Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.8319
This work is licensed under a Creative Commons Attribution 4.0 International License.

TALENT: A Pipeline for Aligning Theses and Dissertations with Strategic Themes Framework to Evaluate National Installed Human Capacity – An Empirical Evaluation on AI in Brazil

Marcelo Augusto Santos Turine [**Federal University of Mato Grosso do Sul** | ***marcelo.turine@ufms.br***]
Solange Oliveira Rezende [**University of Sao Paulo** | ***solange@icmc.usp.br***]
Flávia Cristina Bernardini [**Fluminense Federal University** | ***fcbernardini@ic.uff.br***]
Marcos Paulo Silva Gôlo [**University of Sao Paulo** | ***marcosgolo@usp.br***]
Ricardo Marcondes Marcacini [**University of Sao Paulo** | ***ricardo.marcacini@icmc.usp.br***]

Faculty of Computing, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva, s/n, Bairro Universitário, Campo Grande, MS, 79070-900, Brazil.

Received: 27 May 2026 • **Accepted:** 04 November 2026 • **Published:** 21 November 2026

Abstract. Geopolitical powers, including the United States, the European Union, and China, have established strategic Artificial Intelligence (AI) roadmaps aligned with their respective visions of technological sovereignty and digital competitiveness. A critical determinant of this sovereignty is the installed human capacity required to support advanced AI research, innovation, and workforce development. However, systematically measuring topic-specific research capacity across large nations or macro-regions remains challenging. To address this gap, we introduce **TALENT**, a specialized AI-driven data pipeline designed for the large-scale evaluation and strategic alignment of academic theses and dissertations to map national installed capacity across priority research domains. Combining Natural Language Processing (NLP), semantic embeddings, Large Language Models (LLMs), and human curation, TALENT extracts and semantically analyzes metadata and textual content from master's dissertations and doctoral theses, aligns them with strategic thematic frameworks, and generates granular capacity mappings aggregated by macro-region, state, university, and graduate program. We evaluate TALENT through a large-scale empirical study of the Brazilian National Graduate System (SNPG), using the CAPES National Agenda for Advanced Human Capital Development and the CAPES Sucupira database of theses and dissertations. The evaluation examines both the reliability of the proposed semantic alignment pipeline and its ability to support nationwide analysis of human resource development in Artificial Intelligence. The validation procedure combines AI-assisted assessment with expert-supervised review, providing reference decisions for assessing the quality of thematic mappings while preserving human oversight of ambiguous cases. Beyond methodological validation, the nationwide application enables us to investigate how AI-related expertise is distributed across institutions, regions, graduate programs, and disciplinary areas within the Brazilian SNPG. Our findings reveal that a substantial portion of AI-related graduate education occurs outside traditional Computer Science disciplines, highlighting the growing interdisciplinary diffusion of AI across multiple knowledge areas. Overall, the empirical results demonstrate the potential of TALENT as a scalable analytical framework for mapping national research capacity and illustrate how AI-assisted governance, combined with human validation, can support evidence-based planning, funding policies, and strategic human capital development in emerging economies.

Keywords: Human-Installed Capacity, Natural Language Processing, Large Language Models, Strategic Themes

1 Introduction

Artificial Intelligence (AI) has emerged globally as a strategic technological asset, an instrument of international trade, and a core driver of geopolitical competition [Truby *et al*., 2026]. Currently, three dominant approaches shape the global AI landscape: the United States' market-driven model fueled by venture capital and tech giants [The White House, 2023]; China's state-led techno-nationalism built on massive data centralization and public-private synergy [Choi and Yoon, 2025]; and the European Union's regulatory diplomacy, which prioritizes risk-based, human-centric legal frameworks [European Union, 2024]. For emerging economies, navigating this polarized arena requires strategic pathways to reduce asymmetric digital dependencies and avoid becoming mere targets for data extraction [EIT Digital, 2021]. The Brazilian Artificial Intelligence Plan (PBIA), "AI for Good", represents one such blueprint, blending EU-style ethical guardrails with state-guided goals to address regional inequalities, optimize public services, and establish sovereign digital infrastructure [MCTI, 2024].

This challenging scenario has drawn many countries to pursue AI sovereignty. In Brazil, it has gained even greater urgency with the approval of Brazil's National Strategy for Science, Technology, and Innovation (ENCTI 2024–2034), which establishes a ten-year strategic framework to strengthen scientific sovereignty, digital transformation, sustainable development, reindustrialization, and regional competitiveness. Among its structural priorities, ENCTI explicitly recognizes AI, advanced digital infrastructures, data-driven science, and the education of highly qualified human resources as criti-

cal capabilities for national transformation. In this context, training master's and doctoral students in AI-related fields is no longer merely an academic objective: it is a strategic state asset directly tied to technological autonomy, capacity for innovation, and Brazil's competitive integration into the global knowledge economy. Moreover, the growing complexity of global technological competition and the accelerated diffusion of AI systems for governments, industries, and institutions reinforce the need for evidence-based public policies that can identify, monitor, and strategically guide national capabilities for advanced human resource development [MCTI, 2026].

Nevertheless, achieving true technological sovereignty fundamentally relies on institutional capacity to cultivate advanced human capital capable of building and managing local data chains and (AI) algorithms [EIT Digital, 2021; Valiati and Möller, 2025]. Although strategic policy documents such as the PBIA and ENCTI in Brazil acknowledge a critical deficit in specialized AI talent, they lack operational mechanisms to measure this installed capability across a vast territory. A literature review reveals a significant gap: standardized methods and indicators to quantify localized AI research capacity in developing nations remain scarce. Consequently, understanding the exact installed capacity of the Brazilian National Graduate System (SNPG) in AI domains is essential, not only for academic planning but also for supporting long-term innovations, digital governance, economic resilience, technological sovereignty, and sustainable development.

This challenge is directly connected to the CAPES National Agenda[1] initiative, which identified strategic themes to guide the training of master's and doctoral students across Brazilian federative units, based on input from academia, industry, government, and civil society. Because these priorities include AI, this context raises a central research question: ***how can we automatically analyze the large volume of SNPG master's dissertations and doctoral theses and semantically associate each metadata record with strategic themes to generate indicators with explainable and auditable results?***

To bridge this gap, this paper introduces **TALENT** (**T**heses **A**nalysis & **L**ifecycle **E**valuation for **N**ational Capacity **T**racking), a specialized data pipeline artifact designed to evaluate the lifecycle of academic theses and dissertations and strategically align them, thereby systematically mapping national installed capacity. Our premises for constructing this artifact rest upon three core pillars: (i) textual data and associated institutional metadata from master's dissertations and doctoral theses can be systematically processed by language models to identify granular research topics precisely; (ii) macro-region geopolitical territories (in this study, Brazil represents the macro-region, subdivided into its states, or regions) possess a strategic thematic priority framework, composed of predefined sets of strategic themes of its regions; and (iii) the domain of "Artificial Intelligence" can be dynamically aligned with specific elements of these strategic priority sets through advanced language modeling. By leveraging Natural Language Processing (NLP) techniques and Large Language Models (LLMs), TALENT processes large-scale textual evidence from graduate outputs and aligns it with a strategic thematic priority framework. Our pipeline enables a granular, evidence-based assessment of how topic-specific human resource development is distributed across macro-regions, regions, universities, and individual postgraduate programs, transforming raw academic data into an auditable knowledge base for public policy and strategic planning.

To achieve this goal, we executed the following methodological steps: (i) conducted a literature review to determine the state-of-the-art in assessing human capacity within Artificial Intelligence; (ii) performed a targeted literature review on semantic affinity detection between strategic themes and graduate profiles; (iii) designed the TALENT pipeline tailored to the operational constraints of global AI sovereignty; (iv) conducted an empirical evaluation of TALENT through the systematic extraction and lifecycle analysis of data from academic theses and dissertations defended and registered across the SNPG for the years 2023 and 2024, thereby assessing the country's actual structural readiness for AI technological autonomy; and (v) conducted an ablation study to assess the contribution and operational role of the components of TALENT. Our main contributions are in three key aspects:

1. **The TALENT Pipeline Artifact:** The design, architecture, and implementation of an end-to-end data pipeline that operationalizes national capacity frameworks by combining text embeddings, configurable field-weighting optimizations, and a cost-efficient cascaded language model strategy (smaller screening models paired with selective larger refinement models) for large-scale semantic mapping. It's worth noting that data scientists can use TALENT to identify human-installed capacity at the regional or macro-regional level. The strategic thematic framework will define whether it is a region or a macro-region.
2. **An Auditable Human Curation Process:** A systematic approach to data provenance that pairs machine outputs with expert human curation, incorporating evidence (with {*theme, researcher, affinity, explanation*}) to satisfy institutional compliance, risk assessment, and transparency requirements in strategic monitoring.
3. **A Comprehensive National Empirical Evaluation on AI Capacity:** A nationwide empirical evaluation utilizing the CAPES repository to map the installed capacity of Artificial Intelligence across the SNPG. This analysis provides a cross-disciplinary, granular diagnosis of human resource development across states, universities, and knowledge areas, revealing critical trends in AI applications outside traditional computing fields to support evidence-based public policies.

The results and contributions of the proposed pipeline, while also facilitating evidence-based collective decision-making and increasing institutional transparency and trust in the public sector, reinforce the understanding that governance is not merely management or operational administration. Governance is the ability of public organizations to evaluate, direct, and monitor policies, resources, and processes, and to orient these toward generating public value and delivering results to society Nardes *et al*. [2022].

This work is organized as follows. Section 2 presents the theoretical background. Section 3 discusses related work on

[1]Agenda Nacional: `https://agendanacional.capes.gov.br`

workforce competence in AI. Section 4 reviews affinity detection methods between strategic themes and master's/doctoral graduates. Section 5 describes the proposed pipeline for mapping AI-related strategic themes to theses and dissertations. Section 6 presents the empirical evaluation in the Brazilian context. Finally, Section 7 presents the conclusions and future work.

2 Theoretical Background

Natural Language Processing: Text preprocessing is a traditional stage in Natural Language Processing (NLP). It transforms raw textual data into a structured representation suitable for computational processing [Aggarwal, 2022]. In tasks involving scientific documents, such as theses and dissertations, this step is particularly important because the available texts may contain noise, long descriptions, domain-specific terminology, abbreviations, and heterogeneous writing styles. Classical preprocessing operations include lowercasing, punctuation removal, tokenization, stopword removal, stemming or lemmatization, and term normalization [Aggarwal, 2022]. These preprocessed texts support simple but useful matching strategies, including exact string matching, fuzzy matching, and keyword overlap [Utino and Gôlo, 2025]. For affinity detection between themes and academic profiles, these methods provide a first approximation of semantic relation by checking whether words or expressions associated with a theme appear in titles, abstracts, keywords, or research areas. Although these strategies are efficient and interpretable, they are limited because they depend heavily on lexical overlap and often fail to capture semantic similarity when different terms express related concepts.

Bidirectional Encoder Representations from Transformers (BERT): Transformer-based models introduced a more expressive way to represent text by learning contextual embeddings, where a word's meaning depends on the surrounding words in the sentence [Vaswani *et al*., 2017]. Unlike traditional models, Transformers rely on the attention mechanism, which allows each token to attend to other tokens in the sequence and capture dependencies [Vaswani *et al*., 2017]. BERT is a pioneering model in this architecture and is trained primarily via masked language modeling [Devlin *et al*., 2019]. Formally, given an input sequence of tokens $X = (x_1, x_2, \ldots, x_n)$, the BERT model generates a corrupted version of X, $\hat{X}$, in which tokens are selected randomly (e.g., 15%) and replaced by the `[MASK]` symbol. Also, let $\bar{X}$ be the masked tokens of X. Then, BERT's training consists of reconstructing the masked tokens $\bar{X}$ from $\hat{X}$ [Yang *et al*., 2019]:

$$\max_{\Theta} \log p_{\Theta}(\bar{X}|\hat{X}) \approx \sum_{x_j \in X} c_{x_j} \log \frac{\exp(H_\theta(\hat{\boldsymbol{X}})^T_{x_j} \boldsymbol{e}(x_j))}{\exp(\sum_{x'} H_\theta(\boldsymbol{X})^T_{x_j} \boldsymbol{e}(x'))}, \tag{1}$$

where $c_{x_j} = 1$ indicates that x_j is masked, $\boldsymbol{e}(x_j)$ indicates the embedding of the token x_j, x' is X without x_j, $H_\theta(X) = \{\mathbf{h}_\theta(X)_1, \mathbf{h}_\theta(X)_2, \ldots, \mathbf{h}_\theta(X)_v\}$ is a sequence of hidden vectors mapped by a Transformer.

Large Language Models: Large Language Models (LLMs) extend this representation-learning paradigm by scaling neural language models to billions of parameters and training them on massive corpora composed of billions or trillions of tokens [Minaee *et al*., 2024]. Most autoregressive LLMs are trained with a next-token prediction objective, where the model receives a sequence of tokens and learns to predict the next token conditioned on the previous context [Chang *et al*., 2024]. Formally, given a sequence $X = (x_1, x_2, \ldots, x_n)$, the training objective minimizes the negative log-likelihood [Chang *et al*., 2024]:

$$\mathcal{L}(\Theta) = -\sum_{t=1}^{n} \log P_\Theta(x_t \mid X, x_{<t}), \tag{2}$$

where $x_{<t}$ represents all tokens preceding position t. By repeatedly solving this prediction task at scale, LLMs acquire broad linguistic, factual, and reasoning capabilities, making them adaptable to many downstream tasks through prompting, instruction tuning, or fine-tuning [Chang *et al*., 2024]. In text classification and textual correlation problems, LLMs can analyze text pairs and infer whether they are semantically related, even with little lexical overlap [Li *et al*., 2024]. This ability is especially useful for detecting affinity between researchers and strategic themes, because the relationship between a thesis or dissertation and a theme may depend on implicit concepts, domain knowledge, and contextual interpretation. However, the same scale that gives LLMs strong semantic capacity also increases financial cost and energy consumption, motivating efficient model adaptation and filtering strategies before applying larger models at scale.

AI Sovereignty and AI capacity: The theoretical foundation of AI sovereignty is rooted in macroeconomics, public policy, and geopolitics, evolving from the broader concept of digital or technological sovereignty [EIT Digital, 2021; Truby *et al*., 2026]. It represents a state or region's assertion of supreme authority over the AI systems, data, and digital infrastructures within its territory, intentionally shifting power away from foreign governments and market-dominant technology platforms [Truby *et al*., 2026]. According to the literature, AI sovereignty is not an abstract political claim but a strategic capacity that operates across three interdependent technical dimensions [EIT Digital, 2021]:

- **Algorithmic Autonomy:** The localized capability to audit and regulate machine-learning models, preventing blind reliance on foreign "black-box" models.
- **Data Sovereignty:** Establishing local data ownership models (decentralized or sovereign storage) to counter global data imbalances and structural dependencies.
- **Infrastructure and Cloud Sovereignty:** Secure, federated, and localized computing infrastructure (sovereign clouds) required to execute advanced AI without vulnerability to external surveillance or geopolitical lockouts.

The theoretical scope of AI sovereignty extends into international relations through the emerging concept of "AI Diplomacy" [Truby *et al*., 2026]. As a revolutionary subset of tech diplomacy, AI diplomacy connects governance to international stability, with major global powers (such as the US, China, and the EU) leveraging techno-nationalism,

infrastructure outreach, and regulatory standards to shape global security. In this scenario, AI capacity is theorized as the foundational prerequisite for sovereignty; a region must possess the native technical infrastructure, regulatory frameworks, and human expertise to independently sustain, govern, and negotiate its place within the global AI landscape.

Scientometrics, Research Capacity, and Strategic Intelligence: Scientometrics and bibliometrics provide established quantitative approaches for characterizing scientific activity and the structure of research systems. Common methods include publication and citation analysis, co-citation and bibliographic coupling, co-authorship analysis, and co-word analysis. Together, these methods enable investigation of dimensions such as scientific output, impact, collaboration patterns, intellectual structures, and thematic specialization within research communities, institutions, and countries Zupic and Čater [2015]. Thus, traditional scientometric approaches are particularly valuable for describing what a scientific system produces, how it organizes knowledge, and how researchers and institutions interact.

Research capacity extends beyond scientific output alone. The literature on research capacity strengthening recognizes that capacity can be assessed at individual, institutional, and systemic levels and may involve dimensions such as the acquisition and application of research skills and knowledge, research collaboration, knowledge transfer, research management, and bibliometric outcomes. Pulford *et al*. [2020], for example, reviewed 668 indicators used to evaluate research capacity strengthening and showed that bibliometric indicators represent only one of several dimensions, with substantial emphasis also placed on skills and knowledge development, collaboration, and knowledge transfer. This perspective supports interpreting advanced graduate education, e.g., dissertations and theses, as another important component of research capacity, since master's and Ph.D. education forms highly qualified human resources capable of contributing to specific scientific and technological domains.

Scientometric approaches have also been used to investigate the alignment between scientific activity and external societal or policy priorities. Ciarli and Ràfols [2019], for instance, used bibliometric information to identify countries' research priorities and compared them with societal demands, showing that the thematic orientation of scientific production is not necessarily aligned with societal needs. Their study illustrates the relevance of moving beyond measures of scientific volume or impact toward analyses that explicitly relate research activity to externally defined priorities, which is particularly important in policy-oriented and mission-driven contexts. From this perspective, our work should not be viewed as a replacement for scientometric or bibliometric approaches, but rather as a complementary analytical perspective. Whereas scientometric analyses predominantly characterize scientific output, impact, collaboration structures, and thematic specialization, our work focuses on advanced human-capital formation and its semantic alignment with externally defined strategic priorities.

3 Workforce Competence in AI

We conducted a rapid Scopus review to map the state of the art on workforce competence in using and developing AI technologies, with a specific focus on assessment methods. Our search strategy combined keywords such as "assessment", "human capacity", and "Artificial Intelligence" along with their synonyms. The retrieved literature mainly emphasizes student capabilities, practical AI applications across diverse domains (e.g., medicine and engineering), and specific professional AI competency requirements. Notably, to our knowledge, no study has proposed a method using AI to evaluate human AI capacity within a population or organization. Nevertheless, several studies underscore the strategic importance of measuring regional or national AI capacity as a foundational pillar for digital planning.

For instance, Techanamurthy *et al*. [2025] examined Industry 4.0 readiness among 506 Malaysian Small and Medium Enterprises, evaluating dimensions such as leadership, digital infrastructure, and workforce competency. Their findings showed that while leadership and strategic alignment scored highest, workforce competency scored lowest, highlighting a critical skill gap that hinders technological evolution. Although their study emphasizes targeted workforce training to bridge these readiness gaps, it does not provide a specific metric to measure AI-related workforce readiness.

Similarly, Mollaoglu *et al*. [2026] investigated Industry 4.0 components within the maritime sector. By combining a literature review with expert interviews, they developed a framework to rank key digital transformation factors. Their results identified continuous training as the most influential component, demonstrating that successful digital transformation fundamentally relies on human capabilities rather than on technological infrastructure alone. However, their framework does not define what these specific human capabilities entail or offer a mechanism to measure them.

This gap in the literature highlights a critical misalignment in current digital transformation strategies. While the literature consensus identifies human capability and workforce readiness as the most vital and vulnerable components for technological sovereignty, it lacks concrete tools to measure them. Relying on subjective assessments or generic readiness scores fails to capture the technical expertise actually present in a region. To address this limitation, this work introduces a novel approach that operationalizes the capability framework by evaluating regional AI capacity through the automated analysis of defended theses and dissertations. By leveraging this academic output, our method provides a tangible, data-driven diagnostic tool to assess local AI expertise, bridging the gap between high-level policy guidance and regional technological sovereignty.

4 Affinity detection between themes and master's/doctoral graduates

Affinity detection between strategic themes and master's/doctoral graduates has emerged as a relevant task for evidence-based science policy, particularly in contexts where graduate education systems should be evaluated according to their ca-

pacity to respond to national and regional priorities. In Brazil, the National Agenda for Graduate Education defines strategic themes declared by the 27 federative units and organizes them as a policy instrument to guide the training, retention, and attraction of highly qualified personnel, including master's, doctoral, and postdoctoral researchers [Turine *et al*., 2025]. Recent works have addressed this problem through complementary strategies, including embedding-based similarity thresholds [Miranda *et al*., 2025], lightweight classifiers over semantic embeddings to improve scalable filtering [Utino and Gôlo, 2025], theme-specific one-class learning to handle scarce labeled data and provide interpretability [Gôlo and Utino, 2025], and efficient fine-tuning of language models for direct affinity classification [Sakiyama *et al*., 2025]. Together, these studies indicate that mapping AI capacity requires methods that balance semantic accuracy, scalability, low computational cost, and transparency, since errors in affinity detection may directly affect estimates of how many master's/doctoral graduates are effectively addressing AI-related strategic areas.

Miranda *et al*. [2025] addresses large-scale affinity detection between strategic themes and academic profiles (master's/doctoral graduates) in the Brazilian National Graduate Education System by proposing a BERT–Threshold strategy. The authors propose a pre-filtering module based on contextual embeddings and cosine similarity. In this strategy, strategic themes and their keywords are compared with textual fields extracted from master's and doctoral dissertations and theses, including titles, abstracts, keywords, and graduate program concentration areas. The method evaluates various embedding models, aggregation operators, field-weighting schemes, and decision thresholds to determine whether a profile is relevant to a theme.

Utino and Gôlo [2025] extends the affinity detection problem by evaluating lightweight filtering mechanisms for master's/doctoral graduates and theme prediction under both predictive and environmental constraints. The authors formulate the task as a binary classification problem, distinguishing high-affinity pairs from non-high-affinity pairs, using a real-world dataset of 42,046 theme-thesis/dissertation pairs defended in Brazil in 2023. Their strategy compares three families of methods: lexical baselines based on exact and fuzzy matching, cosine-similarity thresholding over BERT-based embeddings, and lightweight classifiers trained on pooled embedding representations. We obtained the best result using Granite multilingual embeddings combined with K-Nearest Neighbors and max pooling.

Gôlo and Utino [2025] proposes a theme-specific one-class learning strategy for filtering affinity between strategic themes and master's/doctoral graduates. Unlike similarity-based filters, which compare embeddings in an unsupervised and generic way, and unlike binary classifiers, which require many labeled positive and negative examples, the authors model each strategic theme as a specific class of interest. For each theme, the filter is trained on its description, keywords, and a small set of high-affinity profiles, learning a decision region that represents the theme's semantic space. The method uses multilingual embedding models to represent theses and dissertations. It applies Support Vector Data Description in a two-dimensional projected space, enabling both lightweight inference and visual interpretability of the learned decision boundary.

Sakiyama *et al*. [2025] investigates affinity detection as a three-class classification task, categorizing Brazilian theses and dissertations according to their level of affinity to strategic themes: low, medium, or high. Unlike previous filtering-oriented approaches, the authors directly adapt language models to the final classification task and compare efficient fine-tuning strategies for LLMs and transformer encoders under resource constraints. Their main proposal fine-tunes Qwen3-14B-Base with QLoRA. The authors compare their method with zero-shot Qwen models and ModernBERT, demonstrating that the balanced-loss-fine-tuned LLM achieves the best overall classification performance, particularly by improving robustness to minority classes.

Taken together, Miranda *et al*. [2025], Utino and Gôlo [2025], Gôlo and Utino [2025], and Sakiyama *et al*. [2025] show that affinity detection is a central component of AI capacity mapping, since it enables the identification of master's and doctoral graduates whose academic production is semantically aligned with strategic AI-related themes. These works address complementary stages of the same problem, ranging from embedding-based thresholding and lightweight supervised filtering to theme-specific one-class learning and efficient LLM fine-tuning. Their common contribution is to make national-scale capacity estimation more feasible by reducing unnecessary LLM calls and lowering computational costs, energy use, and carbon emissions, while preserving or improving semantic classification quality. At the same time, they highlight persistent challenges for AI capacity pipelines, especially the risk of false negatives, the scarcity of labeled examples for specific themes, and the need to balance classification accuracy, interpretability, and scalability in policy-oriented analyses of installed human capacity.

5 TALENT: Theses Analysis & Lifecycle Evaluation for National Capacity Tracking

We propose a scalable automatic pipeline to assess the degree of affinity between strategic themes within a macro-regional strategic thematic priority framework and academic research outputs produced by graduates of postgraduate programs. Figure 1 illustrates the proposed TALENT pipeline, which takes as input a user-defined research topic, a set of strategic agenda themes, and textual metadata associated with master's dissertations and doctoral theses defended by students in postgraduate programs within the analyzed region. As output, it produces structured mappings as tuples composed of a strategic theme, a researcher (**a master's or doctoral graduate student**), an affinity level, and an explanatory justification. These mappings can support analyses of the installed capacity for human resource development in the research topic at regional, institutional, and postgraduate program levels.

A macro-region is an economic and political entity with its own legal framework and may establish new regulations for the regions that compose it. Each region also has its own legal framework, but it cannot violate the legal structure of the

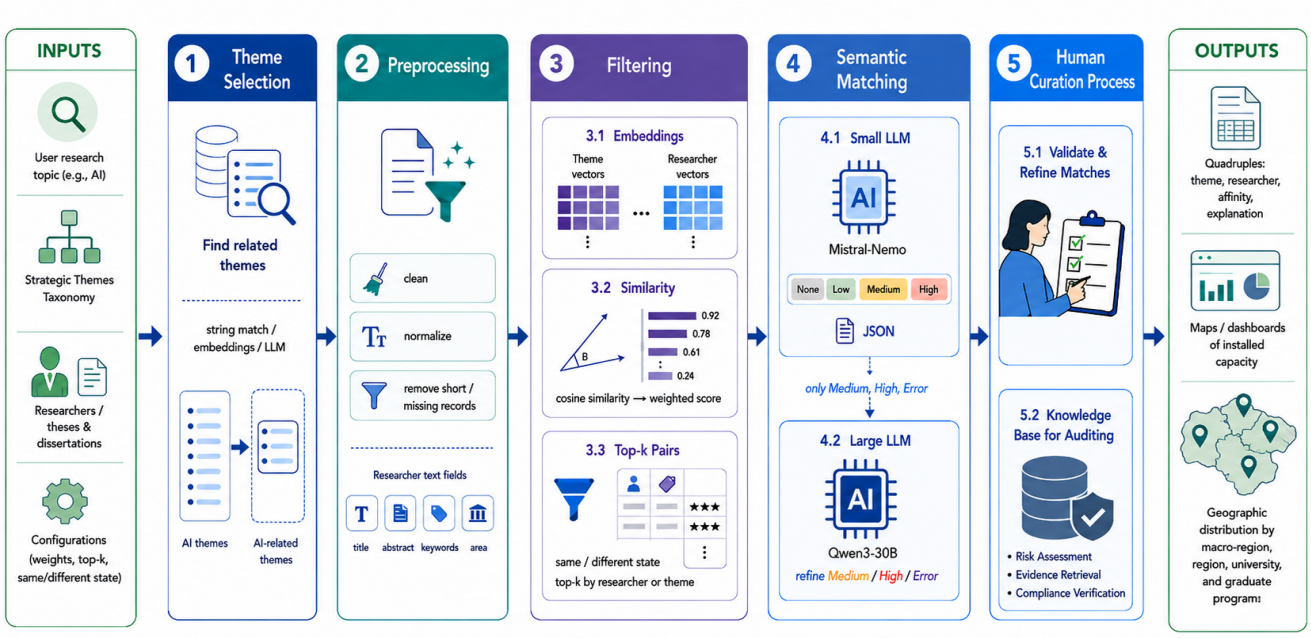

Figure 1. Overview of the proposed pipeline for mapping affinity between strategic themes and academic research outputs. TALENT starts from a user-defined research topic, filters related strategic themes, preprocesses researcher metadata, generates semantic embeddings, computes similarity-based candidate pairs, and applies a cascaded language-model strategy to classify and explain theme–researcher affinity. The final output supports human curation, auditing, and geographic or institutional analyses of installed research capacity.

macro-region to which it belongs. Examples of macro-regions and their respective component regions include: Brazil and its 27 states; the state of São Paulo and its 645 municipalities; the United States and its 50 states; and the European Union and its 27 member countries. Therefore, the proposed pipeline can be adapted for analyses at both regional and macro-regional levels.

The proposed architecture follows a five-stage cascaded design with parameters configured for automatic execution. First, it uses textual preprocessing and dense vector representations to position strategic themes and researchers within a shared semantic space based on a user-defined research topic. Next, it computes similarity scores and applies configurable filters to reduce the number of candidate pairs. Finally, it employs LLMs to perform a more refined semantic assessment of the selected pairs. This design avoids sending all possible combinations of themes and researchers directly to large language models, which would be computationally expensive in large-scale institutional scenarios. Instead, TALENT reserves language-model inference only for the most promising candidates, balancing coverage, computational cost, execution time, and semantic quality.

Stage 1 – Theme selection: TALENT begins with a user-defined research topic or an organization's predefined topic of interest, such as *Artificial Intelligence*, which we use in our empirical evaluation, and which guides the selection of related strategic themes from a broader, policy-oriented knowledge base. The strategic themes are drawn from a database. Given the input topic, TALENT retrieves a subset of semantically associated strategic themes and forwards them to the next stages. This retrieval step can use lexical matching, embedding-based similarity, or language-model-based semantic matching. Because the total number of strategic themes is relatively small, this initial search can employ flexible semantic strategies without compromising scalability, as we observed in our evaluation when filtering the 470 themes to those relevant to the specific topic *Artificial Intelligence*.

Stage 2 – Preprocessing: This stage preprocesses the complementary textual fields associated with each researcher, specifically the dissertation or thesis title, abstract, keywords, postgraduate program, and knowledge area. Each field provides distinct evidence: the title indicates the main object of study, the abstract provides a detailed description of the research, the keywords summarize central concepts, and the program or knowledge area adds disciplinary context. The preprocessing stage cleans, normalizes, and standardizes these fields. Crucially, it filters out and removes records with insufficient textual evidence—such as entries with missing abstracts or very short titles—to prevent incomplete or noisy data from degrading subsequent semantic representations. Thus, this stage ensures each researcher is uniformly represented by textual information that is sufficiently descriptive, comparable, and optimized for semantic matching.

Stage 3 – Filtering: After preprocessing, TALENT generates text embeddings to transform strategic themes and researchers, respectively drawn from the strategic themes and the Researchers/Theses and Dissertations sets, into numerical vectors for direct comparison, capturing semantic proximity even without exact term matches. For each strategic theme, TALENT encodes its name and creates a single descriptor embedding by averaging the individual embeddings of its associated keywords. For each researcher, TALENT generates a multi-field representation by embedding specific academic work (title and abstract) alongside complementary thematic and disciplinary context (keywords, postgraduate program, and knowledge area). Finally, TALENT creates

candidate theme-researcher pairs and computes semantic similarity scores between their corresponding fields using cosine similarity, yielding multiple partial scores per pair.

Let S_{title} denote the similarity between the researcher's title and the theme name, S_{abstract} the similarity between the researcher's abstract and the theme name/abstract, S_{keywords} the similarity between the researcher's keywords and the theme keywords, and S_{area} the similarity between the researcher's knowledge area and the theme name. The final semantic affinity score is computed as: $S_{\text{affinity}} = \alpha S_{\text{title}} + \beta S_{\text{abstract}} + \gamma S_{\text{keywords}} + \delta S_{\text{area}}$, where α, β, γ, and δ are configurable weights. These parameters let the system favor detailed research content by assigning a higher weight to the abstract, or favor direct alignment with strategic descriptors by assigning a higher weight to keywords. To avoid the costly and unnecessary processing of the massive number of candidate pairs generated by the complete Cartesian product between researchers and strategic themes, TALENT filters pairs based on their semantic affinity scores using one of two filtering strategies:

Top-k themes per researcher: Selects the top-k strategic themes for each researcher to identify the themes most closely related to each academic profile.

Top-k researchers per theme: Selects the top-k researchers for each strategic theme to identify the installed research capacity for a given strategic priority.

The parameter k directly controls the trade-off between coverage and computational efficiency. Larger values of k increase recall while minimizing the risk of discarding relevant pairs, at the expense of more language-model calls, whereas smaller values restrict the filtering stage to reduce cost and execution time, making k the central parameter for adapting TALENT to specific institutional constraints.

Stage 4 – Semantic Matching: Following candidate filtering, TALENT executes a semantic assessment by processing the strategic theme, its keywords, and the researcher's textual evidence through language models. The system must assign one of four precise affinity labels: *High* (clear, explicit semantic proximity and descriptor match), *Medium* (relevant proximity but less direct match), *Low* (weak or generic relation), or *None* (no meaningful semantic relation). Then, it generates an explanation grounded in the evidence. The model output is constrained to a structured JSON format with two fields: the affinity label and an explanation that must explicitly cite evidence from both the researcher's data and the strategic theme for any classification other than *None*.

To optimize efficiency without sacrificing quality, TALENT implements a cascaded language model strategy. A smaller language model initially processes all candidate pairs from the filtering stage to perform broad semantic screening. A larger language model is then selectively invoked only for pairs that the smaller model categorizes as *Medium*, *High*, or *Error*. This step ensures the more expensive model is reserved for refining high-potential affinities or recovering from invalid or malformed JSON outputs, combining computational cost reduction with semantic validation in the most critical cases.

Stage 5 – Human Curation Process: Similar to the framework proposed by Karacapilidis *et al*. [2024], TALENT adopts a human-in-the-loop approach to improve reliability, auditability, and explainability in large-scale semantic analysis, combining semantic embeddings, LLMs, and human curation to support higher-education governance and evidence-based policymaking within the SNPG. This final stage of TALENT incorporates human curation by domain experts to validate, revise, and consolidate LLM-generated classifications, strengthening institutional reliability, reducing semantic bias, and improving transparency and traceability in large-scale analytical scenarios. This stage also contributes to building an auditable knowledge base through activities such as risk assessment, evidence retrieval, and compliance verification.

TALENT outputs structured mappings represented as tuples in the form *(theme, researcher, affinity, explanation)*, where the *theme* corresponds to a strategic or policy-related topic, the *researcher* represents a master's or doctoral graduate, and *affinity* indicates the semantic relationship level between the researcher and the theme. The *explanation* provides the textual evidence supporting the generated classification. These structured tuples can be aggregated by macro-region, region, state, university, and postgraduate program to analyze installed research capacity at different geographic levels. In the Brazilian case study, this output provides the evidence needed to map how installed capacity for human resource development in Artificial Intelligence is distributed across the Brazilian National Graduate System, directly supporting strategic planning, public policy formulation, and institutional monitoring.

6 Empirical Evaluation

The aim of this empirical evaluation is to answer the following Research Questions (RQs):

- **RQ1:** Which strategic themes and macro-themes of the CAPES National Agenda present explicit adherence to the topic of Artificial Intelligence (AI)?
- **RQ2:** What is the installed capacity of the SNPG for educating Master's and Ph.D. graduates in Artificial Intelligence in Brazil?
- **RQ3:** To what extent are graduate programs in the Computer Science evaluation area contributing to the installed capacity for educating Master's and Ph.D. graduates in AI-related strategic themes of the CAPES National Agenda?
- **RQ4:** Which major knowledge areas of the SNPG are contributing to the education of Master's and Ph.D. graduates in the AI-related strategic themes mapped within the CAPES National Agenda?

In what follows, we describe the scenario we consider to address these questions, present the results obtained by applying TALENT to our RQs, and discuss the RQ answers.

6.1 Scenario Description

The Coordination for the Improvement of Higher Education Personnel (CAPES), established in 1951, linked to the Ministry of Education, has the mission of promoting the expan-

sion, consolidation, and evaluation of stricto sensu graduate education, acting directly in the structuring and strengthening of the SNPG. CAPES has consolidated itself as the main national agency responsible for supporting master's, Ph.D., and postdoctoral researchers, evaluating the quality of graduate programs, promoting the internationalization of Brazilian science, and fostering public policies oriented toward the country's scientific, technological, and social development. CAPES evaluates all Brazilian graduate programs, and only the ones approved by CAPES are part of the SNPG.

CAPES is also responsible for developing the National Graduate Education Plan (PNPG), reinforcing its strategic role in formulating, coordinating, and monitoring national graduate education policies. PNPG is a main long-term planning instrument for Brazilian graduate education, guiding the strategic directions of CAPES and the SNPG. PNPG 2025–2029 (7th edition) represents a milestone in strategic renewal in response to contemporary challenges in science, innovation, and national development. The document addresses structural themes, including reducing regional asymmetries, intranational and international academic mobility, and equity and diversity, among others.

Among the strategic challenges identified by the PNPG 2025–2029, one of the most relevant is the need for stronger national coordination in Research, Development, and Innovation (R&D&I). Although Brazil has achieved significant installed capacity in the training of Master's and Ph.D. graduates and in scientific production, important limitations remain in translating knowledge into innovation, achieving international competitiveness, and generating the plan recognizes that the historical absence of a National Agenda articulated with the priorities of Brazilian states has hindered convergence among higher education institutions, funding agencies, state governments, the productive sector, and society.

CAPES developed the National Agenda[2] for Advanced Human Resource Development, officially launched in 2025 as a new strategic milestone for the governance of Brazilian public policy on graduate education. Conceived as a complement to the PNPG [CAPES, 2025], the National Agenda seeks to systematically align Master's and Ph.D. education with the socioeconomic, scientific, and technological demands of Brazil's states and regions. By bringing academic research closer to regional strategic priorities, the initiative establishes a new evidence-based planning paradigm grounded in dialogue with the priority public policies of Brazilian territories, functioning as an institutional reference for guiding funding allocation, graduate scholarships, and public policies aimed at national development through the education of highly qualified human resources (Master's and Ph.D. holders).

The National Agenda identified priority strategic themes to guide public investments in advanced human capital, bringing graduate education closer to the real demands of Brazilian territories and society. The Agenda's main purpose is to systematically align Master's and Ph.D. education with the socioeconomic and technological demands of Brazil's different states and regions. By bringing academic research closer to regional needs, the Agenda framework serves as an official "compass" for allocating funding and graduate scholarships. We chose this database because it contains 470 strategic themes collaboratively defined by the 27 Brazilian states, thereby aligning the measurement of AI installed capacity more closely with the socioeconomic and technological needs of Brazil's different macro-regions. Figure 2 shows the distribution of the 470 strategic themes declared across the 27 Brazilian Federative Units, which has the following configuration, in decreased order (states with more themes to the less themes): Ceará (CE) – 24; Amazonas (AM) – 23; Amapá (AP), Mato Grosso do Sul (MS), Rio de Janeiro (RJ), and Roraima (RR) – 21; Bahia (BA) – 20; Distrito Federal (DF), Espírito Santo (ES), Pará (PA), Piauí (PI), Rio Grande do Norte (RN), Sergipe (SE), and Tocantins (TO) – 19; Alagoas (AL) – 17; Goiás (GO), Pernambuco (PE), Rondônia (RO), and São Paulo (SP) – 16; Acre (AC) – 15; Rio Grande do Sul (RS) and Santa Catarina (SC) – 14; Maranhão (MA), Paraíba (PB), and Paraná (PR) – 13; Mato Grosso (MT) – 12; and finally Minas Gerais (MG) – 11.



Figure 2. Brazilian map with the distribution of the number of strategic themes per state.

In this context, our empirical evaluation assesses the SNPG's installed capacity in the research topic "Artificial Intelligence (AI)". For the analysis, we used real-world data from the CAPES Sucupira Platform[3]. Its database contains 4,635 graduate programs (PPGs) offered by 457 Science and Technology Institutions (STIs), totaling 111,699 faculty advisor appointments (since one faculty member may serve as a faculty advisor to multiple graduate programs) and 438,888 students enrolled in Master's and Ph.D. programs. Fig. 3 illustrates the distribution of graduates from 2013 to 2024. For the empirical evaluation, we focus on the two most recent years – 2023 and 2024 – as they were the most recent complete updates on the platform and were used in our evaluation. Among the students enrolled in the SNPG, we observed the following numbers of graduates who successfully defended their master's dissertations or doctoral theses: 2023 – 91,463 from 460 Science and Technology Institutions (STIs), with

[2] *Agenda Nacional da CAPES* in Portuguese. Available at `https://agendanacional.capes.gov.br`.

[3] Available at `https://sucupira.capes.gov.br/`.

4,659 graduate programs; and 2024 – 91,833 from 457 STIs, with 4,635 graduate programs. These numbers underscore the importance of tools for analyzing large volumes of data, including dissertations and theses, such as TALENT. Ultimately, our evaluation uses the academic work produced by these graduates to measure installed Artificial Intelligence capacity across regions in Brazil.

Using TALENT to map the academic works (theses and dissertations) and their authors data, that we call researchers from now on, our premise is to be possible to assess how AI human resource training capacity is distributed across Brazilian regions, states, universities, and graduate programs, generating evidence to support strategic planning, public policy formulation, and institutional monitoring of Brazilian graduate education[4]. So, we investigate how many of the Brazilian SNPG's trainees are highly qualified in AI, specifically Master's and Ph.D. graduates, and how they are distributed over the country.

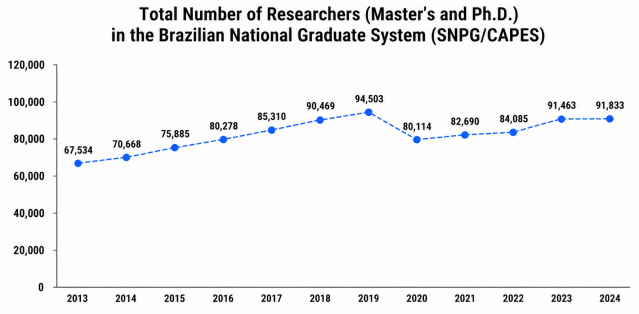

Figure 3. Distribution of graduates in SNPG CAPES/Brazil

6.2 Results and Discussion

Applying Stage 1 – Selection of AI-Related Strategic Themes: As a result of executing the Theme Selection stage of the proposed pipeline using the research topic *"Artificial Intelligence"*, Table 1 shows the 28 strategic themes from different Brazilian Federative Units semantically mapped to the research topic[5]. We marked the topic "Artificial Intelligence" in blue and bold to show that only 14 themes (out of 28) have the topic AI in their title. Table 2 shows the distribution of themes across seven macro-themes, showing the transversal and interdisciplinary nature of AI within the Brazilian national development strategy. It is important to note that a given strategic theme may be associated with multiple macro-themes, as outlined in the CAPES National Agenda.

The semantic mapping process identified strategic themes explicitly or implicitly associated with AI across multiple domains, including digital transformation, data science, information and communication technologies, robotics, bioinformatics, smart agriculture, public governance, and health technologies. This initial filtering stage significantly reduces the semantic search space for subsequent pipeline phases, restricting analysis to themes that more closely align with the research topic and thereby improving computational efficiency and analytical precision.

The pervasive nature of AI is evident across 35% of the macro-themes (7 out of 20) within the CAPES National Agenda, underscoring its importance as an enabling technology. Rather than being confined to digital transformation, it serves as a core driver for strategic domains such as smart cities, urban mobility, healthcare, agriculture, infrastructure, public security, logistics, and industrial innovation. These observations show that AI is no longer restricted to Information and Communication Technologies (ICTs) or computing alone, but has become a foundational element of scientific, technological, economic, and social development. This distribution reinforces AI's role as a core enabling technology that may support socioeconomic transformation, innovation ecosystems, and evidence-based public policymaking across institutions and territories.

In summary, the research topic *Artificial Intelligence* intersects with a broad strategic relevance within the CAPES National Agenda, covering 28 of the 470 strategic themes (unique ID) (5.96%), according to Table 1, and is present in 20 of the 27 Brazilian Federative Units (State) (74.07%). Table 2 presents the distribution of these themes across 7 of the 20 macro-themes (35%). Macro themes 4 and 5 contain the most strategic themes related to Artificial Intelligence. These results highlight AI's transversal and territorial nature, reinforcing its role as a strategic enabling technology across multiple sectors and regions.

Applying Stages 2-4 – From Preprocessing to Automatic Semantic Matching

After the Theme Selection stage related to the *Artificial Intelligence* (defined by the user-defined research topic in TALENT), the mapping phase begins, matching the 28 identified strategic themes to each researcher (Master's and Ph.D. graduates) in the CAPES Sucupira Platform knowledge base.

According to Sucupira, 460 STIs offered 4,659 PPGs, graduating 91,463 Master's and Ph.D. researchers in 2023 alone. In 2024, 457 STIs offered 4,635 PPGs and graduated 91,833 researchers. Given these numbers, TALENT is essential for preprocessing researcher metadata, generating semantic embeddings, computing similarity-based candidate pairs, and applying a cascaded language-model strategy to classify and explain theme–researcher affinity.

We highlight that, in TALENT's output alignments, a researcher may exhibit semantic affinity with multiple strategic themes simultaneously. For example, according to Fig. 4, the researcher[6] defended a doctoral thesis entitled "**Heterogeneous Graphs For Text Representation: An Integrated Approach With Language Models**" in the Graduate Program in **Computer Science and Computational Mathematics** (Code number 33002045004P1) at the **University of São Paulo (USP)**, São Carlos, State of São Paulo. The semantic mapping associated the researcher with multiple AI-related strategic themes, including **ID 156** – ES State, "Information and Communication Technologies", with **Medium** affinity, and **ID 420** – SP State, "Data Science and Artificial Intelligence", with **High** affinity. In addition to these

[4]For studies focused on mapping regional technological capacity, the Sucupira Platform serves as an institutional knowledge base for Brazil, standardizing and aggregating raw academic output into structured data.

[5]We preserved the identifiers (*Theme IDs* and *Macro Theme IDs*)

[6]Due to the Brazilian General Data Protection Law (LGPD), we cannot disclose this personal information in this work.

Table 1. AI-related strategic themes selected from the CAPES National Agenda.

ID	State	Strategic Theme
1	AC	Agribusiness and Information and Communication Technologies
32	AL	Information and Communication Technologies: **artificial intelligence**, digital humanities, and data science
49	AM	Information Security: **artificial intelligence** and traditional knowledge
89	BA	Market Intelligence: data and information analysis and strategy creation
96	BA	Digital Transformation: connectivity, data science, and **artificial intelligence** for social inclusion
119	CE	Information and Communication Technologies: **artificial intelligence**, cybersecurity, and educational technologies
143	ES	Smart Cities: user-oriented services, mobility, and urbanism
156	ES	Information and Communication Technologies
174	GO	Public Safety: evidence-based decision making
199	MT	Digital Transformation and New Jobs
206	MS	Human, Smart, and Sustainable Cities
225	MG	Regional Development: smart cities and asymmetry reduction
242	PA	Technological Integration: cyber-development and **artificial intelligence**
263	PB	Digital Transformation, High-Performance Computing, and Industry 4.0
264	PR	Smart Agriculture: **artificial intelligence** and technologies
272	PR	**Artificial intelligence**: data science, machine learning, and deep learning
324	RJ	Industry 4.0: robotics, bioinformatics, cybersecurity, quantum computing, and **artificial intelligence**
326	RJ	Urban Mobility, Logistics, and Smart Cities
354	RS	Data Science and **Artificial intelligence** Applied to Strategic Sectors
357	RS	Electronics, Bioengineering, and Industry 4.0: **artificial intelligence**, innovation ecosystems, and technological inclusion
386	RO	Data Science and **Artificial intelligence**
410	SC	Industries 4.0 and 5.0: competitiveness and industry-oriented education
414	SC	Information and Communication Technologies: **artificial intelligence**, internet of things, robotics, and bioinformation
418	SP	Biotechnology: human and animal health, microorganisms, and plants
420	SP	Data Science and **Artificial intelligence**
432	SP	Quantum Technologies: computing, sensors, communication, and cybersecurity
436	SE	Communication and **Artificial intelligence** for Social Inclusion
469	TO	Digital Transformation

Table 2. AI-related strategic themes for each Macro-Theme.

ID	Macro-Theme	Associated Themes
4	Digital Transformation, Information and Communication Technologies, and Artificial Intelligence	1, 32, 49, 89, 96, 119, 156, 199, 242, 263, 264, 272, 324, 354, 357, 386, 414, 420, 432, 436, 469
5	Industry, Technological Innovation, and Business	89, 263, 357, 410, 432
9	Health, Quality of Life, and Technologies	324, 418
11	Agriculture, Family Farming, and Food Security	1, 264
14	Smart Cities, Mobility, and Sustainable Urbanism	143, 206, 225, 326
16	Logistics, Engineering, Infrastructure, and Strategic Technologies	326, 357, 432
18	Violence Prevention and Public Security	174

AI-related themes, among the 470 strategic themes of the National Agenda, the researcher was also associated with theme **ID 419** — SP State, "Sustainable and Smart Cities: Territorial Planning, Mobility, and Legal, Social, and Technological Aspects", with **Low** affinity. All remaining strategic themes were classified as having no affinity.

These observations reinforce the interdisciplinary and transversal nature of Artificial Intelligence research within graduate education, where a single thesis may simultaneously contribute to domains such as digital transformation, healthcare technologies, data science, smart agriculture, public governance, or intelligent infrastructure systems. These multidimensional associations also highlight the relevance of semantic approaches based on embeddings and LLMs, as traditional keyword-matching techniques struggle to capture these complex thematic relationships.

TALENT initially generated 2,652,427 candidate strategic theme–researcher pairs for semantic analysis in 2023, which were then classified into affinity levels (*High, Medium, Low,* and *None*). This result highlights the importance of the semantic filtering stages and the cascaded architecture based on embeddings and LLMs, since directly sending more than 2.6 million combinations to large language models would be computationally expensive and impractical in large-scale institutional scenarios. To improve the efficiency and accuracy of affinity mapping results, we adjusted TALENT to consider only mappings classified as *High* or *Medium* affinity, and exclude those classified as *Low* from the final analytical results. This strategy reduces semantic noise and prioritizes stronger and more reliable relations between researchers and themes.

Applying Stage 5 – Human Curation Process: In the Human Curation stage, the mappings automatically classified by the pipeline as **"None"** were prioritized for analysis, totaling **462 dissertations or theses in 2023** and **549 dissertations or theses in 2024**, corresponding to **1,011 mappings with no semantic affinity identified**. In these cases, human curation largely confirmed the absence of a significant thematic relationship, preserving the automatic pipeline's results with-

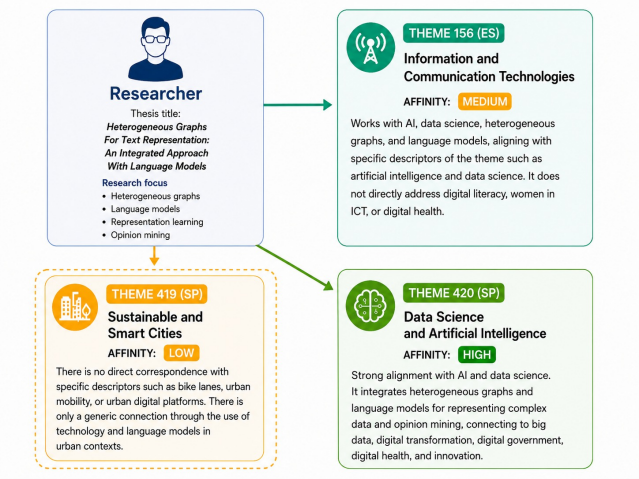

Figure 4. Example of the semantic mapping output for a researcher and three strategic themes. The figure shows the researcher's thesis title, the evaluated strategic themes, the assigned affinity levels (High, Medium, and Low), and a brief rationale for each classification.

out substantive reconsideration. This outcome reinforces the effectiveness of the previous semantic filtering and automatic classification stages performed by embeddings and LLMs.

In addition, human curation conducted sampling-based analyses on the mappings classified with **High** and **Medium** affinity, which represent the most strategically relevant cases for identifying installed capacity in Artificial Intelligence. In 2023, these two categories totaled **26,971 mappings**, while in 2024 they reached **29,196 mappings**, showing increased thematic alignment between the SNPG's academic output and the AI-related strategic themes of the National Agenda – the CAPES strategic thematic priority framework. Based on these samples, the pipeline showed high reliability, indicating semantic consistency between the selected strategic themes and the dissertations and theses automatically classified by the models.

We emphasize that, without human curation, the TALENT pipeline would operate exclusively through automated semantic inference mechanisms based on embeddings and LLMs. Although this fully automated approach would significantly increase scalability and computational efficiency for large-scale institutional analyses, it could also increase the risks of semantic misclassification, hallucinations, contextual inconsistencies, and reduced interpretability in borderline or highly interdisciplinary cases. In this context, specialists involved in Human Curation play a fundamental role in validating strategic alignments, increasing institutional reliability, improving explainability and auditability, and ensuring that the generated mappings remain semantically coherent, trustworthy, and suitable for evidence-based governance and public policymaking.

Output Analysis – Distribution of Human Resource Education in AI: As each researcher can be assigned to multiple strategic themes simultaneously, when counting only the researchers, TALENT found 11,830 out of 91,463 (12.9%) graduates with an affinity for AI-related strategic themes in 2023 (Fig. 5), and 12,459 (13.57%) out of 91,833 in 2024 (Fig. 6). These visualizations show the total number of researchers distributed among the Brazilian federative units and regions that mapped to at least one of the 28 strategic themes related to *Artificial Intelligence*, distinguishing Master's and Ph.D. graduates within each bar for 2023 and 2024, respectively. Each bar represents a Brazilian Federative Unit (*a Brazilian State*) and is grouped into categories. The hatched segment represents the number of Ph.D. graduates, while the unhatched segment represents the number of master's graduates associated with at least one AI-related strategic theme of the respective state. For example, the State of São Paulo (SP) has the largest number of mapped researchers in Brazil, totaling 1,698 master's graduates and 924 Ph.D. with semantic affinity to at least one of the 28 strategic themes of Artificial Intelligence.

Table 3 presents the total number of researchers graduated in 2023 and 2024, and the number of researchers after filtering them (i) by the alignment of their work with AI strategic themes – Number of researchers aligned to AI strategic themes (#Res.–AI themes); (ii) and by the ones from the 88 Computer Science (CS) programs spread over the country – Number of researchers aligned to AI strategic themes from CS graduate programs; (iii) and by the public CS programs – Number of researchers aligned to AI strategic themes from public Computer Science graduate programs; and (iv) or by the private CS programs – Number of researchers aligned to AI strategic themes from private Computer Science graduate programs. These results indicate that approximately one in ten AI-mapped graduates was trained in a graduate program belonging to the Computer Science evaluation area, while nearly 90% originated from programs in other knowledge domains, reinforcing the interdisciplinary diffusion of AI across the Brazilian SNPG.

Table 3. Number of researchers (Graduates, the entire analyzed population), in 2023 and 2024, and numbers (and percentages) after applying filters. **AI Mapped** means the graduates mapped in at least one AI theme. **CS PPGs** means the Computer Science (CS) graduates mapped in at least one AI theme. **Public/Private PPGs in CS** means the Computer Science (CS) graduates in public or private universities mapped to at least one AI theme.

Criterion	2023	2024
Graduates	91,463	91,833
AI Mapped	11,830 (12.9%)	12,459 (13.57%)
CS PPGs	1,168 (9.87%)	1,296 (10.40%)
Public CS PPGs	1,029 (88.10%)	1,132 (87.35%)
Private CS PPGs	139 (11.90%)	164 (12.65%)

We also analyzed the distribution of researchers among major knowledge areas - Table 4. We considered the CAPES division to cluster the graduate programs in major knowledge areas: Applied Social Sciences; Multidisciplinary; Exact and Earth Sciences; Engineering; Health Sciences; Human Sciences; Biological Sciences; Agricultural Sciences; and Linguistics, Literature, and Arts.

6.3 Answering our RQs

Answering RQ1 – Which strategic themes and macro-themes of the CAPES National Agenda explicitly adhere to Artificial Intelligence (AI)? – TALENT discovered that 28 of 470 themes align with AI. It's worth noticing in Table 1 that discovering this set is not an easy task for a human manually

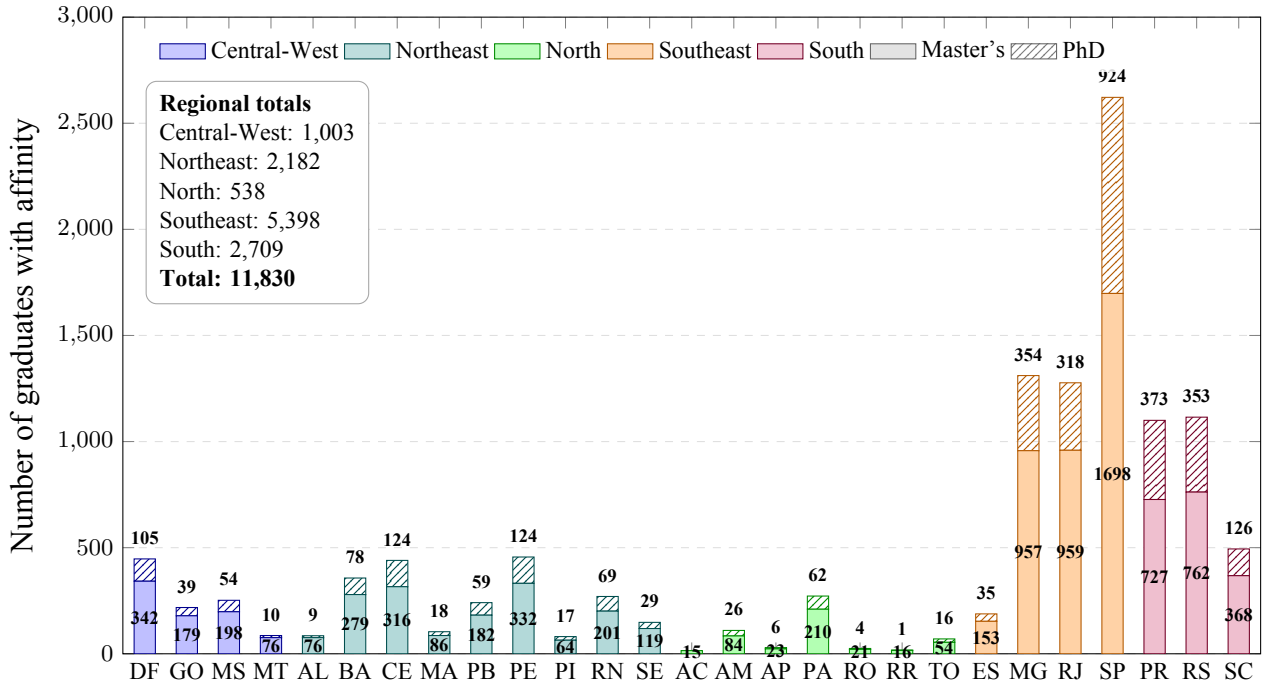

Figure 5. Distribution of graduates with affinity to AI-related strategic themes by Brazilian federative unit and region in **2023**.

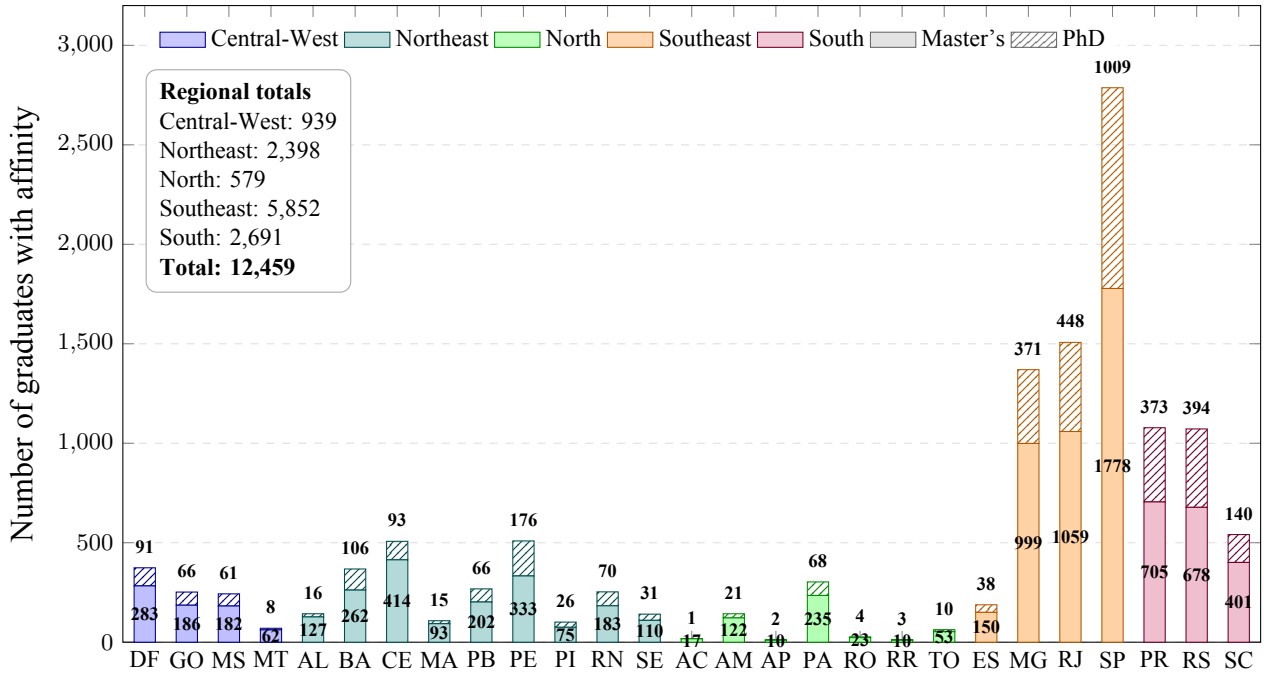

Figure 6. Distribution of graduates with affinity to AI-related strategic themes by Brazilian federative unit and region in **2024**.

conducting it, nor for simple search tools that only conduct syntax matching, when considering that 50% of the mapped themes (14 out of 28) do not explicitly contain the term "Artificial Intelligence" in their titles, although AI-related concepts are present in the themes' associated keywords.

Answering RQ2 – What is the installed capacity of the SNPG in the education of Master's and Ph.D. graduates in Artificial Intelligence in Brazil? –, we could conduct diverse analyses over 2023 and 2024 regarding (i) the distribution of researchers among Brazilian states (Figs. 5 and 6); (ii) number of researchers from Computer Science programs (Table 3); and (iii) the distribution of researchers among large knowledge areas, discussed later, as this is related to RQ4. As expected, in Item (i), we observed the largest number of researchers graduated in the Brazilian Southeast and South regions, respectively. These numbers are expected due to the long-time tradition of many universities in these regions. Nevertheless, it's interesting to observe that the number of researchers who graduated in PE, CE, and BA states in the Brazilian Northeast region, and in the DF in the Central-West region, exceeds 300 graduates per year in AI, which could lead to major innovations in these areas. Regarding Item (ii), it's noticeable that 12.9% of the total researchers concluded their graduate level exploring the AI theme in their academic

Table 4. Number of AI-related graduates by Large Area Of Knowledge.

	2023			2024		
	Master's	**Ph.D.**	***PPGs***	**Master's**	**Ph.D.**	***PPGs***
Applied Social Sciences	2582	785	544	2439	751	552
Multidisciplinary	1484	402	550	1835	556	577
Exact And Earth Sciences	1099	481	216	1169	523	223
Engineering	1039	409	303	1036	440	301
Health Sciences	692	407	424	734	516	415
Human Sciences	839	290	387	827	314	394
Biological Sciences	353	281	160	360	311	176
Agricultural Sciences	314	249	220	290	261	217
Linguistics, Literature, And Arts	95	29	64	62	35	56
Total	8497	3333	2868	8752	3707	2911

work. Given the diversity of themes and the fact that AI can be both an end and a technological means in research, this percentage is astonishing and reflects AI's growing importance across many aspects of human life, including education, work, relationships, and trade. We explore the other aspects of the numbers in Table 3.

Answering RQ3 – To what extent are graduate programs in the Computer Science evaluation area contributing to the installed capacity for educating Master's and Ph.D. graduates in AI-related strategic themes of the CAPES National Agenda? –, we can observe in Table 3 that, in 2023, a total of 1,168 graduated researchers were mapped to the research topic Artificial Intelligence that concluded their work in a CS program, corresponding to 9.87% of the 11,830 researchers which academic works were aligned to AI-related strategic themes. In 2024, this number rose to 1,296 graduates, representing 10.40%. This analysis indicates that Brazil has not only a large number of qualified AI researchers and a high level of installed capacity, but also that approximately 1 in 10 were trained in software development, both developing new AI technologies and incorporating AI into software solutions. The analysis also highlights that most researchers mapped to AI-related themes within Computer Science graduate programs are concentrated in Brazilian public institutions, accounting for 88% of mapped researchers in 2023 and 87% in 2024, underscoring the importance of public academic institutions in building Brazilian AI capacity.

Answering RQ4 – Which major knowledge areas of the SNPG are contributing to the education of Master's and Ph.D. graduates in the AI-related strategic themes mapped within the CAPES National Agenda? – Table 4 shows that the education of Master's and Ph.D. graduates associated with AI-related strategic themes in the CAPES National Agenda is broadly distributed across multiple major knowledge areas of the SNPG, extending far beyond the traditional boundaries of Computer Science programs. The numbers also indicate that Applied Social Sciences represents the largest concentration of AI-related graduates in both years analyzed, totaling 3,367 graduates in 2023 and 3,190 in 2024, distributed across more than 540 PPGs. This behavior suggests the growing incorporation of AI-related themes into domains such as public administration, economics, management, communication, law, and digital governance. The Multidisciplinary area also grew substantially, increasing from 1,886 graduates in 2023 to 2,391 in 2024, while also expanding the number of PPGs involved. This trend highlights the increasing role of interdisciplinary research environments in fostering AI-related education and innovation across heterogeneous domains.

As expected, Exact and Earth Sciences and Engineering also presented strong participation in AI-related education. In 2024, these areas jointly accounted for more than 3,100 Master's and Ph.D. graduates associated with AI-related strategic themes. These findings show the high potential that such researchers can develop AI-based solutions in the future. Another relevant finding is the significant presence of Health Sciences, Human Sciences, Agricultural Sciences, and Biological Sciences in AI-related mappings. The Health Sciences area alone produced 1,250 AI-related graduates in 2024, demonstrating the rapid expansion of AI applications across digital health, bioinformatics, epidemiology, medical data analysis, and healthcare technologies. Similarly, Agricultural Sciences and Biological Sciences demonstrate the increasing adoption of AI in agribusiness, precision agriculture, environmental monitoring, and biotechnology. So, AI is consolidating itself as a transversal enabling technology present across virtually all major knowledge areas.

Overall, based on the answers to all 4 RQs, we understand that Brazil's AI capacity is strongly interdisciplinary and territorially distributed, reinforcing the strategic importance of promoting and fostering public policies to educate Master's and Ph.D. graduates across multiple sectors and regions of the country. The results also show that the development of AI-related competencies is no longer restricted to the traditional, hard domains of Computer Science, but is increasingly expanding into several knowledge areas that incorporate Artificial Intelligence tools, methods, and applications into their scientific, technological, and innovation processes.

The findings also show that Brazil already has a strong academic infrastructure for AI education and research, although important regional asymmetries persist. These disparities underscore the need for scalable, explainable governance mechanisms that continuously monitor national human capital formation. Thus, integrating embeddings, LLMs, and human curation improves transparency, traceability, explainability, and auditability in large-scale analytical processes

related to higher education governance.

From a policymaking perspective, the results are relevant in assessing alignment with the Brazilian National Strategy for Science, Technology and Innovation to 2024-2034, which recognizes AI as a strategic technology for digital sovereignty, industrial transformation, international competitiveness, and evidence-based public administration.

Finally, although this study focuses on the Brazilian context, the proposed architecture is adaptable to other countries, macro-regions, and institutional ecosystems that maintain strategic taxonomies and graduate education repositories, reinforcing its potential as a scalable framework for AI-assisted governance and evidence-based policymaking.

6.4 Ablation Study

This Section presents our ablation study. Ablation studies are commonly employed in artificial intelligence to investigate the contribution of individual components to the behavior and performance of complex computational systems Meyes *et al*. [2019]. In TALENT, we conduct an ablation study to examine the incremental contributions of embedding-based retrieval, Small-LLM filtering, and Large-LLM refinement.

We conducted a stage-wise ablation analysis to assess the incremental contribution and operational role of the main components of the AI-affinity identification pipeline. The analysis was performed using the 2023 data and compared three decision levels: (i) the initial retrieval based on embedding similarity, in which all retrieved pairs are initially treated as having high affinity; (ii) the subsequent reclassification performed by the smaller model; and (iii) the final assessment performed by the larger model. This analysis quantifies how each stage progressively filters unrelated candidates and refines the classification of potentially relevant cases, providing empirical support for the sequential use of the models in the proposed pipeline. Table 5 and Figure 7 show these ablation studies.

Table 5. Number of researcher–theme pairs assigned to each affinity level after each stage of the proposed pipeline for the 2023 data.

	High	**Medium**	**Low**	**None**
Embedding	161292	0	0	0
Small LLM	13382	14563	27154	104202
Large LLM	14516	11359	1995	75

As shown in Table 5, the embedding-based retrieval stage produced 161,292 researcher–theme pairs, all of which were initially treated as high-affinity candidates. This result illustrates the role of semantic retrieval as a broad candidate-generation mechanism that prioritizes coverage over subsequent semantic refinement. Although 160,000 pairs may initially appear substantial, this number should be interpreted in the context of the full candidate space. Considering 28 AI-related topics and approximately 91,000 qualified individuals, the total number of possible researcher–topic combinations exceeds 2.5 million. However, embedding similarity alone is not sufficiently discriminative to support the final affinity decision. After reclassification by the Small LLM, only 27,945 pairs remained in the positive classes, comprising 13,382 High and 14,563 Medium cases. In contrast, 27,154 pairs were classified as Low and 104,202 as None. Thus, the smaller LLM filtered out 131,356 pairs that showed little affinity. The Large LLM then evaluated the remaining positive candidates and retained 25,875 pairs, including 14,516 classified as High and 11,359 as Medium, while downgrading 2,070 pairs to Low (1,995) or None (75).

In summary, the Small LLM reduced the number of candidate pairs requiring Large-LLM refinement from 161,292 to 27,945, corresponding to an 82.7% reduction. Thus, only 17.3% of the embedding-retrieved candidates required further semantic refinement by the more computationally expensive model. Importantly, these values correspond to pairs rather than unique individuals, since the same graduate may be evaluated against multiple AI themes.

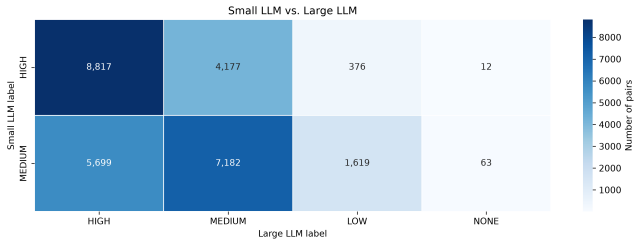

Figure 7. Transition matrix between the affinity labels assigned by the Small LLM and the Large LLM for the candidate pairs retained after the Small LLM stage.

The transition matrix in Figure 7 provides a more detailed view of the differences between the two LLM-based stages and indicates that the larger model acts primarily as a calibration mechanism for the labels assigned by the small LLM. Among the 13,382 pairs initially classified as High by the small LLM, 8,817 remained High, 4,177 were reassigned to Medium, 376 to Low, and 12 to None. Similarly, among the 14,563 pairs classified as Medium, 7,182 remained in the same category, 5,699 were promoted to High, 1,619 were downgraded to Low, and 63 to None. The models therefore produced exactly the same label for 15,999 pairs, corresponding to an exact agreement of 57.2%. However, this agreement becomes substantially stronger when High and Medium are jointly interpreted as indicating meaningful affinity. Under this binary interpretation, 25,875 of the 27,945 pairs that the small LLM considered positive remained positive after evaluation by the large LLM, corresponding to 92.6% agreement. This suggests that the larger model primarily refines the estimated affinity strength within the positive candidate set, while removing only a relatively small fraction of candidates considered positive by the smaller model.

These results support the intended division of responsibilities across the pipeline. The embedding stage prioritizes recall and produces a broad candidate set; the smaller LLM performs the main filtering step, substantially reducing the number of cases requiring further evaluation; and the larger LLM provides a finer-grained reassessment of the remaining candidates. For the subsequent human-curation stage, considering both High and Medium predictions as positive therefore yields a comparatively stable candidate set, with 92.6% agreement between the two LLM-based stages.

Although the differences between the Small and Large LLMs are relatively modest once the candidate set has been

refined, the ablation results highlight the importance of the Small LLM as an intermediate filtering stage. Starting from the 161,292 pairs retrieved by the embedding-based stage, the Small LLM discarded 131,356 pairs as Low or None, preventing more than 100,000 potentially unrelated researcher–theme pairs from being unnecessarily processed by the substantially more expensive Large LLM. This reduction directly lowers computational cost and execution time at scale. At the same time, the embedding stage remains essential because it performs the initial coarse semantic pruning over a candidate space exceeding 2.5 million possible combinations, eliminating pairs that are clearly distant in the semantic space before any LLM inference is required. Therefore, the three stages play complementary roles: embeddings efficiently remove clearly unrelated pairs, the Small LLM reduces the false positives retained by semantic retrieval, and the Large LLM performs a more refined assessment of the remaining potentially relevant cases.

Finally, rather than retraining alternative pipeline architectures, this stage-wise ablation analysis isolates the effect of each successive decision stage by examining how the candidate set and affinity labels change after embedding-based retrieval, Small-LLM screening, and Large-LLM refinement.

6.5 Human Validation Protocol and Mapping Quality Assessment

To evaluate the quality of the mappings produced by TALENT, we selected one postgraduate program in Computer Science as a case study and analyzed all mappings associated with the AI-related strategic themes previously listed in Table 1. We chose one specific graduate program because it allowed us to obtain the curation from the program coordinator of this graduate program. The audited dataset contained 314 unique mappings associated with 106 unique theses and dissertations from 2023 and 2024, including 52 dissertations and 54 theses, that were mapped to the selected AI-related strategic themes. Since each academic work may be associated with multiple strategic themes, the final audit set is represented as mappings in the form *(theme, researcher, affinity, explanation)*. This section therefore presents a case-study evaluation of the human-curation stage on a subset of the dataset.

The postgraduate program coordinator, with domain expertise in Computer Science and Artificial Intelligence, conducted the human curation process with two additional specialists in the field, supported by an *LLM-as-a-judge* strategy.

Recent evidence indicates that LLM-as-a-Judge can provide a scalable and effective strategy for evaluating complex semantic outputs, particularly when the evaluation task is supported by explicit criteria and calibrated against human judgments. Studies show that LLM-based evaluators can achieve agreement levels approaching those of trained experts and, in some settings, substantially reduce evaluation time [Kumar *et al*., 2026; Croxford *et al*., 2025]. However, this reliability is not uniform across tasks, evaluation criteria, languages, and contexts, and therefore does not eliminate the need for expert oversight [Williams *et al*., 2026]. These findings support the hybrid strategy adopted in this study, in which the LLM judge is used as a scalable auxiliary mechanism for inspecting mappings and identifying potentially ambiguous cases, while human experts retain responsibility for validation and for establishing the final reference labels.

We instantiated the judge using Claude Sonnet 4.5 [Anthropic, 2025], selected for its strong performance on public benchmarks among recent LLMs. The LLM judge received the original metadata used by TALENT, the assigned affinity level, and the explanation generated for each mapping. It then produced a Likert-style assessment, an agreement decision, and, when necessary, a suggested corrected level. The human curator used these outputs as decision support, especially to identify mappings with weak semantic evidence or potentially overestimated affinity. We introduced the LLM judge between TALENT's predictions and human curation to reduce the effort required to manually inspect all 314 mappings from scratch. This step is only for this experiment and is not part of our pipeline. Thus, we used the LLM judge to support and standardize the process, while humans supervised the final evaluation.

Figure 8 presents the confusion matrix comparing the original TALENT mappings with the final labels obtained after human curation with LLM judge support. Methodologically, we compare TALENT's output with a human response (after consensus among the three humans), supported by an LLM judge, to simulate the human curation stage of our pipeline. The three specialists reviewed disagreements and reached the final reference label by consensus; therefore, the curated label reflects a human consensus judgment rather than the LLM judge's output. A confusion matrix compares the labels assigned by TALENT with those obtained after human curation, showing on the diagonal cases where both agree and off the diagonal cases where the original mapping level was corrected. TALENT was highly stable for mappings initially classified as *Low*: all six mappings initially classified as *Low* remained *Low* after curation, showing complete agreement for this class within the audited candidate set. Most disagreements occurred between *Medium* and *High*, with 28 mappings upgraded from *Medium* to *High* and 60 mappings adjusted from *High* to *Medium*. These differences have limited practical impact in our setting, since we retain both levels as positive evidence of thematic affinity in the final classification process.

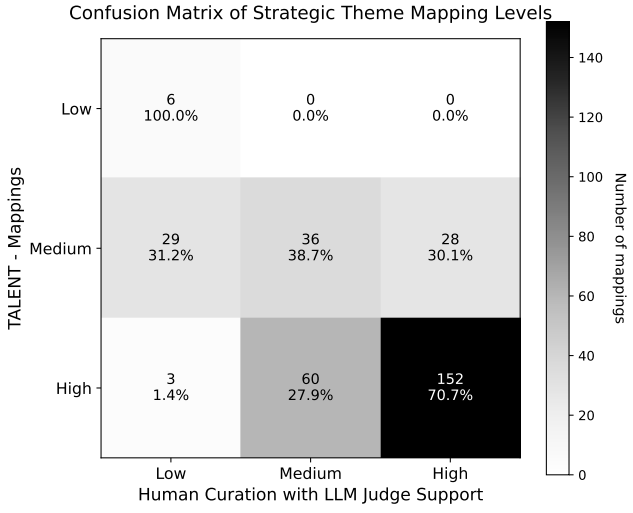

Figure 8. Confusion matrix comparing TALENT mappings with human curation supported by an LLM judge.

The most relevant disagreements concern the *Low* class. In total, 29 mappings originally classified as *Medium* and 3 mappings originally classified as *High* were downgraded to *Low* after curation. These 32 mapping cases indicate possible overestimation by TALENT, where weak or indirect textual evidence may have been interpreted as sufficient thematic affinity. However, these cases do not correspond to 32 distinct academic works, since the same thesis or dissertation may be mapped to multiple strategic themes. In contrast, no *Low* mapping was upgraded to *Medium* or *High*, suggesting that the system favored recall over strict precision for relevant mappings.

When *Medium* and *High* are merged into a single positive class, which corresponds to the criterion used to determine whether a mapping is retained as relevant, TALENT achieves an accuracy of 89.8%. In this binary setting, the system obtained 276 true positives, 6 true negatives, 32 false positives, and no false negatives. Within this post-retrieval audited candidate set, TALENT achieved 89.8% accuracy, 89.6% precision, 100.0% conditional recall, and an f_1-score of 94.5% when Medium and High were merged as positive. These metrics characterize classification performance only within the audited candidate set and must not be interpreted as end-to-end retrieval performance of the complete TALENT pipeline. This behavior is desirable for the proposed application scenario because TALENT relies solely on research metadata, such as titles, abstracts, keywords, research lines, and project descriptions. In this context, a recall-oriented strategy reduces the risk of excluding potentially relevant theses and dissertations from strategic analysis due to incomplete or indirect metadata evidence. The remaining false positives can be addressed in the human curation stage, where specialists review borderline cases and decide whether the semantic evidence is sufficient to support the mapping.

Beyond evaluating individual affinity labels, the curation process also shows how TALENT performs on indicators derived from the mappings. This is important because not every label adjustment affects the final indicators: since both *Medium* and *High* are treated as relevant mappings, changes between these two levels recalibrate the degree of affinity but do not change whether a thesis or dissertation contributes to the indicator. Under this operational criterion, TALENT achieved 89.8% accuracy, with 100.0% recall for relevant mappings. Therefore, the most important corrections cross the relevance boundary between *Low* and *Medium*/*High*, as shown in Figure 9.

Figure 9 shows that most mappings were either preserved or moved only within the positive relevance range. For example, 60 mappings shifted from *High* to *Medium*, and 28 shifted from *Medium* to *High*. Although these changes refine the affinity level, they do not affect the final set of mappings used to compute the indicators. In contrast, only 32 out of 314 mappings (10.2%) produced a real impact on the indicators: 29 mappings were downgraded from *Medium* to *Low*, and 3 from *High* to *Low*. No mapping originally classified as *Low* was upgraded to *Medium* or *High*. This result suggests that TALENT favored recall, preserving potentially relevant mappings for expert review, while human curation primarily removed a small subset of overestimated associations.

This evaluation also focuses on the affinity levels assigned after the candidate mappings have already been produced. Therefore, it does not evaluate the previous semantic filtering stage responsible for retrieving candidate associations between researchers and strategic themes. Previous studies have evaluated that stage with satisfactory performance [Utino and Gôlo, 2025; Gôlo and Utino, 2025; Miranda *et al*., 2025], and we treat it here as an upstream component of the pipeline.

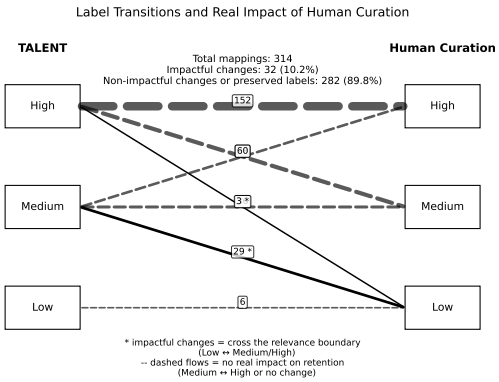

Figure 9. Label transitions between TALENT mappings and human curation. Impactful changes cross the relevance boundary between *Low* and *Medium*/*High*; transitions between *Medium* and *High* do not affect whether a mapping is retained as relevant.

7 Conclusions and Future Work

In this work, we argue that Artificial Intelligence has evolved from a mere object of scientific research in Computer Science, Engineering, and related areas into an indispensable strategic infrastructure for digital governance and evidence-based public policymaking worldwide. However, evaluating the human resource capacity across geographic areas can be very challenging. In this context, we propose TALENT, a pipeline that combines Natural Language Processing, semantic embeddings, LLMs, and human curation to align theses and dissertations with a strategic themes framework and evaluate installed human capacity in a geographic area for a specific topic of interest.

Beyond its technical contribution, our pipeline and empirical evaluation advance the emerging concept of Large Language Models applied to policymaking support in the public sector by showing that the synergy between human and machine reasoning can process massive volumes of unstructured academic data in a computationally feasible manner. This integration guarantees fundamental properties for public administration, such as transparency, explainability, and the auditability of analytical processes. Consequently, the ecosystem transforms complex data into robust strategic intelligence, aligning with global principles of Trustworthy AI and the guidelines of the National Strategy for Science, Technology, and Innovation (ENCTI 2024-2034) to strengthen the country's technological sovereignty.

In the context of Brazilian graduate education, coordinated by CAPES and its Brazilian SNPG, TALENT could produce outputs that demonstrate significant development of AI human capacity in Brazil, which is also deeply interdisciplinary. They also showed that AI growth already extends beyond traditional Computer Science, consolidating AI as a transver-

sal and vital tool for the country's digital transformation and socioeconomic development.

In practice, the results may provide CAPES and other government agencies with precise mechanisms to support high-impact decision-making regarding human-instated capacity in Brazil. Our evaluation focused on AI, but TALENT can be applied to other topics of interest, such as sustainability or climate change, which are important at both the national and global levels. This analytical infrastructure may optimize funding allocation, guide scholarship policies, inform regional development initiatives, and support long-term planning for advanced human resource training. It serves as a planning tool that identifies the scientific capacity installed at each university, state, and region, promoting more competitive growth for the Brazilian innovation ecosystem.

This study has the following main limitations. First, the human-validation experiment was conducted as a case study involving one Computer Science graduate program and should therefore not be interpreted as a representative evaluation of the entire SNPG. Second, the reported classification metrics evaluate mappings after candidate generation and do not measure end-to-end retrieval performance. Third, TALENT estimates installed human capacity from thesis and dissertation metadata and therefore captures only advanced human capital formation, rather than the full multidimensional structure of national AI capacity, which may also involve scientific production, collaboration networks, infrastructure, industry capabilities, and innovation outcomes.

For future work, the research outlines the evolution of the TALENT pipeline by incorporating Knowledge Graphs (KGs) and real-time monitoring systems for national scientific production. Plans are also underway to expand this methodology to other urgent strategic domains, such as energy transition, cybersecurity, digital health, and climate change. Finally, the project is moving toward consolidating the CAPES National Agenda Observatory, an interactive platform with dashboards and predictive analysis for public managers, positioning it as a scalable framework that can be easily adapted to educational and governance repositories in other countries.

Declarations

Authors' Contribution

Marcelo Augusto Santos Turine: Conceptualization, Methodology, Writing – Original Draft, and Writing – Review & Editing. Solange Oliveira Rezende: Methodology, Supervision, and Writing – Review & Editing. Flávia Cristina Bernardini: Writing – Original Draft, Writing – Review & Editing, and Interpretation of Results. Marcos Paulo Silva Gôlo: Software, Formal Analysis, Data Curation, and Writing – Original Draft. Ricardo Marcondes Marcacini: Methodology, Software Validation, and Writing – Review & Editing. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

Acknowledgments

This study was partially funded by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) – Finance Code TED n. 14984/2024 - UFMS/CAPES; and the *Fundação de Apoio à Pesquisa, ao Ensino e à Cultura* (Foundation for the Support of Research, Teaching, and Culture – FAPEC), under grant number 23104.003594/2025-12. The authors also thank the National Council for Scientific and Technological Development (CNPq) for financial support (Grant 317002/2026-0).

Availability of data and materials

The datasets, experimental notebooks, and supplementary results associated with this article are publicly available in the TALENT GitHub repository at `https://github.com/OESNPG/talent`.

Disclosure and AI statement

The authors used ChatGPT, Grammarly, and Gemini to assist in translating the manuscript from Portuguese into English, to improve the understandability and readability of this text, and to support the generation of images present in Figures 1 and 4. After using these services, the authors reviewed and edited the content as needed and take full responsibility for this publication's content.

References

Aggarwal, C. (2022). *Machine Learning for Text*. Springer Publishing Company, Incorporated, United States, 2nd edition. DOI:10.1007/978-3-319-73531-3.

Anthropic (2025). System card: Claude sonnet 4.5. Technical report, Anthropic. Available at: `https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf`.

CAPES (2025). *National Graduate Education Plan 2025–2029*. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasília. Available at `https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/plano-nacional-de-pos-graduacao-pnpg`. Accessed on May 2026.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., *et al*. (2024). A survey on evaluation of large language models. *ACM transactions on Intelligent Systems and technology*. DOI: 10.1145/3641289.

Choi, C. and Yoon, J. (2025). Ai policy in action: the chinese experience in global perspective. *Journal of Policy Studies*, 40(2):1–23. DOI: 10.52372/jps.e685.

Ciarli, T. and Ràfols, I. (2019). The relation between research priorities and societal demands: The case of rice. *Research Policy*, 48(4):949–967. DOI: 10.1016/j.respol.2018.10.027.

Croxford, E., Gao, Y., First, E., Pellegrino, N., Schnier, M., Caskey, J., Oguss, M., Wills, G., Chen, G., Dligach, D., Churpek, M. M., Mayampurath, A., Liao, F., Goswami, C., Wong, K. K., Patterson, B. W., and Afshar, M. (2025). Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*, 8:640. DOI: 10.1038/s41746-025-02005-2.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

EIT Digital (2021). European digital infrastructure and data sovereignty: A policy perspective. Policy report, European Institute of Innovation and Technology (EIT). Available at `https://www.eit.europa.eu/sites/default/files/eit-digital-data-sovereignty-full-report.pdf`. Accessed on May 2026.

European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 May 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. Available at `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689`. Accessed: May 2026.

Gôlo, M. P. S. and Utino, M. Y. R. (2025). One-class lightweight interpretable filtering for academic profiles and strategic themes affinity. In *2025 IEEE/SBC 37th International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, pages 147–154. IEEE. DOI: 10.1109/SBAC-PADW69789.2025.00028.

Karacapilidis, N., Kalampokis, E., Giarelis, N., and Mastrokostas, C. (2024). Generative AI and Public Deliberation: A framework for LLM-augmented Digital Democracy. In *Proceedings of the EGOV-CeDEM-ePart Conference*, pages 368–379, Ghent/Leuven, Belgium. CEUR Workshop Proceedings. Available at: `https://ceur-ws.org/Vol-3737/paper33.pdf`.

Kumar, A., Poungpeth, N., Yang, D., Farrell, E., Lambert, B. L., and Groh, M. (2026). When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8:173–185. DOI: 10.1038/s42256-025-01169-6.

Li, Y., Zhang, Y., Du, Z., and Guo, Z. (2024). Large language model data augmentation for text-pair classification tasks. In *Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition*, pages 427–433. DOI: 10.1145/3704323.3704362.

MCTI (2024). Plano brasileiro de inteligência artificial (PBIA 2024–2026): IA para o Bem *(In Portuguese)*. Available at `https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial`. Accessed on May 2026.

MCTI (2026). *Proposal for the National Strategy for Science, Technology and Innovation (ENCTI 2024–2034): Towards a Fair, Developed, and Sovereign Brazil*. Ministry of Science, Technology and Innovation - Center for Strategic Studies and Management, Brasília, DF. Available at `https://www.gov.br/mcti/pt-br/centrais-de-conteudo/publicacoes-mcti/encti/encti-2024-2034.pdf/view`. Accessed on May 2026.

Meyes, R., Lu, M., De Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint*. DOI: 10.48550/arXiv.1901.08644.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv*. DOI: 10.48550/arXiv.2402.06196.

Miranda, A. C. D., Utino, M. Y. R., Gôlo, M. P. S., Dos Santos, M. A. A., and de Souza, M. C. (2025). Reducing costs in large-scale classification: A hybrid bert–llm strategy. In *2025 IEEE/SBC 37th International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*. IEEE. DOI: 10.1109/SBAC-PADW69789.2025.00026.

Mollaoglu, M., Doganer, B., Demirel, H., Balin, A., and Akyuz, E. (2026). Industry 4.0 in the sustainable maritime sector: A componential evaluation with bayesian bwm. *Sustainability*, 18(8). DOI: 10.3390/su18084078.

Nardes, A., Altounian, C. S., and Vieira, L. A. G. (2022). *Public Governance: Brazil's Challenge*. Editora Fórum, Belo Horizonte. Available at: `https://www.estantevirtual.com.br/livro/governanca-publica-07S-6196-000-BK`.

Pulford, J., Price, N., Quach, J. A., and Bates, I. (2020). Measuring the outcome and impact of research capacity strengthening initiatives: A review of indicators used or described in the published and grey literature. *F1000Research*, 9:517. DOI: 10.12688/f1000research.24144.1.

Sakiyama, K. M., Fujimoto, M. L., Santin, R. V., and Rezende, S. O. (2025). Leandl hpc challenge 2025: Applying large-scale model adaptation techniques. In *2025 IEEE/SBC 37th International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, pages 139–146. IEEE. DOI: 10.1109/SBAC-PADW69789.2025.00027.

Techanamurthy, U., Iqbal, M., and Abdul, R. Z. (2025). Industry 4.0 readiness and strategic plan failures in smes: A comprehensive analysis. *PLoS One*, 20(5). DOI: 10.1371/journal.pone.0324052.

The White House (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. Available at `https://tinyurl.com/USAAIact`. Accessed: May 2026.

Truby, J., Dahdal, A., Brown, R., and Ibrahim, I. (2026). Diplomacy in the age of ai: Legal and strategic approaches

to techno-nationalism, regulatory soft power and the ai chips race. *Research in Globalization*, 12:100335. DOI: 10.1016/j.resglo.2026.100335.

Turine, M. A. S., Souto Maior, A. P., and Marcacini, R. M. (2025). *Agenda nacional para a formação de pessoal de nível superior: temas estratégicos declarados pelas unidades da federação do Brasil*. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasília, DF. DOI: 10.21713/9786501765594. Available at: `https://agendanacional.capes.gov.br/assets/book/book.pdf`.

Utino, M. Y. R. and Gôlo, M. P. S. (2025). Efficient filtering with bert embeddings for researcher–topic affinity prediction in hpc pipelines. In *2025 IEEE/SBC 37th International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, pages 155–162. IEEE. DOI: 10.1109/SBAC-PADW69789.2025.00029.

Valiati, L. and Möller, G. (2025). Cultural and creative industries in the fourth industrial revolution: heterodox economic perspectives on the eu's triple transition. *Frontiers in Communication*, 10. DOI: 10.3389/fcomm.2025.1712537.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 2017 Conference on Advances in Neural Information Processing Systems*, pages 5998–6008. Available at:`https://proceedings.neurips.cc/paper_files/paper/2017`.

Williams, G., Rutunda, S., Nzabakira, F., and Mateen, B. A. (2026). Human evaluators vs. LLM-as-a-judge: toward scalable evaluation of GenAI in global health. *npj Digital Medicine*. DOI: 10.1038/s41746-026-02992-w.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763. Available at:`https://proceedings.neurips.cc/paper_files/paper/2019`.

Zupic, I. and Čater, T. (2015). Bibliometric methods in management and organization. *Organizational research methods*, 18(3):429–472. DOI: 10.1177/1094428114562629.