Proposal: Leveraging Nomic AI's Atlas Platform for Enhanced Classification and Categorization of Research Papers in AGIE

AGIE Project Team

2025-02-01

Table of contents

1 Objective **2**

2 Technical Implementation **2**
2.1 Nomic Embed for Dimensionality Reduction & Semantic Vectorization 2
2.2 Atlas' Neural Search for Cross-Paper Concept Retrieval 2
2.3 Dynamic Knowledge Graph for Interdisciplinary Relationship Mapping 2

3 Workflow Integration **3**
3.1 CSV/PDF Ingestion Pipeline 3
3.2 Multi-Modal Handling (Text, Figures, Equations) 3

4 Expected Outcomes **3**

5 Technical Advantages Highlight **4**
5.1 Open-Source Deepscatter Visualizations 4
5.2 SOC 2-Certified Data Security 4

6 Concrete Examples from Atlas' Wikipedia Biographies & IDEFICS Dataset Analysis **5**

7 Conclusion & Next Steps **6**

1 Objective

To **streamline literature reviews** and **uncover hidden insights** across large, interdisciplinary collections of research papers in AGIE by employing Nomic AI's Atlas platform. This approach will use **AI-driven semantic organization** and **pattern discovery** to systematically **identify trends**, **detect biases**, and **categorize research** with greater accuracy and speed.

---

2 Technical Implementation

2.1 Nomic Embed for Dimensionality Reduction & Semantic Vectorization

- **Contextual Embeddings**: Each paper (or section) is transformed into a high-dimensional vector using Nomic Embed. The embeddings capture nuances in language, domain-specific terminologies, and contextual relationships.

- **Dimensionality Reduction**: These high-dimensional vectors are mapped into a lower-dimensional space for efficient clustering and visualization, enabling **semantic-based** (rather than keyword-based) grouping of research topics such as mentorship frameworks, leadership barriers, or hiring bias studies.

2.2 Atlas' Neural Search for Cross-Paper Concept Retrieval

- **Intelligent Querying**: Researchers can submit text queries (e.g., "mentorship programs for mid-career women in STEM"), and Atlas will retrieve semantically similar papers—even if the keywords differ—ensuring robust discovery of relevant literature across interdisciplinary domains.

- **Scalable Indexing**: The underlying neural search scales seamlessly from **hundreds to tens of millions** of research entries, accommodating the growing corpus of AGIE's national repository.

2.3 Dynamic Knowledge Graph for Interdisciplinary Relationship Mapping

- **Automated Graph Construction**: As papers are ingested, Atlas automatically builds a knowledge graph linking key concepts, methodologies, and authors. Clusters form around themes such as "leadership advancement" or "bias in funding allocation."

- **Live Updates**: Newly added or revised content in the AGIE repository automatically updates the graph, ensuring an up-to-date map of interdisciplinary relationships and **emerging research** hotspots (e.g., evolving definitions of "belongingness").

---

3 Workflow Integration

3.1 CSV/PDF Ingestion Pipeline

- **Batch Upload & Parsing**: A pipeline enabling direct ingestion of CSV summaries and PDF manuscripts from AGIE participants. Atlas parses each document, extracts relevant text sections, figures, and references, and generates embeddings via Nomic Embed.

- **Metadata Enrichment**: Paper titles, authors, publication venues, and domain tags are appended as metadata, facilitating easy filtering (e.g., region, research design, or NIH institute affiliation).

3.2 Multi-Modal Handling (Text, Figures, Equations)

- **OCR & Image Processing**: For PDFs containing images, charts, or figures on gender equity trends, Atlas can apply optical character recognition (OCR) to embed descriptive text.

- **Equations & Structural Data**: Research on statistical models or theoretical frameworks can be preserved for advanced similarity analyses.

---

4 Expected Outcomes

1. **Quantitative Time Reduction in Literature Screening**
    - **Reference (SmarterX Case Study)**: SmarterX reported **hundreds of hours saved** by consolidating their data infrastructure into Atlas. Similarly, AGIE stakeholders can expect **significant reductions** in manual screening time, freeing researchers to focus on deeper analysis.
2. **Improved Categorization Accuracy via Cluster Analysis**

- Utilizing Atlas' **semantic clustering** will surface groups of papers by conceptual similarity rather than superficial keyword overlaps, **raising classification accuracy** and reducing mislabeled or overlooked studies.

3. **Identification of Emerging Research Trends**

   - By continuously updating the knowledge graph, Atlas will detect **new patterns**, gaps in the literature, and **emerging subfields** (e.g., intersectional studies on female faculty experiences in multiple underrepresented groups), aiding AGIE's forward-looking research agenda.

---

5 Technical Advantages Highlight

5.1 Open-Source Deepscatter Visualizations

- **High-Resolution Cluster Exploration**: Deepscatter provides interactive 2D/3D scatterplots of embedded papers, enabling intuitive exploration of **cluster formations** (e.g., mentorship interventions vs. institutional policy reforms).

- **Customizable**: AGIE analysts can color-code by domain, publication year, or author, facilitating meaningful at-a-glance analyses of the research landscape.

5.2 SOC 2-Certified Data Security

- **Confidentiality & Integrity**: Atlas meets SOC 2 Type II standards, ensuring secure handling of sensitive research findings (e.g., unpublished institutional data or proprietary interventions).

- **Access Controls**: Fine-grained permission settings enable safe data sharing across NIH institutes, academic centers, and other AGIE partners without risking unauthorized disclosure.

---

6 Concrete Examples from Atlas' Wikipedia Biographies & IDEFICS Dataset Analysis

1. **Systematic Error Detection (Wikipedia Biographies)**
    - *Clustering by Semantics*: In a large-scale analysis of Wikipedia biographies, Atlas automatically grouped subjects under topical clusters such as "pay equity," leadership advancement," and "Retention."
    - *Identifying Data Gaps*: On inspection, certain groups (e.g., women in sports or minority politicians) were **underrepresented** or misclassified, revealing **systematic bias**.
    - **AGIE Application**: Similarly, for AGIE's repository, Atlas can surface classification discrepancies (e.g., an under-clustered set of studies on intersectional barriers), prompting further validation and **bias mitigation** strategies.
2. **Bias Mitigation (IDEFICS Dataset)**
    - *High-Loss Clusters*: In the IDEFICS training data, Atlas identified clusters of low-relevance or semantically confusing articles (e.g., random "poetry" or "government" text) that contributed to **model inaccuracies**.
    - **Refinement Cycle**: Researchers used these insights to **exclude or correct** problematic data points, improving overall model performance.
    - **AGIE Application**: For classifying research on gender equity, Atlas can highlight **outlier clusters** (papers mislabeled or incongruent with AGIE's scope), allowing curators to **reclassify or remove** spurious content.
3. **Ensuring Domain Alignment**
    - In a broader dataset, Atlas flagged research references that were thematically distant from the corpus (e.g., purely technical AI methods without direct equity focus).
    - **AGIE Application**: Similarly, borderline topics (e.g., purely epidemiological studies with minimal gender-specific analysis) can be pinpointed for additional curation or re-tagging.

---

7 Conclusion & Next Steps

By deploying Nomic AI's Atlas platform within the AGIE initiative, stakeholders can realize **significant efficiency gains** in literature screening, **improved categorization accuracy**, and **early detection of new trends** in gender equity research. The **SOC 2-certified infrastructure**, and **open-source visualization tools** ensure secure, transparent, and cutting-edge capabilities that bolster AGIE's mission to **amplify solutions for advancing gender equity** in academic and clinical settings.