Events	The Latino Cultural Association held a festival...	es, en	['community', 'festival', 'culture']
003	Oral History Interview with Maria Lopez	Oral History Project	20.06.2010	Oral History	Maria Lopez recounts her migration journey...	es, en, de	['oral history', 'migration', 'personal narrative']

To manage and transform the data, `pandas` and `json` libraries from Python will be employed. These tools allow for fast processing of large datasets, create a usable format, and facilitate effective data organization and cleaning. These libraries make it possible to automate the transformation of raw data into organized, consistent data suitable for machine learning applications.

Once processed, the data will be exported into output formats specific for the computational needs for fine-tuning and stored in public domains, for example, Zenodo.org. JSONL files will be generated for AI models, providing a structure that supports alpaca prompt completion pairs for training the model. At the same time, CSV files will be produced for tabular analysis, and project documentation, making it easier for quantitative evaluation and traditional archival research.

This workflow ensures that the archival data are both readable by humans and machine usable.

## 7.4 Fine-Tuning an AI Model

Structuring data for AI fine-tuning involves formatting archival materials into prompt completion pairs or other structured entries that the model can learn from effectively. This process is essential to ensure that the model understands the task, context, and desired output style, especially when adapting a general-purpose GPT to a specialized domain.

Its important to include the following aspects:

1. Prompt Formatting: Each training example is structured as a prompt or input paired with a completion or output.<sup>9</sup> For each archival data, the prompt might include metadata fields followed by a tag for the model to generate either a summary, answer a question, or provide contextual information. This form of structuring will guide the model to produce more coherent and accurate outputs based on the information provided.

```
records = [  
  {  
    "title": "Immigration Dossier of Carlos Reyes",  
    "date": "1965-04-15",  
    "subject": "Immigration",  
    "text": "Carlos Reyes arrived in Basel in 1965 and quickly  
    ↪ became involved in the local community."  
  },  
  {  
    "title": "Oral History Interview with Maria Lopez",  
    "date": "2010-06-20",  
    "subject": "Oral History",  
    "text": "Maria Lopez recounts her experiences migrating from  
    ↪ Peru to Switzerland in the early 1980s."  
  }  
]  
  
with open("archival_finetune_data.jsonl", "w", encoding="utf-8") as  
↪ f:  
  for rec in records:  
    prompt = (  
      f>Title: {rec['title']}\n"  
      f>Date: {rec['date']}\n"  
      f>Subject: {rec['subject']}\n"  
      f>Summarize this record:\n"  
    )  
    completion = rec["text"]  
    entry = {"prompt": prompt, "completion": " " + completion}  
    f.write(json.dumps(entry, ensure_ascii=False) + "\n")
```

Example JSONL line:

```
{"prompt": "Title: Immigration Dossier of Carlos Reyes\nDate:  
↪ 1965-04-15\nSubject: Immigration\nSummarize this record:\n",  
↪ "completion": " Carlos Reyes arrived in Basel in 1965 and  
↪ quickly became involved in the local community."}
```

---

<sup>9</sup>See example of alpaca prompt completion above.

2. Diversity and Representativeness: The dataset is intended to cover a wide range of document types, topics, languages, and styles to avoid bias and ensure that the model generalizes correctly across the corpus.
3. Data Quality and Cleaning: As mentioned before, the data must be cleaned to correct OCR errors, consolidation, and creating a standard format. Cleaning the data correctly will aid the model to perform greatly and reduce the possibility of hallucination.

#### 7.4.1 Fine-Tuning Techniques and Tools

To fine-tune LLMs efficiently and effectively, without using a lot of computational resources, several tools are proposed for this project:

1. LoRA (Low-Rank Adaptation): LoRA is a parameter efficient fine-tuning method that updates a small subset of parameters, lowering computational power and storage requirements. This makes it easier to fine-tune large GPT models without retraining an entire network. LoRA enables faster experimentation and iteration, which is useful when refining models on archival datasets. (Moez, [n.d.](#))
2. Axolotl Framework: Axolotl is an open source fine-tuning framework designed to simplify and accelerate the process of adapting LLMs. Axolotl supports LoRA and other parameter efficient tuning methods. It also provides utilities for data pre-processing (including prompt completion formatting), and integration with popular model architectures. Axolotl can be a useful tool to smooth and accelerate the workflow. (“Axolotl AI Cloud”, [2025](#))

#### 7.4.2 Practical Considerations for Fine-Tuning

1. Hyperparameter Tuning: Adjust learning rates, batch size, and number of epochs carefully to balance between underfitting and overfitting. Smaller learning rates are often preferred for fine-tuning to preserve pre-trained knowledge while adapting to the new domain. (Santopaolo, [2025](#))
2. Evaluation and Validation: Use a continuous validation set to monitor model performance during training. Metrics should evaluate not only language fluency but also accuracy and domain relevance. (“Strategies for Fine-Tuning Large Language Models”, [n.d.](#))

3. Combining Fine-Tuning with Prompt Engineering: Fine-tuning can be complemented by prompt engineering techniques, such as using short examples or instruction tuning, to further improve the model responses without extensive training. (Lester et al., 2021) Using Instruction Tuning helps to align the model's outputs with the desired responses. (S. Zhang et al., 2024)

By structuring the data and using efficient fine-tuning methods such as LoRA and Axolotl frameworks, the model will be able to understand and generate content related to Latin American community archives. By doing this, the model will be able to balance accuracy, efficiency, and scalability, allowing the deployment of AI tools in constrained environments.

### 7.4.3 Code Example to Fine-Tune with LoRA and Axolotl

The first proposed example code is directly fine-tuning the model using a `.yaml` configuration document and Axolotl.

1. Prepare the data into JSONL format:

```
{"prompt": "Title: Immigration Dossier of Carlos Reyes\nSubject:\n  Immigration\nDate: 1965-04-15\nText:", "completion": "\n  Carlos Reyes arrived in Basel in 1965 and..."}
```

2. Installation of Axolotl and other necessary dependencies: `pip install axolotl`
3. Confirmation of the right `PyTorch` and `transformers` are installed:

```
pip install torch transformers datasets
```

4. Example configuration file (config.yaml):

```
model:\n  # Pretrained base model name or path\n  pretrained_model_name_or_path: "gpt5"\n\ntrain:\n  # Path to your training data in JSONL format\n  dataset: "data.jsonl"\n  # Number of training epochs\n  epochs: 3
```

```

# Batch size per device
per_device_train_batch_size: 4
# Learning rate for LoRA fine-tuning
learning_rate: 3e-4
# Gradient accumulation steps to simulate larger batch size
gradient_accumulation_steps: 8
# Maximum sequence length for inputs
max_seq_length: 512
# Validation split (optional)
validation_split_percentage: 10

lora:
# Enable LoRA fine-tuning
enabled: true
# LoRA rank (low-rank adaptation dimension)
r: 8
# LoRA alpha (scaling factor)
alpha: 16
# LoRA dropout rate
dropout: 0.1
# Target modules to apply LoRA (depends on model architecture)
target_modules: ["c_attn"]

optimizer:
# Optimizer type
type: "AdamW"
# Weight decay
weight_decay: 0.01

logging:
# Log training progress every N steps
logging_steps: 50
# Output directory for checkpoints and logs
output_dir: "./lora_finetuned_model"

```

## 5. Running the Fine-Tune via the Operating Systems Terminal with Axolotl:

```
axolotl train --config config.yaml
```

When manually fine-tuning using LoRA, here is an example that includes the `peft` library from Hugging Face. This is an example using OpenAI's "gpt5".

```

from datasets import load_dataset
from transformers import AutoTokenizer, AutoModelForCausalLM, Trainer,
↪ TrainingArguments
from peft import LoraConfig, get_peft_model

```

```

# Load dataset
dataset = load_dataset('json', data_files='data.jsonl')

# Load tokenizer and model
model_name = "gpt2"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

# Tokenize function
def tokenize_function(examples):
    inputs = [ex["prompt"] + ex["completion"] for ex in examples]
    model_inputs = tokenizer(inputs, max_length=512, truncation=True)
    return model_inputs # 512 for smaller texts

tokenized_dataset = dataset.map(tokenize_function, batched=True)

# Configure LoRA
lora_config = LoraConfig(
    r=8,
    lora_alpha=16,
    target_modules=["c_attn"], # GPT-2 attention module
    lora_dropout=0.1,
    bias="none", # LoRA Parameters for Fine Tuning and hallucination
    ↪ avoidance
    task_type="CAUSAL_LM"
)

# Apply LoRA to model
model = get_peft_model(model, lora_config)

# Training arguments
training_args = TrainingArguments(
    output_dir="./lora_gpt2",
    per_device_train_batch_size=4,
    num_train_epochs=3,
    logging_steps=50,
    save_steps=200,
    evaluation_strategy="steps",
    save_total_limit=2,
    learning_rate=3e-4,
    weight_decay=0.01,
    gradient_accumulation_steps=8,
    fp16=True,
)

# Initialize Trainer
trainer = Trainer(
    model=model,

```

```

    args=training_args,
    train_dataset=tokenized_dataset["train"],
)

# Train
trainer.train()

```

Next, here is the suggested code for fine-tuning the proposed model for the project, Meta's Llama 3.

The following prerequisites are necessary to fine-tune Llama 3 with LoRA, accessing the model via Ollama.

1. Ollama provides the bases to deploy Llama 3 model in a local way, however it doesn't directly expose LoRA fine-tuning API's.
2. First it is necessary to download Llama 3's base weights compatible with Hugging Face's `transformers` library.
3. Using `peft` (Parameter-Efficient Fine-Tuning) or Axolotl can support LoRA for Llama 3.
4. After the fine-tuning process, the updated weights should be merged or guided to enable inference through Ollama or other pipelines.

This code runs outside Ollama but uses the same model weights Ollama relies on.

```

from transformers import AutoTokenizer, AutoModelForCausalLM, Trainer,
↳ TrainingArguments
from datasets import load_dataset
from peft import LoraConfig, get_peft_model

# Load tokenizer and base Llama 3 model compatible with Ollama's weights
model_name = "meta-llama/Llama-3-8b-hf" # Replace with actual HF repo
↳ or local path
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

# Configure LoRA
lora_config = LoraConfig(
    r=16,
    lora_alpha=16,
    target_modules=["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj",
↳ "up_proj", "down_proj"],

```

```

    lora_dropout=0.0,
    bias="none",
    task_type="CAUSAL_LM"
)

# Integrate LoRA into model
model = get_peft_model(model, lora_config)

# Load fine-tuning dataset - your prepared prompt-completion pairs in
↪ JSON or JSONL format
dataset = load_dataset("json", data_files={"train": "data/train.jsonl",
↪ "validation": "data/val.jsonl"})

# Tokenize input
def tokenize_fn(examples):
    return tokenizer(examples["prompt"], truncation=True,
↪ max_length=512, padding="max_length")

tokenized_dataset = dataset.map(tokenize_fn, batched=True)

# Define Trainer and Training Arguments
training_args = TrainingArguments(
    output_dir="./llama3-lora-finetuned",
    per_device_train_batch_size=4,
    num_train_epochs=3,
    learning_rate=3e-4,
    logging_steps=50,
    evaluation_strategy="steps",
    save_steps=200,
    weight_decay=0.01,
    gradient_accumulation_steps=8,
    fp16=True,
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["validation"]
)

# Train with LoRA fine-tuning
trainer.train()

# Save LoRA adapter weights separately
model.save_pretrained("./llama3-lora-adapter")

```

Merge LoRA's weights into the base model using **PEFT** utilities, then save a new model directory that Ollama can load.

```
merged_model = model.merge_and_unload()
# Merge PEFT adapters into base weights
merged_model.save_pretrained("./llama3-lora-merged")
```

Finally, here is the suggested code for fine-tuning Llama 3 using Axolotl:

```
# Model settings
base_model: 'NousResearch/Llama-3.2-1B' #Hugging Face Llama3 base model
model_type: LlamaForCausalLM
tokenizer_type: LlamaTokenizer
is_llama_derived_model: true

# Precision and Memory Options
load_in_8bit: false
load_in_4bit: true # 4-bit quantization for efficient memory usage
strict: false

# Dataset Configuration
dataset:
  -path: 'data/archival_data.jsonl' # Instruction tuning JSONL
  type: alpaca # Alpaca-style prompt-completion format
dataset_prepared_path: 'preprocessed_data'
val_set_size: 0.1 # 10% validation split

# Fine-Tuning Method and LoRA Config
adapter: lora
lora_model_dir: '' # Fresh training, no pre-existing adapter
lora_r: 16
lora_alpha: 32
lora_dropout: 0.05
lora_target_modules:
  - 'q_proj'
  - 'v_proj'
  - 'o_proj'

# Training Hyperparameters
gradient_accumulation_steps: 4
micro_batch_size: 2
num_epochs: 3
optimizer: adamw_bnb_8bit
lr_scheduler: cosine
learning_rate: 0.0002

# Training options
train_on_inputs: false
group_by_length: false
fp16: true
gradient_checkpoint: true
logging_steps: 10
```

```
# Output directory for the fine-tuned model  
output_dir: './llama3_lora_finetuned'
```

```
axolotl train llama3_lora_finetuned.yaml
```

Exporting and using the fine-tuned model with Ollama:

```
from peft import PeftModel  
from transformers import AutoModelForCausalLM  
  
base_model_name = "NousResearch/Llama-3.2-1B"  
lora_path = "./llama3_lora_finetuned"  
  
model = AutoModelForCausalLM.from_pretrained(base_model_name,  
↳ load_in_4bit=True)  
model = PeftModel.from_pretrained(model, lora_path)  
  
# Save merged model locally for Ollama  
model.save_pretrained("./llama3_ollama_ready")  
  
# Then convert/save in Ollama compatible format (GGUF) using Ollama CLI  
↳ or other tools  
  
# Run the model locally  
ollama add ./llama3_ollama_ready  
ollama run llama3_ollama_ready
```

## 7.5 Deployment and Integration

After the fine-tuning process, it is necessary to deploy and make the model publicly available and achieve FAIR data principles. For this project, several ways have been proposed. These include deploying the model via a Cloud Server, API access, developing an HTML User Interface, or by deploying the model via external source frameworks like Langflow, Zenodo or GitHub.

To effectively deploy and integrate the model, it is essential to follow a series of steps to transform the fine-tuned model into a practical tool. This will involve making it accessible to users, which can be achieved by cloud hosting it in environments like AWS, Azure, or Google Cloud that offer flexible work infrastructures.

### 7.5.1 Cloud Hosting and API Wrapping

Typically, developers "wrap" the model into a web API (Application Programming Interface) using frameworks like FastAPI or Flask. The API acts as a bridge, taking in queries from different front-end applications and responding with answers powered by the model. It is also possible to use containerization tools such as Docker to package the model environment; this will ensure that the performance between several platforms is consistent, making updating and scaling a more simple task. (Z. Ahmed, 2024)

Using FASTAPI, a web framework for building API in Python, provides several benefits for LLMs:

1. **Asynchronous Support:** Async programming helps handle multiple requests efficiently. This is a good tool for managing multiple simultaneous requests in LLMs. For example, using chatbots where multiple interactions are happening at the same time. With FASTAPI, each request is taken independently.
2. **Automatic Documentation:** FASTAPI generates interactive API documentation through Swagger and ReDoc. These tools help identify possible needs for testing and debugging endpoints interacting with the LLM.
3. **Performance:** FASTAPI helps minimize the time response for LLMs.
4. **Easy Integration:** FASTAPI integrates machine learning libraries, making it easier to deploy LLMs into different environments, many of which can also be accessed by HTTPS requests.

On the other hand, Flask is a lightweight web application framework. ("Flask Documentation (3.1.x)", n.d.) In an article published by Sekar, 2024, the author shows how to implement Flask with Llama3. To do so, the implementation performed by the author required 5 steps:

1. Install Ollama to source the model.
2. Creating a `-vn` all the files created will be easier to access and manage.
3. Utilize Ollama's Python API and LangChain to access the LLM and create conversations: the author explains how building the LLM interaction logic with LangChain, specifically using the package `ChatPropmtTemplate`, it is possible to include a conversation context and user queries. This allows for maintaining conversational history and generating responses through programming.

4. Create the front-end using Bootstrap: Bootstrap is a CSS framework that helps create easy web user interfaces by using providing several components like buttons, drop downs, forms, alerts, tabs, etc. (“How to use Bootstrap with Flask”, 2021)
5. Connect the Flask front-end with the model back-end: This step requires connecting the HTML front-end to the back-end endpoints to send user inputs and receive the AI generated content. This should result in a complete web application with a chat style able to interact with an LLM.

(Sekar, 2024)

### 7.5.2 HTML Interface and Langflow

Developing a User Interface (UI) will shape the way users interact with archival data and the AI system. By adding simple text input boxes, chatbots, or more advanced functions like easy and safe user log-ins, the design aims at intuitive querying, exploration, and display of archival content. As a proposal, creating HTML pages with a minimal responsive design powered by Bootstrap will be able to host conversational chat widgets that query the back-end GPT model.

As mentioned above, Bootstrap is a framework that will facilitate the creation of a minimalist looking but functional user interface. Within its advantages, Bootstrap does not require high programming skills, allowing any developer to build fast prototypes that can be developed into functional applications. The framework is also built to adapt and display properly in desktop or mobile devices, making it also a good option for deployment across several browsers like Chrome, Safari or Firefox. (“How to use Bootstrap with Flask”, 2021)

The following HTML code uses Bootstrap to create a simple usable interface but does not include the integration of the model yet.

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width,
  ↪ initial-scale=1.0">
  <title>Latin American Heritage GPT</title>
  <!-- Bootstrap 5 CDN -->
  <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0/dist/css/b_
  ↪ ootstrap.min.css"
  ↪ rel="stylesheet">
```

```

<link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.4.1/css/bootstrap.min.css">
  <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.7.1/jquery.min.js"></script>
<script src="https://cdn.jsdelivr.net/npm/jquery@3.7.1/dist/jquery.slim.min.js"></script>
  <script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.1/dist/umd/popper.min.js"></script>
  <script src="https://cdn.jsdelivr.net/npm/bootstrap@4.6.2/dist/js/bootstrap.bundle.min.js"></script>

  <!-- Custom CSS -->
  <link rel="stylesheet"
    href="/Users/paolamlechuga/Desktop/GIU_GPT/css/style.css">
  <!-- Google Fonts -->
  <link rel="preconnect" href="https://fonts.googleapis.com">
  <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin>
  <link href="https://fonts.googleapis.com/css2?family=DM+Sans:ital,opsz,wght@0,9..40,100..1000;1,9..40,100..1000&display=swap"
    rel="stylesheet">
  <link rel="stylesheet" href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/6.5.0/css/all.min.css">

</head>
<style>

table, th, td {
  border:0px solid black;
}
.p1 {
  font-family: 'DM Sans', sans-serif;
  font-size: 25px;
  font-weight: 700;
  font-stretch: 10%;
  font-display: swap;
  margin: 0;
}
.p2 {
  font-family: 'DM Sans', sans-serif;
  font-size: 10px;
  font-weight: 500;
  font-stretch: 10%;
}
.p3 {
  font-family: 'DM Sans', sans-serif;
  font-size: 25px;
  font-weight: 500;
  font-stretch: 10%;
}

```

```

        font-display: swap;
    }
.p4 {
    font-family: 'DM Sans', sans-serif;
    font-size: 20px;
    font-weight: 300;
    font-stretch: 10%;
    font-display: swap;
}
p.groove {border-style: groove;}
p.ridge {border-style: ridge;}

p.a {text-align:right; width:15%;}

.form-inline input {
    vertical-align: middle;
    margin: 5px 100px 5px 0;
    padding: 10px;
    background-color: #fff;
    border: 1px solid #ddd;
}
.form-inline {
    display: flex;
    flex-flow: row wrap;
    align-items: center;
}
.grid-container {
    display: grid;
    grid-template-columns: 90% 10%;
    gap: 1px;
    background-color: #fff;
    padding: 10px;
}
.grid-container > div {
    background-color: #fff;
    color: #000;
    padding: 10px;
    font-size: 30px;
    text-align: left;
}

.borde { border-color: #a7a7a7}

</style>

<body>
    <header class="navbar navbar-expand-lg navbar-light bg-light px-3 ">
<h1 > <p class="p1"> LATIN AMERICAN HERITAGE </p>

```

```

</h1>

<div class="ms-auto d-flex align-items-center">
  <span class="language-label text-light me-2"> Language:
  </span>
  <div class="language-options d-flex">
    <a href="#" class="btn btn-sm btn-light me-1">ES</a>
    <a href="#" class="btn btn-sm btn-light me-1">PT</a>
    <a href="#" class="btn btn-sm btn-light me-1">DE</a>
    <a href="#" class="btn btn-sm btn-light me-1">EN</a>
    <a href="#" class="btn btn-sm btn-light me-1">FR</a>
    <a href="#" class="btn btn-sm btn-light">About</a>
    <a href="#" class="btn btn-sm btn-light">Dataset
      ↪ Info</a>
  </div>
</div>
</header>

<form>
  <div class="col-lg-offset-1 col-lg-10 input-group input-group-lg">
    <input type="text" class="form-control" placeholder="Dame la
      ↪ receta de arroz chaufa si vivo en Suiza">
    <div class="input-group-btn">

<button style="font-size:24px"> <i class="fas fa-microphone"
  ↪ aria-hidden="true"></i></button>
  </div>
</div>
</form>
</div>

<div class="container-fluid">

  <div class="row col-lg-offset-1">

<!-- Sidebar -->
  <!-- <nav class="col-md-3 col-lg-2 d-md-block bg-light sidebar
  ↪ py-3"> -->
<nav class="col-md-2 col-lg-2 d-md-block bg-light sidebar py-3">
  <div class="sidebar-sticky">
    <h5 class="mb-3">Cultural Filters</h5>
    <ul class="nav flex-column">
      <li class="nav-item"><a class="nav-link"
        ↪ href="#">Food</a></li>
      <li class="nav-item"><a class="nav-link"
        ↪ href="#">Migration History</a></li>

```

```

        <li class="nav-item"><a class="nav-link"
        ↪ href="#">Music</a></li>
        <li class="nav-item"><a class="nav-link"
        ↪ href="#">Political Context</a></li>
        <li class="nav-item"><a class="nav-link"
        ↪ href="#">Urban Memory</a></li>
        <li class="nav-item"><a class="nav-link"
        ↪ href="#">Community Voices</a></li>
    </ul>
</div>
</nav>

<!-- Main Content -->
<!-- <main class="col-md-9 ms-sm-auto col-lg-10 px-md-4 py-4"> -->
<main class="col-md-9 ms-sm-auto col-lg-10 px-md-5 py-1">
<div class="row">

<!-- <div class="container p-3 my-3 border"> -->

<div class="container border p-3 my-3">

    <div class="grid-container">
        <div class="p2" style="font-size:15px">Model</div>
        <div class="p2" style="font-size:15px">ES</div>

    </div>
<div class="chatbox mb-4">
    <div class="chat-messages p-3 mb-3 bg-white rounded
    ↪ shadow-sm">
        <!-- Example chat bubbles -->
        <div class="chat-bubble bot mb-2 p3">
            Claro, aqui esta la receta de arroz chaufa
            ↪ basada en los diarios de Andres del
            ↪ Monte, un inmigrante peruano que vive en
            ↪ Basilea, Suiza.

        </div>
    </div>
</div>

<!-- </div> -->

<!-- <div class="row"> -->

    <table style="width:100%">
        <tr>

```

```

<td><h2 class="mb-4 p4">Visual Archive</h2>

  <div class="card-body">
    <!-- <h5 class="card-title">Arroz Chaufa</h5> -->
    <p class="card-text">“Este plato siempre me
      ↳ conecta con mi hogar.”
    </p>
  </div>
</td>
<td>
  <!-- <div class="card"> -->
  <div >
    
  </div>
</td>
</tr>
</table>

<!-- <div class="row"> -->

<a href="#" class="btn btn-outline-dark borde" >View Recipe</a>
  <!-- More cards... -->
</div>
</div>
</main>

<!-- Footer -->
<footer class="footer mt-auto py-3 bg-light d-flex
↳ justify-content-between align-items-center px-4">
  <small class="text-muted">This is a Fine-tuned model still under
↳ community review</small>
  <div class="footer-buttons">
    <button class="btn btn-light btn-sm me-2">Contact</button>
    <button class="btn btn-light btn-sm
↳ me-2">Contribute</button>
    <button class="btn btn-light btn-sm me-2">Citation</button>
    <button class="btn btn-light btn-sm">Data Ethics</button>
  </div>
</footer>

<!-- Bootstrap JS Bundle -->
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0/dist/js/b_
↳ ootstrap.bundle.min.js"></script>
</body>
</html>

```

As another option, visual workflow frameworks like Langflow enable rapid and easy prototyping of the AI logic, combined with a front-end like StreamLit, Gradio, or React.js, or using embedded widgets can integrate the model easily into existing websites.

Langflow (LF) is a visual, low-code development platform designed to quickly create and deploy AI applications; this platform allows RAG systems, chatbots, and multi-agent frameworks.

The main features of this platform include:

- Visual Workflow Builder: LF provides drag-and-drop components to create complex apps without the need of programming.
- Rapid Prototyping: Instant experimentation with different prompts, models, and data sources for fast iteration.
- Multi-agent and Multi-tool support: Easy orchestration of several agents, route queries, or assign sub tasks using nodes and tool-calling logic.
- Native Retrieval Augmented Generation Functions: Integration of vector-stored data and retrieval pipelines to ground responses in archival or domain-specific data.
- Extensive Integrations: With LF it is easy to integrate databases, API's, Python functions, or custom tools. It is also possible to embed the LF chat widget or import other sourced workflows.
- Rapid Observing and Debbing: In LF, it is possible to use functions like LangSmith and LangFuse for workflow monitoring and event logging.
- Flexible Deployment: LangFlow offers a wide range of deployment options, whether using a cloud server, Docker, or other platforms.
- API and Widget Publication: With LF it is possible to publish flow projects as APIs, for code sharing, and accessing other users widgets.

Here is a first proposal on how to include the LangFlow embedded chat widget while leaving the first HTML proposal intact:

```
<!-- Langflow Embedded Chat Widget -->
<script src="https://cdn.jsdelivr.net/gh/logspace-ai/langflow-embedded-
↪ chat@v1.0.7/dist/build/static/js/bundle.min.js"></script>
```

```

<div style="width:100%; height:600px; margin-top:20px;">
  <langflow-chat>
    host_url="http://localhost:7860" <!-- LangFlow URL where the
      ↪ instance is running on -->
    flow_id="YOUR_FLOW_ID" <!-- Unique identifier of the LF flow
      ↪ created -->
    api_key="YOUR_API_KEY" <!-- Langflow API key if authentication is
      ↪ used -->
    language="en" <!-- Set default language -->
  </langflow-chat>
</div>

```

Secondly, there is also the option to make the existing form send requests to Langflow API via JavaScript:

Instead of adding the LangFlow widget and modifying the existing Bootstrap input box, it is possible to hook the input box with the Langflow API with a JavaScript `fetch` function.

The process works as followed: First, give the input box from the original HTML file an `id` for easy selection:

```

<input type="text" class="form-control" id="userInput"
  ↪ placeholder="Dame la receta de arroz chaufa si vivo en Suiza">

```

Then, add a submit handler for the form or button:

```

<form id="queryForm">
  <!-- your input and button -->
</form>

<script>
document.getElementById("queryForm").addEventListener("submit", async
  ↪ function(event) {
    event.preventDefault();
    const input = document.getElementById("userInput").value;
    if (!input) return;

    // Show "loading" or disable button here as needed

    // Call Langflow API endpoint
    const response = await
      ↪ fetch("http://localhost:7860/api/v1/run/YOUR_FLOW_ID",
    // Replace YOUR_FLOW_ID with the actual LangFlow id
    {
      method: "POST",
      headers: {

```

```

    "Content-Type": "application/json",
    "Authorization": "Bearer YOUR_API_KEY"
    // Replace with actual LangFlow API KEY
  },
  body: JSON.stringify({
    input_value: input,
    input_type: "text",    // or "chat" depending on your flow
    output_type: "text"
  })
});
const data = await response.json();

// Process and display the response in your chat area
const chatMessages = document.querySelector(".chat-messages");
const userBubble = `

Finally, styling the inserted chat bubbles to match the current chat box will provide a user interface with the original HTML front-end retrieving queries directly from the model back-end.



### 7.5.3 Archival Database Integration



Integrating archival databases can also improve the utility of the Web application. This involves connecting the AI model and user interface to the archival data sources, whether they are in a SQL database, NoSQL storage, or vector search engines for RAG. Such integration will allow the model to fetch relevant documents in real time, as well as incorporate contextual metadata, and provide precise, natural responses directly from the source corpus. (Gao et al., 2024)



For SQL databases like MySQL or SQLite, its possible to use an object relational mapper. Using SQLAlchemy enables connecting databases through programming, execute queries, and retrieve records found the in data as Python objects.



NoSQL or document oriented databases like MongoDB require to use other libraries to fetch documents via APIs.



48


```

Finally, RAG retrieval helps enrich the responses provided by the model with the archival data used for the fine-tuning process. By providing not only the metadata but also the processed data or externally sourced databases, hallucination can be prevented enhancing the model's ability to reason. The process of using both techniques is called RAFT (Retrieval-Augmented Fine-Tuning).

To implement this, it is necessary to follow the following steps:

1. **Embedding and Indexing:** This step requires processing the archival data to turn them into vector representations using embedding models. Once the vectors are created, they can be indexed in a vector database such as FAISS or Pinecone. (noauthor \_ faiss \_ 2025)
2. **Retrieval:** This step occurs should occur when the user makes a query, the system will generate an embedding for the query and search the vector database for the most relevant entries.
3. **Augmentation and Generation:** The original query provided by the user, along with the retrieved information, is formatted into a single prompt that is then sent to the model. The model should then generate a summary, answer, or output based on the fine-tuned data and the archival vector database.

In a study performed by T. Zhang et al., 2024, the authors experimented using RAFT as a method for accurate information retrieval. With RAFT, the model can learn domain-specific knowledge through fine-tuning without getting distracted while retrieving information. With this study it is established how implementing RAG techniques into the model and user interface will improve the performance of the model.

As mentioned above, RAG techniques can be implemented using FASTAPI, linking them directly to the existing HTML/JavaScript code. Here is the proposed code to do so:

```
# Save this as rag_service.py and install dependencies:
pip install fastapi uvicorn transformers sentence-transformers faiss-cpu

from fastapi import FastAPI
from pydantic import BaseModel
from transformers import pipeline
from sentence_transformers import SentenceTransformer
import faiss
import numpy as np

app = FastAPI()
```

```

# -- Archival data sample --
archival_data = [
    {"title": "Immigration Dossier", "text": "Carlos Reyes arrived in
    ↪ Basel in 1965..."},
    {"title": "Oral History with Maria", "text": "Maria Lopez recounts
    ↪ her experiences migrating from Peru..."},
    {"title": "Newsletter 1970", "text": "The Latino Cultural
    ↪ Association organized a major festival..."}
]

texts = [doc["text"] for doc in archival_data]

# -- Embedding Model & Index --
embedder = SentenceTransformer('all-MiniLM-L6-v2')
embeddings = np.array(embedder.encode(texts))
index = faiss.IndexFlatL2(embeddings.shape[1])
index.add(embeddings)

# -- LLM Pipeline (swap to Ollama for production) --
llm = pipeline("text-generation", model="llama3")

# -- Request Structure --
class QueryRequest(BaseModel):
    query: str
    k: int = 2 # number of retrieved documents

# -- RAG Endpoint --
@app.post("/rag_generate/")
def rag_generate(req: QueryRequest):
    query_embedding = np.array(embedder.encode([req.query]))
    D, I = index.search(query_embedding, req.k)
    retrieved = [texts[i] for i in I[0]]
    context = "\n".join(f"[{i+1}] {txt}" for i, txt in
    ↪ enumerate(retrieved))
    prompt = f"User query: {req.query}\nRelevant archival
    ↪ info:\n{context}\n\nAnswer:"
    response = llm(prompt, max_new_tokens=150)[0]["generated_text"]
    return {"retrieved": retrieved, "response": response}

```

Running the code:

```
uvicorn rag_service:app --reload --port 8000
```

Integration of the RAG functions to the HTML code:

```

<!-- Add this to the HTML page to any part of the body -->
<form id="ragForm">
    <input id="userInput" class="form-control" placeholder="Ask about the
    ↪ archive..." />

```

```

<button type="submit">Ask</button>
</form>
<div id="resultArea"></div>

<script>
document.getElementById("ragForm").onsubmit = async function(e) {
  e.preventDefault();
  const query = document.getElementById("userInput").value;
  const response = await fetch("http://localhost:8000/rag_generate/", {
    method: "POST",
    headers: {"Content-Type": "application/json"},
    body: JSON.stringify({query: query, k: 2})
  }).then(res => res.json());

  // Display retrieved context and model response
  document.getElementById("resultArea").innerHTML =
    `<b>Context:</b><br>${response.retrieved.map(x =>
    ↪ `<div>${x}</div>`).join("")}<br><b>Answer:</b>
    ↪ ${response.response}`;
};
</script>

```

#### 7.5.4 Large-Scale deployments

For larger-scale deployments, container orchestration with platforms like Kubernetes (K8S) is possible. Kubernetes is an Open Source, FAIR platform that supports load balancing, automatic fail-over, and horizontal scaling to handle multiple simultaneous users. It also offers robust security features such as role-based access control and network isolation, this is helpful to protect sensitive historical and personal data. (“Overview on Kubernetes”, n.d.)

For AI tools with low latency and high data usage and security, K8S offers:

- Scalability and Load Balancing: K8S can add or remove containers based on metrics obtained from the CPU, GPU, and request rates. If many users interact with the GPT at the same time, the app increases container replicas to handle the energy and power demand. On the other hand, the app also allows user traffic to be evenly distributed among several containers, preventing the back-end to over-saturate or fail while being extracted.
- High Availability and Automatic Failover: If something crashes on the app, K8S detects the failure and creates a new instance to automatically replace it.

- **Robust Security Controls:** Administrators can define user and service accounts permissions. Access to the archival data, model, and sensitive information can be isolated depending on the admin needs. K8S also supports network segmentations. For example, a database containing personal information can be restricted so only certain authorized containers can query it.
- **Operational Flexibility and Monitoring:** K8S manages custom AI resources, such as scheduling inference jobs and vector-based reloads.

To implement Kubernetes in this project, it is necessary to take several steps:

1. **Containerize the components:** For the FastAPI manifest, create a Dockerfile directly on terminal:

```
FROM python:3.10-slim
WORKDIR /app
COPY . /app
RUN pip install -r requirements.txt
CMD ["uvicorn", "rag_service:app", "--host", "0.0.0.0", "--port",
↪ "8000"]
```

For the Model deployment in Ollama:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: ollama
  namespace: ollama
spec:
  replicas: 1
  selector:
    matchLabels:
      app: ollama
  template:
    metadata:
      labels:
        app: ollama
    spec:
      containers:
        - name: ollama
          image: ollama/ollama:latest
          ports:
            - containerPort: 11434
# Expose with Service:
apiVersion: v1
```

```
kind: Service
metadata:
  name: ollama
  namespace: ollama
spec:
  type: NodePort
  selector:
    app: ollama
  ports:
  - port: 80
    targetPort: 11434
```

2. Write the K8S YAML Manifests: For each component, create both a Deployment (for the code/images) and a Service (to expose it inside/outside the cluster). Typical files are:

- `fastapi-deployment.yaml` and `fastapi-service.yaml`
- `langflow-deployment.yaml` and `langflow-service.yaml`
- `ollama-deployment.yaml` and `ollama-service.yaml`

Each deployment specifies resources like CPU/GPU, RAM. Services can be ClusterIP, NodePort, or LoadBalancer types depending on the cloud or local setup.

3. Cluster Setup: Create a K8S cluster using the preferred cloud or local solution.
4. Network and Orchestration: Configure the secure HTTP routing, and traffic distribution. The use the Horizontal Pod Auto scaling for scaling backend/model-serving pods in response to traffic. Finally secure containers and connections using Kubernetes secrets and RBAC for API keys, database credentials, and private endpoints.
5. Connect Everything: The HTML/JS front-end should point to the appropriate API endpoints or NodePorts on the Kubernetes cluster, in this case Langflow and FastAPI. Finally, Langflow, the back-end, and vector database can be connected with the Ollama service inside the cluster by referring to its service DNS name:

```
http://ollama.ollama.svc.cluster.local:80
```

With this, it is possible to set environment variables for URLs directly with the K8S manifests.

### 7.5.5 Security Considerations

Monitoring and maintenance are also important aspects of the deployment process, these include logging user interactions, tracking performance metrics, and updating the model or datasets based on user feedback and new data. Continuous evaluation helps mitigating the model from hallucinating and generating bias, making sure the system stays accurate, inclusive, and culturally appropriate.

## 7.6 Documentation and Maintenance

As an important part for the development of this project and after deployment, is necessary to create a maintenance and documentation strategy that addresses technical operations and user clarity.

To ensure that, the next steps should be taken into consideration:

- Documentation of the Codes and Workflows
- Creation of a Maintenance Strategy
- Documentation of User Interaction
- Data Governance and Provenance
- Automation and consistent Review

### 7.6.1 Documentation of the Codes and Workflows

For this step, it is essential to write down all the processes and resources used. Creating a repository on GitHub, and storing it in the Swiss National Data and Service Center for the Humanities (DASCH), divided by several directories that should include all the codes used, configurations, `.yaml` manifests, Docker files, Langflow flows, and front-end assets. Here is the proposed directories for the GitHub repository:

- `/backend`: For the FastAPI and Python Fine-tuning Codes
- `/frontend`: For all the HTML, JS, CSS, and Bootstrap assets
- `/k8s`: To store the Kubernetes charts and containers
- `/langflow`: Documentation of the flows created in LangFlow

- /docs: For all the mermaid workflow diagrams and LLM usage guides.

Including different directories, each of the previously mentioned should also contain a `README.md` file that establishes an overview of the project, to which audience is intended for, and usage guidelines.

This file should also include a summarize of each component used and how they interact with each other. For example:



Figure 2: System flow from user query to output

In addition, the `README.md` document should also include the list of all the technologies used, such as FastAPI, Ollama, LangFlow, Pinecone and Kubernetes, a setup document on how to install all the necessary dependencies, as well as configuration of the virtual environments needed to run the several codes and platforms, case scenarios to fix potential errors, and contact information to report any issues encountered during the project.

With this, it can be ensured that the different stages of the project are correctly documented for further preservation and long term usage.

### 7.6.2 Maintenance Strategy

It is essential to perform constant updates for all the components of the project, including Ollama, Langflow, the front-end, and any vector databases. As a proposal, rebuilding Docker containers monthly or when security alarms are executed, is recommended. In addition to this, keeping the Python and JSON codes updated, as well as their libraries, can help to improve the stability of the deployment. This can be performed by assigning a data warden in an institution for regular checkups. To make this process easier, automated tools like Dependabot (for GitHub) or GitLab's Dependency Scanning can be used. These tools automatically create pull requests to update the dependencies listed in the configuration files whenever new versions or security updates are available. (dependabot, n.d.)

Before running the updates, it is important to validate them in a dedicated test environment. This ensures that no regressions or issues are introduced or created in the original documentation. Maintaining an automated diary of every version control system or release will help keep track of every update and the reason for why the updates, security checks, or additions were made.

On the other hand, regular updates of the model and embedding indexes are important to maintain the accuracy of the whole system. Scheduling periodic reviews to assess whether retraining, fine-tuning again, or switching model is necessary based on feedback, performance reviews, or new incoming archival data. Detailing the documentation of all training runs, new fine-tuning, or RAG processes, including hyperparameters, datasets, and version information, will ensure the possibility of reproducing the project and identifying errors.

For the embeddings and vector database, automated processes can be established to detect newly or modified archival documents. These scripts should perform the re-embedding of just the updated or new documents to optimize power use. Archive and old versions of the vector indices should be kept in case it is necessary to restore the state of the databases. All changes to models and data should be carefully documented as previous or updated versions. Before the deployment of new versions into the original project, a benchmark comparison test should be performed and accepted to prevent failures while running the project.

Automation of backups of all the data, including the databases and workflow configurations, can be effectuated into reliable storage locations such as cloud storage or Network Attached Storage (NAS) devices. NAS offers the possibility to allow multiple users to store and share files via WiFi or Ethernet. (Susnjara and Smalley, 2023) The backup data can be stored in JSON formats that include metadata timestamps, user and session identifiers, requests, and component details for easier access with the help the Research and Infrastructure Support department from the University of Basel.

Physical archiving can be performed for this project, by backing up into hard drives and storing them in specific storage rooms, institutional or governmental requirements can be covered to prevent sensitive data loss.

In order to prevent failure over the model, user interface, or data retrieval, it is possible to set alerts to detect error patterns, including crashes in the server, unnecessary restarts, web application loading errors, or failed queries. By using communication tools like Slack, or email, fails to prompt or query data can be identified. Maintaining an incident diary that describes common failures, diagnostic steps, solving procedures, and recovery workflows will make problem solving more efficient with a reduced amount of time spent and lower impact into the finished project.

This approach helps maintain a reliable, secure platform that can efficiently serve the users while keeping safe and up to date the model, data, and infrastructure.

### 7.6.3 User Documentation

By providing a clear user guide for potential users, the benefits for engaging and optimizing of the model are higher. Creating a step by step instructions manual that walks the user through the functionalities, like submitting queries, navigating through the databases, interpreting information, adding visual aid on how to navigate the platform, and perform accurate input prompting will provide the user a better result each time. Providing user guides within the platform and as a part of the official repository will become beneficial for the project and future users.

Adding a FAQ (Frequently Asked Question) section in the user interface to address possible common question and mistake scenarios, such as how to improve prompting, or what to do when the results are not the desired ones, how to send feedback on hallucination or bias, or error reports will aim to improve user satisfaction and reduce support loads.

As for the repository, adding a `CHANGEDIARY.md` file can be helpful for maintaining institutional and government transparency. This document should include all updates, including new features, bug fixes, performance improvements, security updates, and data changes. Each entry in the file should include a date, a version number, a summary of any changes done and references related in case any requests or issues are specifically found. This document also provides a guidelines for all the administrators of the project to follow any updates or changes performed by someone else. This helps with preventing correction duplicates or missed errors lost during several review processes.

Adding these documents to the repository can also provide new developers, researchers, or future collaborators with good idea on how to join or support the project. These can cover how to set up the environments, installing dependencies, configure local services like Ollama or the Langflow web app, and running all the front-end assets. With all the mentioned above, including coding standards, workflow procedures, and instructions on how to use, extend or update the system. Attached to that, including examples of datasets, queries, and debugging tips will ensure others installations run smoother.

As for institutes and other administrators, it is important to establish in a document the names of all the responsible people for deploying, operating, and securing the platform. This should include a guide on how to set up the clusters, use of Kubernetes and Docker, access and management controls, what to do in case of failure during back ups, data recovery, and network security configurations.

By creating this user and developer guides, the project is grounded a solid base for

long term operation and archival. This documentation ensures clarity and continuity as the project might change over time.

#### 7.6.4 Data Governance and Provenance

Keeping a clear documentation of the data schema and metadata is necessary to have a good foundation for the reliability, usage, and long term archival of the project. This documentation should include data fields, types, relationships, and allowed formats. For example, replicating the same format styles as the JSON database files which include document ID, title, creation or extraction date, language, and source provenance. Records concerning data transformation should also include how the raw data is obtained, cleaned, processed, embedded, or converted to formats suitable for posterior processes.

To ensure interoperability, all data stored in the documentation must stick to a defined ontology that includes fields like:

- record\_id: string value and unique identifier
- title: string value
- author: string value
- date: ISO 8601 date format
- language: ISO language codes
- document\_type: for example, oral\_history, official\_record, newspaper, photograph, etc.
- text: string value
- source: string value, summary of provenance
- metadata: JSON format with additional values like location and tags

Schema template for the archival records:

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Latin American Heritage Archival Record",
  "type": "object",
  "required": ["record_id", "title", "date", "language",
    ↪ "document_type", "text", "source"],
  "properties": {
```

```

"record_id": {
  "type": "string",
  "description": "Unique identifier for the archival record"
},
"title": {
  "type": "string",
  "description": "Title or headline of the archival document"
},
"author": {
  "type": ["string", "null"],
  "description": "Author or creator of the document"
},
"date": {
  "type": "string",
  "format": "date",
  "description": "Creation or publication date in ISO 8601 format"
},
"language": {
  "type": "string",
  "description": "ISO 639-1 language code, e.g. 'es', 'en', 'pt'"
},
"document_type": {
  "type": "string",
  "description": "The type of document: e.g., 'oral_history',
  ↪ 'official_record', 'newspaper', etc."
},
"text": {
  "type": "string",
  "description": "Full transcribed or digitized text content for the
  ↪ record"
},
"source": {
  "type": "string",
  "description": "Provenance information, institutions or
  ↪ collections"
},
"metadata": {
  "type": "object",
  "description": "Additional metadata fields",
  "properties": {
    "geographic_location": { "type": "string" },
    "subject_tags": {
      "type": "array",
      "items": { "type": "string" }
    },
    "rights": { "type": "string", "description": "Copyright or usage
    ↪ rights status" }
  },
  "additionalProperties": true
}

```

```
    }  
  },  
  "additionalProperties": false  
}
```

Clear metadata documentation will enable handling and querying data consistently, as well as facilitating tracking the origin of the data, making it possible to find summaries or AI generated responses back to the original sources. It is important to note that in humanities and history research, source attribution is required to be strongly influenced by ethical practice. In a document written by The Finnish National Board on Research Integrity TENK, it is established that researchers in the area of humanities, arts, and social sciences should follow the next principles:

- Core Principles:
  - Autonomy: Voluntary, informed consent; exceptions only where legally permitted.
  - Avoid Harm: Minimize risks; balance potential harm with scientific value.
  - Privacy: Safeguard confidentiality and personal data at all stages.
- Ethical Review: Recommended via the creation of an ethics committees for sensitive research; review must occur before data collection; researchers or developers carry the ultimate responsibility.

(“Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review”, 2009)

Furthermore, documenting policies and workflows that include updating, versioning, and deleting archival records will ensure the integrity of the dataset over time, supporting data quality, and keeping institutional standards.

As for ethical considerations regarding the maintenance and documentation, the data should follow, as the mentioned above EU regulations, governing policies about privacy, consent, and data rights. These policies should be stated in the official repository and user interface. Describing how the data will be managed, and how sensitive information would be handled, will help the participants know their rights and the potential uses of the data. In cases where participants want to withdraw consent for using their data, the system should be able to remove the information through data removal requests. These

requests and procedures should also be documented in the repository as well as in the official documentation records.

By doing a rigorous documentation and enforcing ethical standards, participants and users will gain trust to contribute and use the system. This will also ensure that the platform respects human dignity and legal frameworks while also allowing historical and cultural exploration.

This approach to data governance and provenance protects the quality, transparency, and ethical foundation of the project, encouraging accurate education and community engagement.

## 7.7 Model Capabilities and Limitations

Building a domain-specific GPT model as proposed aims to offer a wide range of possibilities that enhance the everyday user experience and the richness of academic research.

One of the main strengths of this is the support for conversational engagement. Users can prompt highly complex questions to the model and receive contextual answers extracted from the corpus provided to the model. This quality combines the gap between unstructured archival content and existing inaccessible or "hard to find" knowledge. With this, the model seeks to empower researchers and community members to interact with natural language and real content in a more intuitive way, preventing the extensive task of using traditional search engines.

Another important function present in this proposal is summarization, for example, standardizing Tags through LLMs. The model will be able to compress different types of documents, depending on their size or format, into concise and coherent summaries. This allows users to access essential information quickly without having to browse through large amounts of materials, making the everyday user experience faster and more accurate in-depth research.

Thematic linking also makes a difference between the model and traditional search engines. It can identify connections among related archival items, such as identifying a link between a specific photograph and its background story found in the corpus. This makes it easier to conduct deeper exploration of the archives, helping users uncover relationships, narratives, and generate relevant and useful content that might otherwise remain undiscovered or hidden among different sources.

Spatharioti et al., [2025](#), mention in their conference paper *Effects of LLM-based Search*

on *Decision Making: Speed, Accuracy, and Overreliance* the large drawbacks when using traditional web search. Although it is considered more convenient to have access to different sources, the information gathered during these searches are considered time consuming and challenging to summarize. The relevant information usually is found, as mentioned in the paper, as “instant answers” or snippets, resulting in users having to click and scroll through different links to find the information needed.

For example, someone who recently moved from Peru to Switzerland might be interested in cooking Chaufa rice or Peruvian fried rice, a specialty found only in their country. This person can access the GPT model and create the following prompt:

```
"Dame la receta para cocinar arroz chaufa si vivo en Suiza." a  
"English translation: "Give me the recipe to cook peruvian fried rice if I live in Switzerland."
```

The model then will provide an answer based on the corpus, where it will browse Peruvian recipes as well as testimonies of Peruvian immigrants living in Switzerland who previously mentioned in an interview or had any written documents related to the dish. Like this:

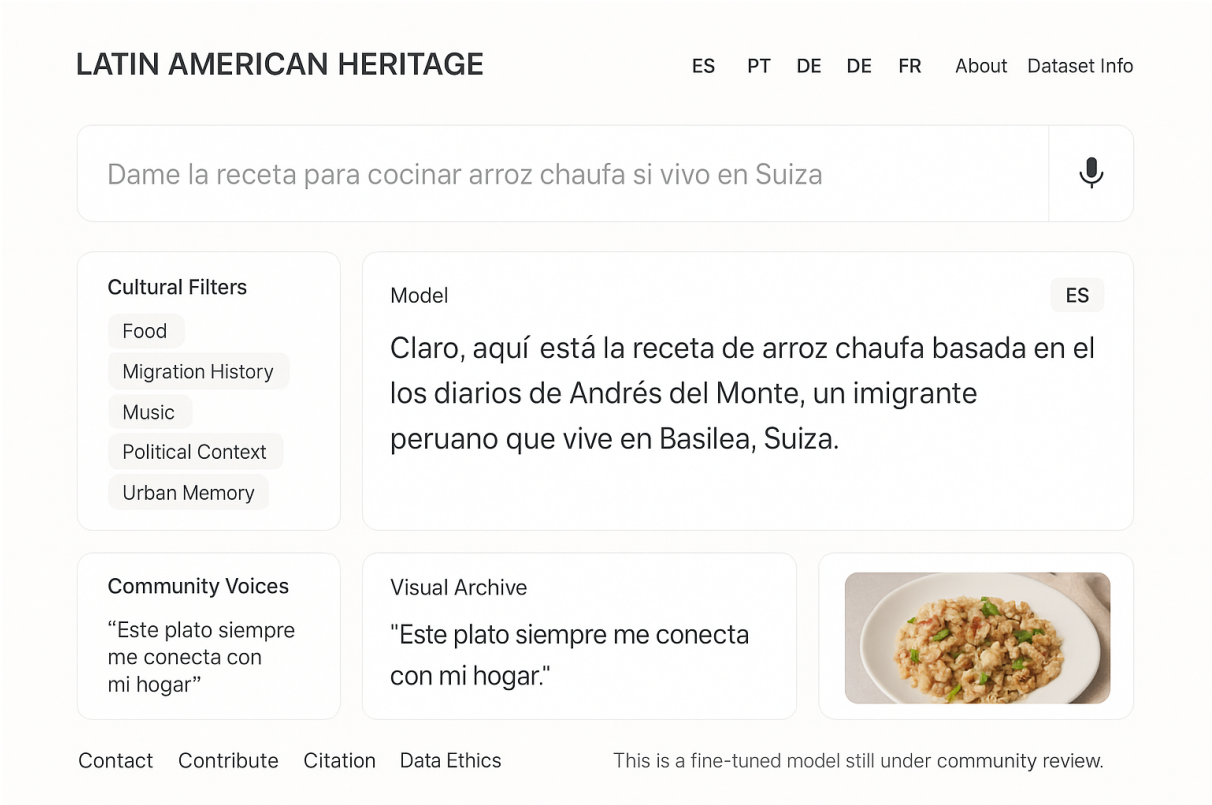


Figure 3: Visual example of the question and response <sup>10</sup>

<sup>10</sup>This image has been generated with AI; English translation: Sure, here is a recipe of Peruvian fried rice based on the diaries of Andrés del Monte, a Peruvian immigrant living in Basel, Switzerland.

Multilingual support is another important aspect when comparing a domain-specific LLM with traditional search engines. Given the possibility of having a wide linguistic diversity in the corpus, important when it comes to not excluding multicultural communities, the model will have the ability to process and generate text in multiple languages with the purpose of broadening the access and inclusiveness of diverse users. By fine-tuning the model with other languages besides English, such as Spanish, Portuguese, French, and German, we can prevent possible bias coming from the lack of linguistic diversity lost when everything is translated into English or other universal languages.

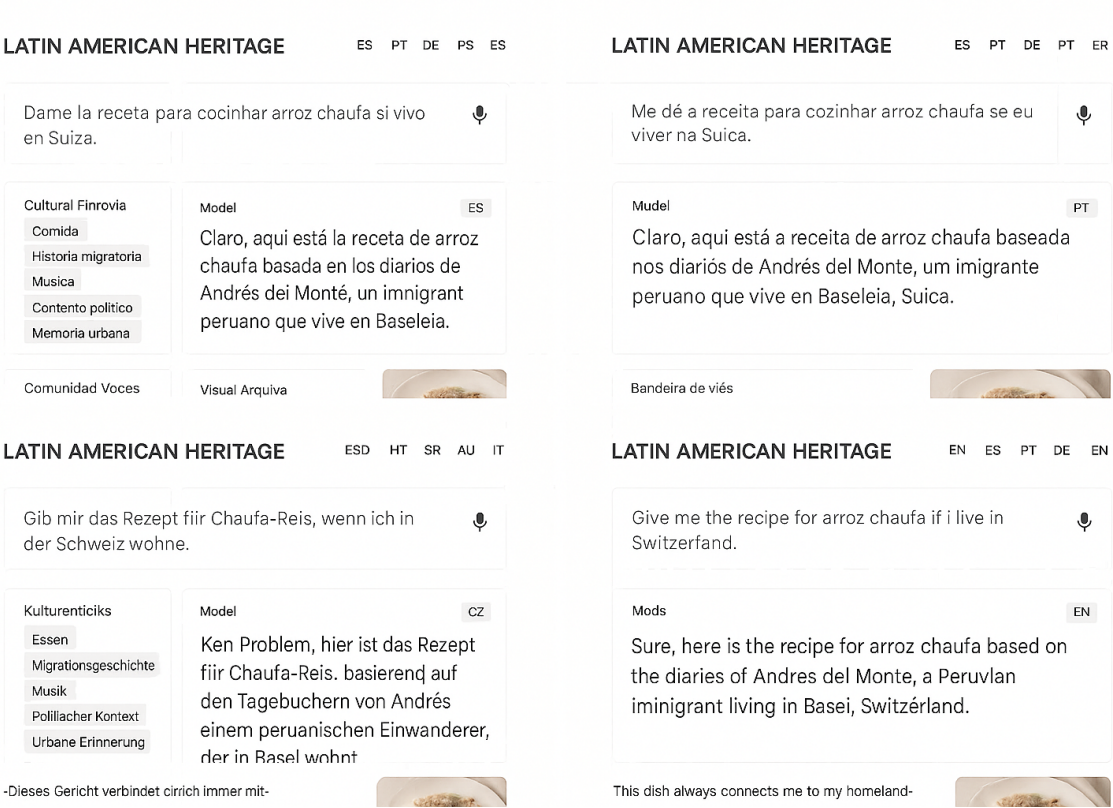


Figure 4: Example of the several multilingual options <sup>11</sup>

Finally, the model aims to illustrate strong content generation skills. It should be able to produce time reports, biographical sketches, and thematic overviews extracted from the dataset. These outputs will help with precise contextualization of historical events, creation of individual biographies, and other related themes, making content more engaging and meaningful to the user.

However, building this proposal might come with several important limitations that must be carefully considered during the development and deployment of the model. The model's accuracy is basically constrained by the provided data. The output results are

<sup>11</sup>This image has been generated with AI

only as reliable as the quality, representativeness, completeness, and structure of the data used to fine-tune the model. If the sources are fragmented, uneven, or biased, the GPT’s outputs may reflect these deficiencies, potentially excluding certain perspectives or misrepresenting historical records.

There is also a high risk potential for hallucination. The model, when confronted with gaps in information or ambiguous input, it can generate plausible responses that are not actually grounded in the source material. This tendency can lead to the propagation of inaccuracies, particularly in the form when users assume the generated content is fully accurate. In the case of hallucinations, it is important to mention the different types of results that can be provided by the GPT.

In a survey conducted by Huang et al., [2025](#), the authors describe two categories of hallucination:

- Factuality Hallucination, which divides into two different types:
  - Factual Contradiction: This corresponds to situations where the LLMs output contains facts that can be proved with real information, but also present certain contradictions. This type of hallucination happens when using diverse sources. This type of hallucination can furthermore be divided into two other subcategories: “entity error”, which refers to situations where the generated text contains wrong entities, like misinterpreting who the creator of the telephone is, as the survey mentions. The other subcategory is “relation error”, which occurs when the generated text outputs a wrong relationship between entities, such as the example mentioned on the survey, where Edison is mentioned as the inventor of the light bulb instead of an improver of the first light bulb.
  - Factual Fabrication: This happens when the LLM’s output has facts that cannot be verified against real information. This type of hallucination can also be subdivided into two categories: “unverifiability hallucination” and “overclaim hallucination”. The first mentioned occurs when the statements provided by the LLMs are entirely fake or can’t be verified by other sources. As the authors exemplify, when asking a model about the impacts of the construction of the Eiffel Tower, the output states it had a direct impact in the extinction of the Parisian Tiger, which cannot be truth considering the species does not exist and there are no historical claims regarding this topic. “Overclaim hallucination”

happens when there is lack of universal validity due to bias. For example, claiming the Eiffel Tower’s construction is recognized as an event that sparked the green architecture movement, when there is no real consensus or enough evidence that claims this statement.

- **Faithfulness Hallucination:** This happens when LLMs are trained to align with user instructions. When the use of LLMs changes into a more user centered application, the consistency with user-provided instructions and contextual information is important. This type of hallucination can also be divided into three subcategories: “instruction inconsistency”, “context inconsistency”, and “logical inconsistency”.
  - “Instruction inconsistency” happens when the outputs deviates from the user’s commands. Some deviations might be unintentional, as mentioned in the paper, the example of asking the model to translate a question into another language. The model misinterprets the input, and instead of translating as the user desired, the output is a direct answer to a fragment of the input, which the model interprets as a question.
  - “Context inconsistency” happens when the output is unfaithful with the user’s provided information. For example, modifying the output summary from *The Great Lakes of Africa* to *the mountain range in Africa*. Here, the model simplifies and contradicts the real location of the Nile River.
  - “Logical inconsistency” happens when the outputs of the LLMs reveal internal logical contradictions, such as reasoning tasks. This inconsistency is found when comparing the reasoning steps with the final outcome. As an example mentioned by the authors, by solving the equation

$$2x + 3 = 11 \tag{1}$$

the reasoning of the model of dividing both sides of the equation by 2 is correct, the final answer of

$$x = 4 \tag{2}$$

is inconsistent with the reasoning chain, creating a wrong result.

(Huang et al., 2025)

Contextual gaps are another risk, especially in the absence of rich metadata or precise

annotation. Without sufficient context, the model can misinterpret relationships or overlook subtle but significant variations, decreasing the depth and reliability of the output. Employing only data acquired, for example, from Spanish diaries about the Mexican conquest, will not generate sufficient historical context to understand the Mexican conquest and Independence from the Mexican perspective.

On a practical level, the fine-tuning and deployment of LLMs require significant computational resources. High-performance hardware and technical expertise are required both to adapt the model to become domain-specific and to serve users efficiently, which can pose barriers for smaller organizations or community projects if they wanted to contribute to the process of fine-tuning.

Finally, the challenge of language and cultural bias is a remaining topic of discussion. The model may reinforce the biases that exist in its training data or that have been inadvertently encoded during general pre-training. For example, Tay, a Twitter bot trained to understand and learn the language through its environment. Tay started tweeting casually and engaging with followers, however, as Tay continued learning, it started posting offensive tweets, resulting in having to shut down 16 hours later. Tay is the result of reinforcing bias by unconsciously extracting mean tweets.(Ganesan, 2023) Attention to diversity in data sources and continuous bias assessment are vital to ensure that the tool serves its intended communities with fairness and cultural sensitivity.

A study developed by Naous et al., 2024 demonstrates how large language models, specifically comparing GPT-4 and JAIS-Chat (an Arabic specific LLM) when asking to complete sentences in Arabic regarding "going for a drink", they will primarily prioritize alcoholic and westernized beverages, before comprehending that going for a drink in the Muslim culture excludes alcoholic beverages. With this example, the author proposes that by neutral prompting or pre-training the model with Arabic data, is possible to mitigate cultural bias in LLMs.

## 7.8 Ethical and Practical Considerations

It is important to consider the ethics and practice of the development and use of AI tools for archival research, especially when handling sensitive or personal data. According to the European Commission, sensitive data corresponds to those revealing racial or ethnic origin, political inclination, religious or philosophical beliefs, memberships, biometric and health related data, and data concerning the person's sexual orientation. ("What personal

data is considered sensitive?”, n.d.) For this study and the development of the project, it is recommended to create an ethics committee that regulates the different aspects that will be mentioned below.

Privacy and consent must be considered a priority. Sensitive personal information, particularly that coming from interviews or recent records, requires careful management, including the possibility of making some of the data inside the records anonymous if the collection is also considered to be publicly accessible. This will ensure respect for individual autonomy and protection of the privacy of those willing to share their stories. Data anonymization can be performed in several ways; however, in order to preserve the integrity of the data, for this proposal, performing pseudonymization is considered the best option. The pseudonymization process keeps the data integrity as is without disrupting the fine-tuning process and will still ensure data privacy policies. (“What is Data Anonymization | Pros, Cons & Common Techniques | Imperva”, n.d.)

Community involvement is equally something to be considered. Engaging with members of the community throughout the project will provide respectful representation and cultural precision. By having constant dialogue, collaborative decision making, and responsiveness to community concerns, this will help to build trust and ensure the outcome project aligns with the values and expectations of those whose data are being used, histories are being preserved and interpreted, and from those who seek to make use of the model.

Transparency will help with ethical practice and academic guidelines. Clear documentation of data sources, annotation protocols, and model limitations should also be made publicly available. This will allow users to understand the origin of the data and the methodologies used, and this will promote trust and facilitate reproducibility of results.

To mitigate bias, it is required to carefully identify, monitor and address potential biases in the archival material, as well as the outputs generated by the AI model. (N. A. Ahmed, n.d.) This could involve diversifying the training data, applying rigorous annotation standards, and regularly testing the model responses. Strategies for mitigating bias will include:

1. Conscious data curation: Ensure that the training data used for the model has been accurately curated from diverse sources. Human and machine involvement in the annotation and processing of the data will help mitigate possible unbalance or under-representation.

2. Employment of transfer learning techniques during the Fine-Tuning process: Jain, [2025](#), describes transfer learning as a technique in machine learning that allows models to leverage knowledge from a specific task to improve the performance on another task. Using smaller datasets to fine-tune the model and run the fine-tuning process for smaller, newer tasks will help prevent data over saturation.
3. Implementing metric evaluation: Methods like human evaluation, automatic evaluation, or hybrid evaluation can result in anticipating, detecting, or filtering biases in the model. Examples of metrics could be:
  - Perplexity: Evaluate the model's ability to predict the next word in an output.
  - Accuracy: Evaluation of how correct the outputs of the model are. For example, using sentiment analysis to identify topic generations.
  - BLEU scores: Identify how well machines are able to translate as humans do the multiple languages proposed in the project.
  - Counterfactual fairness: To evaluate whether the model's predictions would change if sensitive attributes are modified, added, or extracted from the dataset. For example, excluding indigenous histories. (Karzhev, [n.d.](#))
4. Applications of logical reasoning: A study performed by MIT CSAIL demonstrated that using logical reason parameters into their own small pre-trained model, like tagging the word "doctor" as neutral, the small model presented a less stereotyping. (Gordon, [n.d.](#))

Sustainability is also something necessary to consider for ensuring the long-term preservation and reliability of the digital archive and the AI model. Maintenance plans should include regular updates to incorporate new materials, retain the model as required, and establish mechanisms for user feedback. This continuous improvement will help to keep the tool relevant and reliable over time. In a survey performed by Singh et al., [2025](#), the authors propose several sustainable practices to consider when working with AI models. These include energy efficient training, sustainable hardware, life cycle optimizations, sustainable deployment, and end-of-life management. This can be developed by performing tasks like distributing the computational load, using low-power GPU's, enhancing the model to be reusable, using low-resource cloud servers for deployment, and understanding the process of reusing and recycling after the hardware is no longer functional.

Finally, the deployment and design of the graphic user interface should be managed considering at all time user accessibility and ethical integrity. The platform should be available to a wide and diverse audience, not just academic researchers but also everyday online users. This includes providing a reliable multilingual platform, intuitive graphic user interfaces, and resources for people with different levels of digital literacy. With this, the deployed interface aims to make access to the preserved cultural heritage and knowledge easier and more engaging.

By designing a simple and clear graphic interface with minimalist design, users can complete tasks more efficiently. In an article published on Medium, the importance of visual hierarchy and consistency is stated. A consistent interface will also create a better user experience. By prioritizing certain contrasts, font sizes, colors, and placements, the user will feel more in control when accessing the model. For this, it is also important to consider implementing user feedback, in which the GUI and the model can be continuously monitored and refined. Understanding how users interact with the model and the GUI will prevent waste of digital and hardware resources. (“Best Practices for Designing User-Friendly Interfaces for UI/UX Designers”, [2024](#))

## 8 Conclusions

This project aims to successfully propose and create a tool that bridges new technologies with the cultural richness of Latin American communities in Switzerland. By creating an AI-powered archival platform that documents, preserves, and provides access to the histories and materials of Latin American immigrants living in Basel, this project seeks to help migrants *"feel like home away from home"*. Through new, groundbreaking digital technologies, this proposal aims to provide ways for communities to remain connected to their heritage and to make their histories more widely known and accessible.

With the development of large language models and the rise of tools like ChatGPT as mainstream search and learning engines, this project connects technological innovation with traditional cultural preservation techniques. Thanks to the widespread use of GPT-based systems, developers and researchers are now able to find more efficient and accurate ways to handle archival information, from fine-tuning and pre-training LLMs for domain-specific tasks, to developing tooling that enables faster, more relevant access to deeply contextual content. These advancements not only smooth the process of finding and organizing information, but also make the experience more personalized and engaging for

all users.

Today, LLMs transform the way users interact with information, allowing for the creation of specific narratives, informal explorations, and conversational access to historical materials. This offers a new perspective on how history can be told, experienced, and learned, moving beyond traditional libraries and texts to a dynamic, interactive approach where stories are brought to life through context-aware AI.

Throughout the development process, a special emphasis has been placed on ethical responsibility and inclusive practices. Informed consent, age limits, options for anonymity, and strong privacy protections are important aspects of the data collection and management strategy. The platform intends to be built on a diverse array of sources, including official records, oral histories, personal documents, community newsletters, and academic media, all curated and documented with respect for provenance and cultural context. Role access controls, detailed audit performances, and compliance with international data protection laws should further reinforce the trustworthiness and security of the platform.

The technical foundations of the project ensure scalability, sustainability, and robustness. Containerization orchestration, regular model and infrastructure updates, real-time monitoring, and automated backups will allow the platform to grow and adapt to future demands while maintaining its availability and reliability. Comprehensive documentation, including user guides, change logs, manual guides, and data governance policies, ensures transparency and empowers both end users and future contributors.

More broadly, this project stands as a model of interdisciplinary collaboration, demonstrating how developers, researchers, and community members can work together to preserve and share cultural heritage in new and creative ways. While challenges remain, such as evolving privacy expectations and the need for ongoing community participation, the project's design aims to be flexible and responsive, with mechanisms for continued improvement and growth.

As a final conclusion, this archival AI platform not only hopes to save the memories and stories of Latin American communities in Basel, but also opens new ways for research, education, and cultural exchange. It exemplifies how advanced technology can be exploited with empathy, responsibility, and vision to enrich both academia and society, helping migrants, future generations, historians, and anyone interested, truly understand their and other cultures.

## References

- ¿What is Kubernetes? [Section: docs]. (n.d.). Retrieved August 5, 2025, from <https://kubernetes.io/es/docs/concepts/overview/what-is-kubernetes/>
- About CULTURESCAPES. (n.d.). Retrieved July 25, 2025, from <https://culturescapes.ch/en>
- Ahmed, N. A. (n.d.). Understand and Mitigate Bias in LLMs. Retrieved July 27, 2025, from <https://www.datacamp.com/blog/understanding-and-mitigating-bias-in-large-language-models-llms>
- Ahmed, Z. (2024, April). What is FastAPI and its LLM Applications? Retrieved July 29, 2025, from <https://medium.com/@zahmed333/what-is-fastapi-and-its-llm-applications-4b30ae2d43a5>
- Ai on how new and evolving technologies will impact professions. (2024, July). Retrieved July 26, 2025, from <https://oralhistory.org/wp-content/uploads/2024/04/AI-in-OH-Symposium-Program.pdf>
- Alevi Communities In Western Europe: Identity And Religious Strategies. (2011, January). In *Yearbook of Muslims in Europe, Volume 2* (pp. 561–591). BRILL. <https://doi.org/10.1163/ej.9789004184756.i-712.789>
- Al-Gamal, A. A., & Mohammed Ali, E. A. (n.d.). Corpus-based Method in Language Learning and Teaching. Retrieved August 8, 2025, from [https://www.researchgate.net/publication/332876550\\_Corpus-based\\_Method\\_in\\_Language\\_Learning\\_and\\_Teaching](https://www.researchgate.net/publication/332876550_Corpus-based_Method_in_Language_Learning_and_Teaching)
- Algan, Y. (2012). *Cultural Integration of Immigrants in Europe*. OUP Oxford. <https://books.google.ch/books?id=hUFBOx2-gS8C>
- Al-Hasan, T. M., Sayed, A. N., Bensaali, F., Himeur, Y., Varlamis, I., & Dimitrakopoulos, G. (2024). From Traditional Recommender Systems to GPT-Based Chatbots: A Survey of Recent Developments and Future Directions [Publisher: MDPI AG]. *Big Data and Cognitive Computing*, 8(4), 36. <https://doi.org/10.3390/bdcc8040036>
- Augusta Raurica: Experience - Preservation - Research. (n.d.). Retrieved July 25, 2025, from <https://www.augustaurica.ch/en/offers/digital-stories-from-the-ruins>
- Axolotl ai cloud [original-date: 2023-04-14T04:25:47Z]. (2025, July). Retrieved July 26, 2025, from <https://github.com/axolotl-ai-cloud/axolotl>
- Baraka, R. (n.d.). Computer Vision for Cultural Heritage Preservation: Unlocking the Past with Advanced Imaging Technology - Comet. Retrieved July 25, 2025, from

- <https://www.comet.com/site/blog/computer-vision-for-cultural-heritage-preservation-unlocking-the-past-with-advanced-imaging-technology/>
- Basel as cultural capital. (n.d.). Retrieved July 25, 2025, from <https://eurovision-basel.ch/wp-content/uploads/2024/07/Cultural-capital-Basel.pdf>
- Belcic, I. (2024, October). What is RAG (Retrieval Augmented Generation)? | IBM. Retrieved August 8, 2025, from <https://www.ibm.com/think/topics/retrieval-augmented-generation>
- Bergman. (2024, March). What is Fine-Tuning? | IBM. Retrieved August 8, 2025, from <https://www.ibm.com/think/topics/fine-tuning>
- Best Practices for Designing User-Friendly Interfaces for UI/UX Designers. (2024, October). Retrieved July 27, 2025, from <https://medium.com/@uidesign0005/best-practices-for-designing-user-friendly-interfaces-for-ui-ux-designers-0b761c85ce48>
- Bolzman, C. (1997a). Collective identity, associative dynamics and social participation of migrant communities in Switzerland: The search for a local citizenship. *Migraciones. Publicación del Instituto Universitario de Estudios sobre Migraciones*, (2), 75–98. Retrieved May 25, 2025, from <https://revistas.comillas.edu/index.php/revistamigraciones/article/view/4886>
- Bolzman, C. (1997b). Identidad colectiva, dinámica asociativa y participación social de las comunidades migrantes en Suiza: La búsqueda de una ciudadanía local. *Migraciones*, (2), 75–98. Retrieved May 25, 2025, from <https://dialnet.unirioja.es/servlet/articulo?codigo=195528>
- Bolzman, C. (2004). Les migrations latino-américaines dans l'Europe urbaine: Quels enjeux sociaux et éducatifs ? *L'éducation en débats : analyse comparée*, 2, 32–56. Retrieved May 25, 2025, from <https://oap.unige.ch/journals/ed/article/view/447>
- Bolzman, C. (2013). Reflexions sobre la perspectiva intercultural a partir de la figura de l'estranger. *Educació social. Revista d'intervenció socioeducativa*, (54), 49–60. <https://doi.org/10.34810/EducacioSocialn54id267191>
- Bolzman, C., Fibbi, R., & Vial, M. (1999). Modes of social and occupational insert, practices cultural identity and possessions. The example of young adults of Spanish and Italian origin in Switzerland. *Migraciones. Publicación del Instituto Universitario de Estudios sobre Migraciones*, (6), 61–84. Retrieved May 25, 2025, from <https://revistas.comillas.edu/index.php/revistamigraciones/article/view/4444>
- Caballar. (2025, May). What Is Model Deployment? | IBM. Retrieved August 8, 2025, from <https://www.ibm.com/think/topics/model-deployment>

- Cain, C., & Haque, S. (n.d.). Organizational Workflow and Its Impact on Work Quality. In *ResearchGate*. Retrieved August 8, 2025, from [https://www.researchgate.net/publication/49843267\\_Organizational\\_Workflow\\_and\\_Its\\_Impact\\_on\\_Work\\_Quality](https://www.researchgate.net/publication/49843267_Organizational_Workflow_and_Its_Impact_on_Work_Quality)
- Chakrabarti, K. C. (2024, July). The Role of AI in Cultural Preservation and Heritage [Section: AI]. Retrieved July 25, 2025, from <https://itmunch.com/the-role-of-ai-in-cultural-preservation-and-heritage/>
- Cultural Capital of Switzerland. (2024, April). Retrieved July 25, 2025, from <https://www.bs.ch/en/schwerpunkte/portrait/arts>
- D'Amato, G. (n.d.). Swiss Federalism and its Impact on Integration Policies.
- Data Archive for AI | Significance, Benefits, and Use cases. (2024, January). Retrieved July 26, 2025, from <https://platform3solutions.com/blog/how-to-harness-the-power-of-data-archive-for-ai/>
- dependabot. (n.d.). GitHub - dependabot/dependabot-core: Dependabot's core logic for creating update PRs. Retrieved August 7, 2025, from <https://github.com/dependabot/dependabot-core>
- Derclaye, E. (2002). What is a Database?: *A Critical Analysis of the Definition of a Database in the European Database Directive and Suggestions for an International Definition* [Publisher: Wiley]. *The Journal of World Intellectual Property*, 5(6), 981–1011. <https://doi.org/10.1111/j.1747-1796.2002.tb00189.x>
- DeVore, V. (2016, November). Grassroots projects bring locals and immigrants together. [https://www.swissinfo.ch/eng/society/part-of-society\\_grassroots-projects-bring-locals-and-immigrants-together/42596288](https://www.swissinfo.ch/eng/society/part-of-society_grassroots-projects-bring-locals-and-immigrants-together/42596288)
- Diaz Castillo, R. (n.d.). AFIRMACION DE LA IDENTIDAD CULTURAL EN AMERICA LATINA, PRESERVACION Y DESARROLLO DE LOS VALORES QUE LA COMPONENTEN.
- Duc-Quang, N. (n.d.). Defining the 25% foreign population in Switzerland. [https://www.swissinfo.ch/eng/society/migration-series-part-1-\\_who-are-the-25-foreign-population-in-switzerland/42412156](https://www.swissinfo.ch/eng/society/migration-series-part-1-_who-are-the-25-foreign-population-in-switzerland/42412156)
- Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review. (2009). Retrieved August 8, 2025, from <https://www.tenk.fi/sites/tenk.fi/files/ethicalprinciples.pdf>

- European Parliament. Directorate General for Internal Policies of the Union. (2017). *European identity: Research for CULT Committee*. Publications Office. Retrieved May 25, 2025, from <https://data.europa.eu/doi/10.2861/70563>
- Flask Documentation (3.1.x). (n.d.). Retrieved July 30, 2025, from <https://flask.palletsprojects.com/en/stable/>
- Four out of ten Swiss residents have migration background. (n.d.). <https://www.swissinfo.ch/eng/society/four-out-of-ten-swiss-residents-have-migration-background/48955814>
- Ganesan, K. (2023, March). What went wrong with Tay, the Twitter bot that turned racist? Retrieved August 15, 2025, from <https://www.opinosis-analytics.com/blog/tay-twitter-bot/>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024, March). Retrieval-Augmented Generation for Large Language Models: A Survey [arXiv:2312.10997 [cs]]. <https://doi.org/10.48550/arXiv.2312.10997>
- Geschichte | GGG Migration. (n.d.). Retrieved July 25, 2025, from <https://www.ggg-migration.ch/geschichte/>
- Ghosh, M., & Arunachalam. (n.d.). Introduction to Artificial Intelligence. In *ResearchGate*. [https://doi.org/10.1007/978-981-16-0415-7\\_2](https://doi.org/10.1007/978-981-16-0415-7_2)
- Gordon, R. (n.d.). Large language models are biased. Can logic help save them? | MIT CSAIL. Retrieved July 27, 2025, from <https://www.csail.mit.edu/news/large-language-models-are-biased-can-logic-help-save-them>
- Grimm, T. (2009, April). IDENTIFYING AND PRESERVING THE HISTORY OF THE LATINO VISUAL ARTS: SURVEY OF ARCHIVAL INITIATIVES AND RECOMMENDATIONS. Retrieved July 25, 2025, from [https://www.chicano.ucla.edu/files/crr\\_06April2005.pdf](https://www.chicano.ucla.edu/files/crr_06April2005.pdf)
- Haas, M. (n.d.). Cultural pluralism | EBSCO Research Starters. Retrieved July 25, 2025, from <https://www.ebsco.com/research-starters/religion-and-philosophy/cultural-pluralism>
- Hagmann. (2017, April). Magnet Basel - was Fremdenpolizeiakten erzählen. Retrieved May 25, 2025, from <https://blog.staatsarchiv-bs.ch/magnet-basel-fremdenpolizeiakten-erzaehlen/>
- Hagmann, D. (2017, May). Von unschätzbarem Wert: Fremdenpolizeiakten im Staatsarchiv. <https://blog.staatsarchiv-bs.ch/von-unschaetzbarem-wert-fremdenpolizeiakten-im-staatsarchiv/>

- How to use Bootstrap with Flask [Section: Developer Tools]. (2021, October). Retrieved July 30, 2025, from <https://learningactors.com/how-to-use-bootstrap-with-flask/>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions [arXiv:2311.05232 [cs]]. *ACM Transactions on Information Systems*, 43(2), 1–55. <https://doi.org/10.1145/3703155>
- Hugo, G. (2025). Migrants in society: Diversity and cohesion.
- Ibrahim, M. (n.d.). AI for Art & Heritage Conservation | Ultralytics. Retrieved July 25, 2025, from <https://www.ultralytics.com/blog/ai-in-art-and-cultural-heritage-conservation>
- Jacquez, F., Vaughn, L. M., & Hardy-Besaw, J. (2024). Immigrant Perspectives of Social Connection in a Nontraditional Migration Area. *Healthcare*, 12(6), 686. <https://doi.org/10.3390/healthcare12060686>
- Jain, S. (2025, February). Transfer Learning: LLM generalization to similar problems. Retrieved July 27, 2025, from <https://medium.com/@sulbha.jindal/transfer-learning-llm-generalization-to-similar-problems-1d3b2bf28c6e>
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning [arXiv:2104.05314 [cs]]. *Electronic Markets*, 31(3), 685–695. <https://doi.org/10.1007/s12525-021-00475-2>
- Jenkins, J., & Bustos, A. (2022). *Patrimonio Efímero. Memorias, cultura popular y vida cotidiana*. El Colegio de San Luis. [https://books.google.ch/books?id=Dzp\\_EAAAQBAJ](https://books.google.ch/books?id=Dzp_EAAAQBAJ)
- Jossen, M. (2018). *Undocumented Migrants and Healthcare: Eight Stories from Switzerland*. Open Book Publishers. <https://books.google.ch/books?id=XM5dDwAAQBAJ>
- Karzhev, S. (n.d.). LLM Evaluation: Metrics, Methodologies, Best Practices. Retrieved July 27, 2025, from <https://www.datacamp.com/blog/llm-evaluation>
- Kerner, S. M. (n.d.). GenAI search vs. traditional search engines: How they differ. Retrieved July 26, 2025, from <https://www.techtarget.com/whatis/feature/GenAI-search-vs-traditional-search-engines-How-they-differ>
- Khandelwal, P. (2025, April). Understanding Prompt Engineering and Its Different Types. Retrieved July 25, 2025, from <https://dev.to/piyushpk/what-is-prompt-engineering-and-types-of-it--5blm>

- Kivindy, a. (2023, December). What are Generative Pre-trained Transformers (GPTs)? Retrieved August 8, 2025, from <https://medium.com/@anitakivindy/what-are-generative-pre-trained-transformers-gpts-b37a8ad94400>
- La tienda latina basel. (n.d.).
- Leonard, D., & Lugo-Lugo, C. (2015). *Latino History and Culture: An Encyclopedia*. Taylor & Francis. <https://books.google.ch/books?id=FPBnBwAAQBAJ>
- Lester, B., Al-Rfou, R., & Constant, N. (2021, September). The Power of Scale for Parameter-Efficient Prompt Tuning [arXiv:2104.08691 [cs]]. <https://doi.org/10.48550/arXiv.2104.08691>
- Leybold-Johnson, I. (2017a, June). Basel police archives reveal lives of past city residents. Retrieved July 25, 2025, from [https://www.swissinfo.ch/eng/society/immigration-stories\\_basel-police-archives-reveal-lives-of-past-city-residents/43258236](https://www.swissinfo.ch/eng/society/immigration-stories_basel-police-archives-reveal-lives-of-past-city-residents/43258236)
- Leybold-Johnson, I. (2017b, June). Basel police archives reveal lives of past city residents. [https://www.swissinfo.ch/eng/society/immigration-stories\\_basel-police-archives-reveal-lives-of-past-city-residents/43258236](https://www.swissinfo.ch/eng/society/immigration-stories_basel-police-archives-reveal-lives-of-past-city-residents/43258236)
- Maaithili, B., & Phil, M. (n.d.). Heritage conservation & preservation through AI & NLP [Accessed: 2025-8-15].
- Mathari, A. (2024, April). What cities like Geneva are doing to welcome migrants and refugees. Retrieved August 15, 2025, from <https://www.swissinfo.ch/eng/international-geneva/what-cities-are-doing-to-welcome-migrants-and-refugees/75107116>
- Metadata. (n.d.). Retrieved August 8, 2025, from [https://www.cdema.org/virtuallibrary/images/atdmgm\\_geonode\\_manual\\_Metadata.pdf](https://www.cdema.org/virtuallibrary/images/atdmgm_geonode_manual_Metadata.pdf)
- Migrant integration. (2020, September). Retrieved July 25, 2025, from <https://www.migrationdataportal.org/themes/migrant-integration>
- Moez, A. (n.d.). Mastering Low-Rank Adaptation (LoRA): Enhancing Large Language Models for Efficient Adaptation. Retrieved July 26, 2025, from <https://www.datacamp.com/tutorial/mastering-low-rank-adaptation-lora-enhancing-large-language-models-for-efficient-adaptation>
- Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024, March). Having Beer after Prayer? Measuring Cultural Bias in Large Language Models [arXiv:2305.14456 [cs]]. <https://doi.org/10.48550/arXiv.2305.14456>
- Nosotras basel. (n.d.). Retrieved August 14, 2025, from <http://www.nosotrasbasel.org/ueber-uns-de.html>

- Nuestra Identidad | Misión católica de lengua española de basilea. (n.d.). Retrieved July 25, 2025, from <https://www.misiondebasilea.ch/nuestra-identidad>
- Olla Común. (n.d.). <https://www.k5kurszentrum.ch/events/olla-comun/>
- Overview on Kubernetes. (n.d.). Retrieved August 5, 2025, from <https://kubernetes.io/docs/concepts/overview/>
- Pajo, P. (2025, April). Vector Embeddings Unveiled: A Comprehensive Exploration of Their Creation, Types, Applications, Challenges, and Future Directions in Machine Learning. <https://doi.org/10.13140/RG.2.2.15544.05129>
- Patton, S. (2024, October). AI Meets Archives: The Future of Machine Learning in Cultural Heritage. Retrieved July 26, 2025, from <https://www.clir.org/2024/10/ai-meets-archives-the-future-of-machine-learning-in-cultural-heritage/>
- Prados-Peña, M. B., Pavlidis, G., & García-López, A. (2025). New technologies for the conservation and preservation of cultural heritage through a bibliometric analysis [Publisher: Emerald]. *Journal of Cultural Heritage Management and Sustainable Development*, 15(3), 664–686. <https://doi.org/10.1108/jchmsd-07-2022-0124>
- Prutsch, M. J. (n.d.). Investigación para la Comisión CULT - Identidad europea.
- Richardson, J. (2024, June). Artificial Intelligence Museum Audio Guide: AI Revolution at the Smithsonian American Art Museum. Retrieved August 15, 2025, from <https://www.museumnext.com/article/artificial-intelligence-museum-audio-guide-ai-revolution-at-the-smithsonian-american-art-museum/>
- Roche, M. (n.d.). Historical Research and Archival Sources. Retrieved July 8, 2025, from <https://www.sjsu.edu/people/kathrine.richardson/courses/Geog145/s1/Chapter-9---Historical-Research-and-Archival-Sources.pdf>
- Sachsen Gessaphe, K. (2011). Kulturgüterschutz und politische Entwicklung in Mexiko im Jubiläumsjahr 2010: Jahreskongress 2010 der Deutsch-Mexikanischen Juristenvereinigung e.V. in Zusammenarbeit mit dem Ibero-Amerikanischen Institut Preußischer Kulturbesitz ; [14. September 2010 in Berlin]. *Ibero-Online.de*, 11. Retrieved May 25, 2025, from [https://publications.iai.spk-berlin.de/receive/riai\\_mods\\_00000434](https://publications.iai.spk-berlin.de/receive/riai_mods_00000434)
- Sánchez-Miranda, N. A., Julca, M. R., Rosas-Prado, C. E., & Cerna, J. M. R. (2022). Conservación y preservación del Patrimonio Cultural: Una revisión a partir de la identidad latinoamericana: *Revista de Filosofía*, 39, 157–168. <https://doi.org/10.5281/zenodo.7297801>

- Santopaolo, G. P. (2025, January). Understanding Key Hyperparameters When Fine-Tuning an LLM. Retrieved July 26, 2025, from <https://genmind.ch/posts/understanding-key-hyperparameters-when-fine-tuning-an-llm/>
- Sanz, N., Arce, V., Manuel, J., Mexico, U., & de la Frontera Norte, C. (2016). *Migración y cultura*. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina en México. <https://books.google.ch/books?id=Zk28DgAAQBAJ>
- Sekar, A. (2024, August). An implementation of LLMs and Gen AI utilizing a Flask REST API. Retrieved July 29, 2025, from <https://medium.com/@anirudhsekar2008/an-implementation-of-llms-and-gen-ai-utilizing-a-flask-rest-api-677769db8224>
- Serrano, I., Fernández García, M., Ordoñez, Á., Bajo Marcos, E., & Miguel Somavilla, S. (n.d.). The integration of migrants. Retrieved May 25, 2025, from <https://www.immerse-h2020.eu/publications/>
- Singh, A., Patel, N. P., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025, January). A Survey of Sustainability in Large Language Models: Applications, Economics, and Challenges [arXiv:2412.04782 [cs]]. <https://doi.org/10.48550/arXiv.2412.04782>
- Skublewska-Paszowska, M., Milosz, M., Powroznik, P., & Lukasik, E. (2022). 3D technologies for intangible cultural heritage preservation—literature review for selected databases. *Heritage Science*, 10(1), 3. <https://doi.org/10.1186/s40494-021-00633-x>
- Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S., & Mansoor, W. (2023). Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions [arXiv:2307.14107 [cs]]. *Journal of King Saud University - Computer and Information Sciences*, 35(8), 101675. <https://doi.org/10.1016/j.jksuci.2023.101675>
- Soum, P. (n.d.). A Definitive Guide to Fine-Tuning LLMs Using Axolotl and Llama-Factory - Superteams.ai. Retrieved July 25, 2025, from <https://www.superteams.ai/blog/a-definitive-guide-to-fine-tuning-llms-using-axolotl-and-llama-factory>
- Sound, F. P. +. (2018, December). Immersive Multimedia Show in Roman Colosseum. Retrieved July 25, 2025, from <https://prolight-sound-blog.com/immersive-multimedia-roman-colosseum/>
- Spatharioti, S. E., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025). Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3706598.3714082>

- Strategies for Fine-Tuning Large Language Models. (n.d.). Retrieved July 26, 2025, from [https://www.capellasolutions.com/blog/strategies-for-fine-tuning-large-language-models?utm\\_source=chatgpt.com](https://www.capellasolutions.com/blog/strategies-for-fine-tuning-large-language-models?utm_source=chatgpt.com)
- Stünzi, R., Fibbi, R., & D'Amato, G. (2025). Uneven Pathways to Local Power: The Political Incorporation of Immigrants' Descendants. *Politics and Governance*, 13, 9293. <https://doi.org/10.17645/pag.9293>
- Susnjara, S., & Smalley, I. (2023, August). What Is Network Attached Storage (NAS)? | IBM. Retrieved August 7, 2025, from <https://www.ibm.com/think/topics/network-attached-storage>
- Swarm mode. (800). Retrieved August 8, 2025, from <https://docs.docker.com/engine/swarm/>
- swissinfo.ch, S. W. I. (2023, November). Four out of ten Swiss residents have migration background. Retrieved May 25, 2025, from <https://www.swissinfo.ch/eng/society/four-out-of-ten-swiss-residents-have-migration-background/48955814>
- SWR Landesschau Rheinland-Pfalz. (2023, May). Erstaunlich! Roboter meißelt Skulptur. Retrieved August 15, 2025, from <https://www.youtube.com/watch?v=yICSiZzTqXk>
- Vargas, L. (2009). *Latina Teens, Migration, and Popular Culture*. Peter Lang Pub. <https://books.google.ch/books?id=z1AphAiJe3AC>
- Vaštakas, L., Hanscam, E., Einsiedler, J., & Alissandrakis, A. (n.d.). Cultural Heritage Search with Large Language Models.
- Venkatesh, P., Ahtsham, Z., Aafaq, K., & Arsalan, S. (n.d.). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities (Version 1.0). Retrieved July 25, 2025, from <https://arxiv.org/html/2408.13296v1#Ch2.S1>
- What are the GDPR consent requirements? [Section: News & Updates]. (2019, January). Retrieved August 11, 2025, from <https://gdpr.eu/gdpr-consent-requirements/>
- What is a Container? | Docker. (n.d.). Retrieved August 8, 2025, from <https://www.docker.com/resources/what-container/>
- What is Data Anonymization | Pros, Cons & Common Techniques | Imperva. (n.d.). Retrieved July 26, 2025, from <https://www.imperva.com/learn/data-security/anonymization/>

- What personal data is considered sensitive? - European Commission. (n.d.). Retrieved August 15, 2025, from [https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive\\_en](https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en)
- Wilde, J. (n.d.). Cultural Barriers in Immigration: Can Immigrants Maintain Their Identity? <https://jinheewilde.co/cultural-barriers-in-immigration-can-immigrants-maintain-their-identity/>
- Wilde, J. (2025, February). Cultural barriers in immigration: Can immigrants maintain their identity? [Accessed: 2025-8-15].
- Yuri, K. (n.d.). Tradition meets AI in Nishijinori weaving style from Japan's ancient capital. Retrieved July 26, 2025, from <https://abcnews.go.com/Technology/wireStory/tradition-meets-ai-nishijinori-weaving-style-japans-ancient-124061289>
- Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2024, December). Instruction Tuning for Large Language Models: A Survey [arXiv:2308.10792 [cs]]. <https://doi.org/10.48550/arXiv.2308.10792>
- Zhang, S., Peng, S., & Hou, J. (n.d.). Archives Meet GPT: A Pilot Study on Enhancing Archival Workflows with Large Language Models.
- Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., & Gonzalez, J. E. (2024, June). RAFT: Adapting Language Model to Domain Specific RAG [arXiv:2403.10131 [cs]]. <https://doi.org/10.48550/arXiv.2403.10131>
- Zhang, Z., Cao, H., Wang, X., Zhang, Y., & Sheng, Q. Z. (2024). Ontology-Driven Archival Knowledge Graph Construction Leveraging Large Language Models.

# A Appendix

## A.1 Example of Consent Form:

### **Informed Consent Form for Participation in the Latin American Heritage Archival Research Project**

Project Title:

Latin American Heritage Archival AI GPT Model

Principal Investigator(s): [Name(s)], [Institution/Affiliation]

Contact: [Email], [Phone number]

#### **Purpose of the Project**

You are invited to participate in a project that aims to document, preserve, and make accessible the archival materials and histories of the Latin American communities in Basel and Switzerland. Your participation will help build an AI-powered tool to facilitate access to these valuable cultural records.

#### **Description of Participation**

If you agree to participate, you will be asked to:

- Provide an oral history interview, personal documents (e.g., letters, diaries, photographs), or other materials relevant to the project.
- Share information about your experiences, family history, and new community.

#### **Use of Data**

- Your contributions (interviews, documents, images, transcripts) will be digitized and stored securely.
- Data will be used to train and improve an AI tool to provide contextualized access to historical records.
- Your data may be combined with other archival sources, but personal identifiers can be anonymized upon your request.
- Access to sensitive data will be restricted and controlled to protect your privacy.

#### **Confidentiality and Privacy**

You may choose to:

- Use a pseudonym or remain anonymous in the records and AI outputs.
- Request redaction of specific personal details.
- The research team will safeguard your data using encryption and access controls.
- Your personal data will not be shared with third parties except as required by law or with your explicit permission.

**Voluntary Participation and Withdrawal**

- Your participation is completely voluntary.
- You may refuse to answer any question or decline to provide any materials.
- You have the right to withdraw your participation and request removal of your data from the project at any time, without any negative consequences.

**Risks and Benefits**

- There are minimal risks associated with participation. Some topics might be sensitive or personal—please let the interviewer know if you feel uncomfortable at any time.
- Benefits include contributing to the preservation of important community histories and helping develop public resources for future generations.

**Consent**

By signing below, you indicate that:

- You have read and understood this consent form.
- Your questions have been answered satisfactorily.
- You voluntarily agree to participate and provide consent for the use of your materials as described.

Participant Name: \_\_\_\_\_

Date: \_\_\_\_\_

Signature: \_\_\_\_\_

If you are under 18 years old please provide your:

Legal Guardian or Parent Name: \_\_\_\_\_

Guardian Signature: \_\_\_\_\_

Date: \_\_\_\_\_

**Contact for Questions or Complaints**

If you have questions about the project or your rights as a participant, please contact:

[Principal Investigator Name]

[Email Address]

[Phone Number]