Collective Intelligence with Large Language Models for the Review of Public Service Descriptions on Gov.br

Rafael Marconi Ramos[2,9,*][a], Pedro Carvalho Brom[2,7][b], João Gabriel de Moraes Souza[1,6][c], Li Weigang[2,*][d], Vinícius Di Oliveira[2,8][e], Silvia Araújo dos Reis[1,4][f], Jose Francisco Salm Junior[1,10][g], Vérica Freitas[5][h], Herbert Kimura[1,4][i], Daniel Oliveira Cajueiro[1,3][j], Gladston Luiz da Silva[1][k] and Victor Rafael R. Celestino[1,4,*][l]

[1]*LAMFO - Lab. of ML in Finance and Organizations, University of Brasília, Campus Darcy Ribeiro, Brasília, Brazil*

[2]*TransLab, Department of Computer Science, University of Brasília, Campus Darcy Ribeiro, Brasília, Brazil*

[3]*Department of Economics, University of Brasília, Campus Darcy Ribeiro, Brasília, Brazil*

[4]*Department of Business Administration, University of Brasília, Campus Darcy Ribeiro, Brasília, Brazil*

[5]*School of Business and Management, Uberlandia Federal University, Uberlândia, Brazil*

[6]*Department of Production Engineering, University of Brasília, Campus Darcy Ribeiro, Brasília, Brazil*

[7]*Department of Mathematics, Federal Institute of Education, Science and Technology of Brasília, Campus Estrutural, Brasília, Brazil*

[8]*Federal District Secretariat of Economy, Brasília, Brazil*

[9]*Euro-Americano University Center (Unieuro), Brasília, Brazil*

[10]*University of the State of Santa Catarina, Florianópolis, Santa Catarina, Brazil*

Keywords: Multi-Agent Systems, LLM, MoE, Generative AI in Government, Text Rewriting and Simplification, Gov.br.

Abstract: This paper presents an intelligent multi-agent system to improve clarity, accessibility, and legal compliance of public service descriptions on the Brazilian Gov.br platform. Leveraging large language models (LLMs) like GPT-4, agents with specialized contextual profiles simulate collective deliberation to evaluate, rewrite, and select optimal service texts based on ten linguistic and seven legal criteria. An interactive voting protocol enables consensus-based editorial refinement. Experimental results show the system produces high-quality texts that balance technical accuracy with linguistic simplicity. Implemented as a Mixture of Experts (MoE) architecture through prompt-conditioning and rhetorical configurations within a shared LLM, the approach ensures scalable legal and linguistic compliance. This is among the first MoE applications for institutional text standardization on Gov.br, establishing a state-of-the-art precedent for AI-driven public sector communication.

1 INTRODUCTION

Public service descriptions are a critical interface between governments and citizens. Poorly structured or obscure texts undermine transparency, accessibility, and citizen trust. In Brazil, laws such as Law 13.460/2017 and standards like ISO 24495-1:2023 mandate that public communication be comprehensible, inclusive, and legally compliant. Improving these texts thus has normative and practical implications, impacting citizen satisfaction and the effectiveness of digital service delivery. The Gov.br platform centralizes approximately 5,000 systems from 180 institutions, serving over 180 million users, amplifying the challenge of ensuring standardized, ac-

[a] https://orcid.org/0000-0001-7422-3213
[b] https://orcid.org/0000-0002-1288-7695
[c] https://orcid.org/0000-0003-0685-3082
[d] https://orcid.org/0000-0003-1826-1850
[e] https://orcid.org/0000-0002-1295-5221
[f] https://orcid.org/0000-0002-1646-4454
[g] https://orcid.org/0000-0002-8492-1645
[h] https://orcid.org/0000-0003-3035-9738
[i] https://orcid.org/0000-0001-6772-1863
[j] https://orcid.org/0000-0001-5898-1655
[k] https://orcid.org/0000-0001-9650-2993
[l] https://orcid.org/0000-0001-5913-2997
* Emails of corresponding authors.

Ramos, R. M., Brom, P. C., Souza, J. G. M., Weigang, L., Di Oliveira, V., Reis, S. A., Salm Junior, J. F., Freitas, V., Kimura, H., Cajueiro, D. O., Luiz da Silva, G. and Celestino, V. R. R.
Collective Intelligence with Large Language Models for the Review of Public Service Descriptions on Gov.br.
DOI: 10.5220/0013831100003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 301-312
ISBN: 978-989-758-772-6; ISSN: 2184-3252

cessible communication across diverse entities and audiences with varying cultural and educational backgrounds (De Melo et al., 2024).

This work addresses the challenge of enhancing clarity, accessibility, and legal conformity of Gov.br service descriptions, which often suffer from bureaucratic jargon and inconsistencies. We propose a multi-agent system leveraging a Mixture of Experts (MoE) (Shen et al., 2023) within a single LLM instance. Prompt-based role conditioning simulates specialized agents (technical, creative, critical) that evaluate, rewrite, and select optimal texts based on linguistic and legal criteria, incorporating voting and feedback loops for consensus-driven refinement.

Our approach builds on recent research on deliberative reasoning and multi-agent LLMs. The Tree of Thoughts framework (Yao et al., 2023) shows that exploring alternative thought paths with structured evaluation improves coherence and correctness. Multi-agent debate systems (Guo et al., 2024; Du et al., 2023) demonstrate that collaborative deliberation enhances factuality and mitigates hallucinations. By orchestrating specialized editorial agents followed by a consensus evaluator, our system applies these principles at scale to hundreds of Gov.br service descriptions, advancing the state of the art in AI-driven public sector communication.

The paper is organized as follows: Section 2 reviews related work on LLMs, collective intelligence, and public sector communication; Section 3 details the multi-agent architecture, agent roles, evaluation criteria, and MoE strategy; Section 4 describes prototyping, validation, and tools; Section 5 discusses results, scalability, and societal implications; and Section 6 presents conclusions and future directions.

2 RELATED WORK

Recent advances in large language models (LLMs) have inspired a growing body of research on their application in public sector communication, collaborative text generation and structured evaluation of outputs. This section reviews relevant work across three interrelated areas: (1) the use of LLMs in public administration for text simplification and citizen engagement, (2) multi-agent and deliberative prompting strategies to simulate collective reasoning and (3) evaluation frameworks that incorporate legal, ethical and linguistic dimensions. Together, these strands of literature provide the foundation upon which our proposed system builds.

2.1 State of the Art (2024–2025)

Recent research underscores a convergent trajectory that combines sparse Mixture-of-Experts (MoE) language models with multi-agent orchestration, while public-sector guidelines converge toward stronger governance of generative AI solutions.

Sparse Mixture-of-Experts LLMs. The *Mixtral 8, 7B* model pioneered efficient expert routing in 2024, activating only two specialists per token yet matching or surpassing dense competitors such as Llama 2 70B across several benchmarks (Jiang et al., 2024). OpenAI's *GPT-4o* generalised this routing paradigm to a multimodal setting, sustaining GPT-4-level reasoning with lower latency and cost (OpenAI, 2024). *DeepSeek-V2* (236B parameters) subsequently introduced memory-efficient routing, reporting a 42.5% reduction in training expenditure when compared with dense baselines (AI, 2024). A NAACL-2025 analysis of four popular MoE models further revealed that routers systematically prefer experts with higher output norms and that expert diversity rises with depth, offering practical guidelines for load balancing and expert allocation (Lo et al., 2025).

Multi-Agent LLM Frameworks. LLM-based multi-agent systems have evolved from single-agent prompting to explicitly defined collectives. A 2024 survey introduced a five dimensional taxonomy actors, interaction type, structural topology, strategy and coordination protocol documenting hallucination reductions of up to 30% when agents debate or vote (Guo et al., 2024). Commercial deployments, exemplified by Reflection AI's *Asimov*, leverage cascades of retriever and reasoning agents to tackle enterprise codebases, surpassing single-agent baselines in human preference studies (Reflection.AI, 2025). Early 2025 work extended these ideas with SCIBORG, a finite-state automata memory layer that delivers a 12% gain in task completion over prompt-only baselines (Muhoberac et al., 2025), while an urban-scale survey mapped agent applications in planning, public safety and environmental management, outlining trustworthiness criteria essential for government adoption (Han et al., 2025).

Public-Sector Adoption and Governance. The *State of AI in GovTech 2024* reported that 56% of state and local agencies already pilot generative-AI solutions, primarily in content simplification and citizen chatbots (Center for Public Sector AI, 2024). Internationally, the GOV.UK Design System revised its content-style guidance in 2025 to align with ISO

24495-1, thereby reinforcing plain-language standards for digital government services (Government Digital Service, 2025). Regulatory momentum accelerated in 2025: California became the largest U.S. court system to formalise generative-AI policies, requiring safeguards for confidentiality, bias mitigation and disclosure (Sloan, 2025). The *AI Index 2025* records a 40% year-on-year rise in AI-related regulations and notes that 78% of surveyed organisations now embed AI in daily operations (Stanford Human–Centered AI Institute, 2025).

Implications for `gov.br`. Collectively, these developments indicate a period of *consolidation*: MoE architectures are becoming better understood, multi-agent frameworks are integrating persistent memory and domain specificity and governance mechanisms are crystallising. The prototype proposed for `gov.br`, which combines Mix of Agents (MoA) LLMs with a memory aware multi-agent workflow accords with the technical and regulatory direction set by the 2024–2025 literature.

2.2 LLMs in Public Administration

Large Language Models (LLMs) are increasingly being deployed to support communication, document drafting and information accessibility in public administration. Applications range from text simplification and translation of legalese into plain language to the automation of citizen-facing interfaces (Devaraj and Li, 2023; Sallam and Farouk, 2023). Several public institutions, including Brazil's Gov.br platform, have begun experimenting with natural language processing (NLP) tools to standardise service descriptions and reduce bureaucratic opacity (Melo and Castro, 2023).

LLMs are effective in simplifying complex administrative and legal texts without compromising meaning, particularly when aligned with plain language principles such as those set out in ISO 24495-1 (Guo and Zhang, 2023). These models also contribute to legal drafting and compliance workflows by aligning generated outputs with formal structures and normative standards (Hendrycks, 2023). However, most current implementations rely on monolithic or single-agent pipelines, lacking deliberative collaboration or persona-based specialisation.

The integration of Large Language Models with Back-Translation (LLM-BT), as highlighted in recent research (Weigang and Brom, 2025), presents a significant opportunity to enhance public administration. LLM-BT enables improvements in efficiency, transparency and accessibility across governmental operations. Its capabilities in text validation, translation and scientific terminology standardization address essential demands in legal, regulatory and public communication domains. By leveraging LLM-BT's lightweight, explainable and accurate NLP features, public institutions can streamline workflows, foster citizen engagement and reinforce the principles of good governance. Embracing LLM-BT thus represents a strategic step toward modernizing public services, ensuring they are both effective and equitable.

2.3 Simulated Deliberation and Multi-Agent Architectures

A growing body of research explores the use of simulated deliberation through multi-agent prompting. Role-based prompting, where LLMs assume distinct editorial or evaluative stances (e.g., lawyer, critic, layperson), has been shown to improve diversity and quality in text generation (Schick et al., 2023; Park, 2023). This has led to the emergence of the "model-as-committee" paradigm, in which multiple agents evaluate, refine and vote on candidate responses (Liu, 2023; Du et al., 2023).

Such systems are often organised around deliberation protocols such as majority voting, self-criticism and iterative revision cycles. For instance, Self-Refine applies a critique-and-revise loop to improve coherence and factual accuracy (Madaan et al., 2023b; Madaan et al., 2023a; Chen, 2023). Constitutional AI encodes normative constraints into LLM prompting as constitutional rules that guide iterative corrections (Bai, 2022).

Nevertheless, these approaches have primarily been applied in creative or open-domain tasks. There remains a significant gap in adapting these techniques to institutional communication, particularly in domains requiring structured, legally compliant and citizen-accessible documentation. Few, if any, systems have orchestrated domain-specific personas (e.g., legal analyst, plain language expert) in a coordinated deliberative workflow to improve government service texts.

Recent advancements in artificial intelligence have reinvigorated interest in MoE architectures, in which different subnetworks are activated dynamically based on the input. In this project, a structurally similar strategy is adopted to simulate collective intelligence among domain-specific agents. MoE models partition input space into semantically coherent regions, each handled by expert modules specialized in specific subdomains (Zoph et al., 2022). A dynamic router assigns each input to the most appropriate experts at inference time, enabling compu-

tational efficiency and improvements in output quality. Inspired by this paradigm, our approach orchestrates specialized agents, focused on legal, linguistic and user experience dimensions, into a deliberative workflow tailored to the Gov.br context. Unlike traditional MoE implementations that rely on parameter-isolated submodules, our model simulates specialization through rhetorical conditioning and task-aligned prompting within a shared LLM backbone. This design ensures adaptability to diverse normative contexts while maintaining semantic cohesion and interpretability.

In parallel, recent empirical studies from the University of Brasília have explored domain adaptation of LLMs to Portuguese and public administration contexts, yielding promising results (Oliveira et al., 2024; De Melo et al., 2024). These works demonstrate significant reductions in token error rates and enhanced coherence in text revision tasks, thereby supporting the choice of models employed in this study.

2.4 Evaluation Frameworks for Natural Language Generation

Evaluating LLM-generated outputs poses a persistent challenge, especially when outputs must satisfy regulatory, ethical and linguistic constraints. Conventional metrics like BLEU and ROUGE (Papineni et al., 2002; Lin, 2004) are insufficient for capturing clarity, structural organisation or legal adequacy. Recent benchmarks, such as HELM (Bommasani et al., 2023; Liang et al., 2022), propose multi-dimensional evaluation frameworks to assess accuracy, robustness and fairness. Similarly, models guided by constitutional principles are evaluated for alignment with human feedback and regulatory expectations (Bai, 2022). Complementing these approaches, anonymous crowd-sourced pairwise comparisons of LLM outputs, such as those collected on the LM Arena leaderboard (LMArena, 2025), provide an alternative mechanism to evaluate human preferences across multiple models and tasks, highlighting strengths and weaknesses that conventional metrics may overlook.

In public communication, evaluation frameworks based on plain language laws, such as the Plain Writing Act or ISO 24495-1, emphasise readability, tone, inclusion and ethical standards (Action, P. L. and Network, I., 2021). Metrics like BERTScore (Zhang, 2020) and BLEURT (Sellam, 2020) are increasingly used to assess semantic fidelity and pragmatic quality in text generation tasks.

2.5 Contribution and Research Gap

While the literature illustrates the potential of LLMs in text simplification, deliberative generation and structured evaluation, few systems integrate all three dimensions in a cohesive architecture for institutional review. In particular, no known frameworks simulate collective intelligence through deliberative agent roles for revising public service descriptions with embedded legal and linguistic compliance.

This work addresses that gap by introducing a multi-agent LLM framework that simulates deliberation among specialised rhetorical profiles. Our system operationalises ISO 24495-1 and Brazilian Law 13.460/2017 as normative anchors for rewriting and evaluation, offering a novel integration of collective reasoning, legal alignment and plain language enforcement in digital governance. The design further draws inspiration from recent advances in structured reasoning in LLMs, such as the Tree of Thoughts framework (Yao et al., 2023), which demonstrates the effectiveness of deliberative search and multi-step evaluation in improving coherence and task performance.

3 METHODOLOGY

The proposed system is designed to automate the revision of public service descriptions by leveraging language models configured to simulate collective reasoning. This section details the architecture, workflow and evaluation procedures adopted. The methodology is organized into six sequential modules: (1) acquisition and embedding of service data; (2) semantic retrieval based on user queries; (3) evaluation of textual quality according to linguistic and legal criteria; (4) iterative rewriting through simulated agents; (5) automated cross-evaluation with consensus voting; and (6) interactive user validation and feedback. Throughout the entire pipeline, the system adheres to legal and normative standards such as the Brazilian Law no. 13.460/2017 (Brasil, 2017), ISO 24495-1 (Brasil, 2023) and the plain language framework proposed by Fischer (Fischer, 2022).

Figure 1 summarizes the full processing pipeline from data acquisition to user interaction. The system begins by consuming data from the official *Gov.br* API (publicly available at https://www.gov.br/pt-br/api) and a structured URL list in CSV format. Valid service links are filtered and corresponding HTML pages are downloaded in parallel, with redirection handling. These pages are parsed, extracting relevant information such as service title, description, steps,

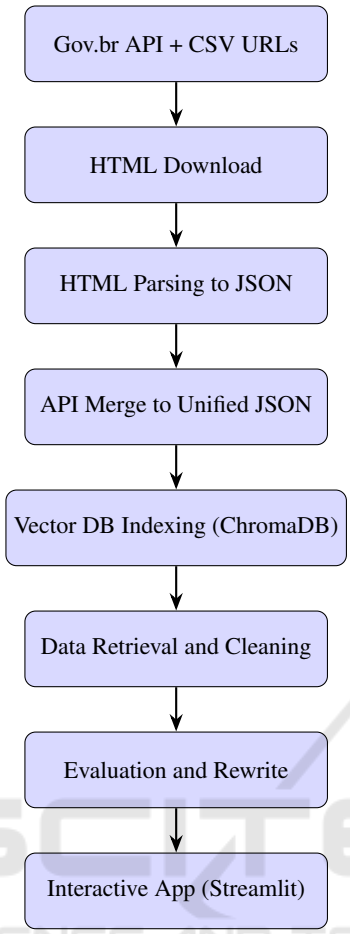


Figure 1: Overview of the end-to-end processing pipeline, from data acquisition through the Gov.br API to interactive evaluation and rewriting.

eligibility and legal references.

The extracted content is then merged with structured fields from the API to form a unified JSON dataset. This dataset is transformed into dense vector representations via the *text-embedding* model and indexed in a vector database optimized for semantic search with embedded metadata. Each record includes document text and fields such as title, category, contact and canonical sections (e.g., *What is it?*, *Who can use it?*). This semantic indexing facilitates fast and context-aware retrieval for downstream rewriting tasks.

3.1 Architectural Overview and Motivation

The proposed system adopts an iterative architecture designed to simulate collective intelligence through a Large Language Model (LLM) configured with distinct evaluative personas. The core objective is to automate the review and rewriting of public service descriptions from the Gov.br platform, ensuring alignment with principles of plain language and legal conformity.

The methodological foundation of this prototype draws upon the theoretical framework of Simulated Agents with Graded Evaluation in Iterative Loop (SAGE-ILoop), a protocol that configures multiple virtual agents with contrasting cognitive profiles (e.g., technical, creative, critical) to evaluate and revise text proposals. The text revision strategy is grounded in the principles of the Brazilian framework *Método Comunica Simples*, developed by Fischer (Fischer, 2022), which advocates for clarity, empathy and communicative accessibility in public service communication. This approach aligns with both the ISO 24495-1:2023 standard for plain language (Brasil, 2023) and the guidelines established by Brazilian Law no. 13.460/2017 (Brasil, 2017), ensuring that revised content is not only legally compliant but also linguistically accessible to a diverse population.

By consolidating the process into a single-model architecture configured with distinct simulated personas, the system avoids the computational overhead associated with multi-model ensembles, while preserving diversity of judgment through controlled variations in parameters (e.g., temperature, top-p and prompt role-play). These agents perform parallel rewritings followed by mutual evaluation and the most suitable version is selected via majority voting. If evaluator confidence is low or disagreement persists, the system initiates a new iteration of rewriting and evaluation, with a maximum of five cycles.

Figure 2 illustrates the overall evaluation architecture of the proposed system, in which a MoE approach is simulated through a set of specialized rhetorical agents orchestrated within a shared LLM environment. These agents are configured via prompt-based conditioning to assume expert roles focused on different perspectives, legal, linguistic and user experience. Each expert generates a rewritten version of the original service description based on its specialization. This multi-agent process acts as a meta-evaluation layer responsible for comparing the proposed rewrites and selecting the most suitable one according to predefined quality criteria, such as clarity, legal accuracy and accessibility.

The MoE-inspired design leverages diversity in configuration rather than isolated parameters: agents vary in temperature, top-p sampling, rhetorical focus and decision-making strategy, simulating a functional diversity analogous to traditional MoE models. A central router dispatches the input to all rhetorical agents in parallel, while an ensemble of evaluation

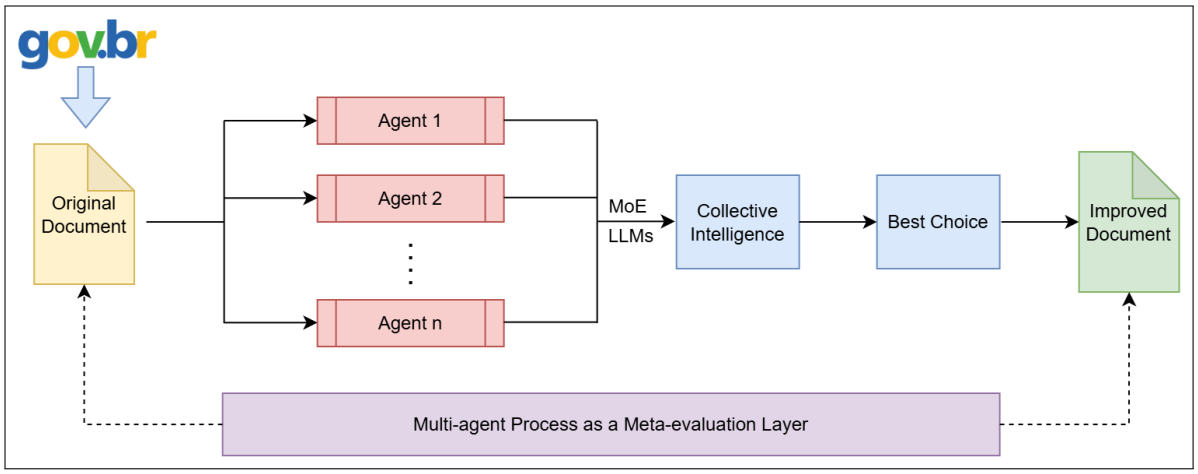


Figure 2: Multi-agent evaluation flow, highlighting the expert agent router and the meta-decision module.

agents performs comparative analysis across outputs. The meta-decision module then synthesizes the evaluators' feedback using a weighted voting strategy, producing a final selection that reflects collective judgment. This design provides interpretability, flexibility and domain-aligned adaptability, crucial factors in institutional contexts like Gov.br, where legal, linguistic and citizen-facing requirements must coexist.

3.2 Semantic Retrieval and User Interaction

Once the service database is indexed, the system enables semantic search based on user queries. The user initiates interaction through a prompt indicating the topic or purpose of the desired document. This input is transformed into a vector representation using the same embedding model employed during indexing, ensuring alignment in the latent space.

The query vector is submitted to the ChromaDB (Chroma Inc., 2025) engine, which returns the top-5 most semantically similar service records based on cosine similarity. These retrieved entries serve as contextual references for subsequent rewriting, ensuring lexical and structural coherence with existing public service descriptions.

The user interface, implemented in *Streamlit* (Streamlit Inc., 2025), presents the most relevant documents and allows the user to select one for editing. The selected content is displayed alongside editable fields, including title, body and an optional legal reference URL. If a valid legal URL is provided, the system extracts its content and appends it to the working text before evaluation. Legal references correspond to URLs of official legislation (e.g., planalto.gov.br) or institutional regulations that underpin the described service.

This interaction paradigm supports guided reauthoring while maintaining flexibility for manual intervention, legal contextualization and iterative refinement.

3.3 Multicriteria Evaluation and Agent-Based Rewriting

Once a service description is submitted or edited, the system initiates a rewriting and evaluation pipeline mediated by simulated agents. These agents are configured with distinct rhetorical and cognitive profiles, each tailored to represent a particular editorial stance. While all agents share a common language model backbone, either *gpt-4o-mini* or *gpt-4.1-nano*, heterogeneity is introduced through controlled variation in prompt conditioning, temperature and sampling strategies, see Table 1. This approach avoids the computational overhead of ensemble methods while preserving diversity in stylistic and evaluative perspectives.

Each agent receives the same structured prompt, designed to guide the rewriting process according to ten evaluation criteria and six plain language guidelines. The prompt enforces a binary assessment for each criterion (true/false), Table 2, followed by brief improvement suggestions where necessary. Subsequently, the agent must produce a rewritten version of the input text, adhering to a standardized structure:

Prompt Schema Excerpt:

```
Evaluate the text below according to the ten criteria
provided with strictly [specialization] bias. Respond
with a JSON in the format

{'1': true, '2': false, ...}
```

```
For each criterion marked as false, provide a brief
comment with suggested improvements. Then, rewrite
the text in accordance with the six guidelines. The
rewritten version must contain the following sections:

- What is it?
- Who can use this service?
- Steps to access the service
- Other Information
- Legislation if applicable
```

The output of this stage consists of multiple rewritten candidates, each annotated with compliance scores derived from the ten binary criteria. These candidates are then subjected to a cross-evaluation and consensus mechanism.

Agent Specializations and Evaluation Criteria

Agent diversity is operationalized through three distinct specializations, each simulating a different editorial perspective:

- **Technical:** emphasizes legal formality, precision and domain-specific terminology;
- **Creative:** focuses on fluency, engagement and accessibility for non-expert users;
- **Critical:** adopts a rigorous reviewer stance, stressing internal consistency and compliance with ethical and structural norms.

These specializations are reinforced by differentiated generation parameters, as shown in Table 1.

Table 1: Simulated Agent Specializations.

Agent Type	Viewpoint	Temp.	Top-p
Technical	As a domain specialist	0.0	0.1
Creative	As a layperson	1.0	1.0
Critical	As a policy evaluator	0.0	0.0

Table 2: Plain Language Evaluation Criteria.

ID	Criterion
1	Respectful and polite language
2	Cultural and social sensitivity
3	Simplicity and accessibility
4	Courtesy and empathy
5	Presumption of user good faith
6	Representativeness and inclusiveness
7	Clarity and structural organization
8	Information security and data protection
9	Transparency and ethical communication
10	Use of plain syntax and active voice

The evaluation criteria span stylistic, structural, ethical and legal-linguistic dimensions. Each criterion contributes equally to the assessment of the rewritten text. Table 2 lists the criteria applied during evaluation.

This integrated framework enables the generation of diverse, guideline-compliant rewritings, systematically evaluated under a unified schema to support high-quality, user-centered public communication.

3.4 Automated Voting and Iterative Refinement

After the submission of a service description, the system initiates an internal rewriting and voting process coordinated by the SAGE-ILoop mechanism. This framework simulates deliberative decision-making using only a single language model instance with varied configurations.

The rewriting stage involves three simulated editor agents, each defined in the system's config list. These agents differ in rhetorical specialization, sampling configuration and viewpoint:

- Technical agent: *gpt-4o-mini*, *as a specialist*, temperature = 0.0, top-p = 0.1;
- Creative agent: *gpt-4o-mini*, *as a layperson*, temperature = 1.0, top-p = 1.0;
- Critical agent: *gpt-4.1-nano*, *as a startup evaluator*, temperature = 0.0, top-p = 0.0.

Each agent rewrites the original text based on a common prompt, generating three distinct candidate versions. These are then passed to a panel of two evaluator agents defined in the *evaluators* list:

- Evaluator A: *gpt-4o-mini*, *as a specialist*, temperature = 0.0, top-p = 0.1;
- Evaluator B: *gpt-4o-mini*, *as a layperson*, temperature = 1.0, top-p = 1.0.

Each evaluator receives all three rewritten texts and is prompted to select the best one based on the ten criteria and six normative guidelines. Their decision and justification are parsed to determine the winning candidate.

Example of Simulated Voting Interaction:

```
Evaluator A (specialist):
- Text 1 is precise but lacks introductory clarity.
- Text 2 is accessible yet imprecise.
- Text 3 is well-structured and normatively sound.

Selected: Text 3
- Rationale: Combines clarity with legal adequacy.
```

```
Evaluator B (layperson):
- Text 1 is overly technical.
- Text 2 is informal but approachable.
- Text 3 is readable, structured and respectful.

Selected: Text 3
- Rationale: Most balanced version for general readers.
```

In the current implementation, the process is executed in a single voting round with forced termination ($k = 5$), without actual iterative retries. However, the architecture is designed to support up to five refinement loops in case of evaluator disagreement or low confidence, as a mechanism that may be activated in future iterations of the system.

The selected version is presented to the user, along with the competing alternatives and their compliance scores, enabling further editing or approval.

3.5 Interactive Feedback and User Validation

After the best rewritten version is selected through the SAGE-ILoop voting process, the system presents the final output to the user in a structured interface built with Streamlit. The user has access to the following elements:

- The final selected version, pre-formatted and downloadable as a `.txt` file;
- All alternative rewritings, each expandable with its corresponding compliance score;
- A panel for user feedback, including star-based rating, open comment field and preferred suggestion selection.

The interface is designed to support both review and iterative editing. If the user is unsatisfied with the selected output, they may return to any of the suggestions and trigger a new editing cycle. Additionally, a field is provided to optionally include a legal reference URL, typically pointing to `planalto.gov.br`; if valid, the corresponding legal text is scraped and included as context in the next evaluation.

User interactions, including ratings, written feedback and chosen version, are stored in a local `SQLite` database via structured insertion commands. The schema captures: (1) original user input; (2) all rewritten suggestions; (3) automated scores; (4) evaluator choice; (5) user-selected version; (6) rating (1 to 5); and (7) textual comments.

This final module closes the loop between system-generated suggestions and human-in-the-loop validation, allowing for both quantitative monitoring and qualitative insight into model performance in public administration contexts.

4 DETAILED EXECUTION: PHASE GUIDE

A detailed guide on the practical implementation of the methodology outlined in 3 is presented below. This includes both the initial prototyping efforts and the validation procedures adopted to test and refine the proposed system.

4.1 Prototyping

This subsection clarifies how the system processes information through successive user-interface screens, ensuring that readers understand the end-to-end data flow. The prototyping phase translated the architectural vision of a deliberative, agent-based rewriting framework into a working application tailored to the Brazilian federal platform *Gov.br*. Development followed a modular, iterative approach that integrates data acquisition, semantic processing, multi-agent evaluation and user interaction components.

A central feature is a simulated collective reasoning protocol that employs a single LLM configured with multiple rhetorical personas. Three agents, *technical*, *creative* and *critical*, were parameterised with distinct generation settings and instructed to evaluate and rewrite public-service descriptions against ten plain-language and seven legal-compliance criteria. These agents generated alternative text versions that were subjected to a structured voting procedure, thereby emulating consensus deliberation.

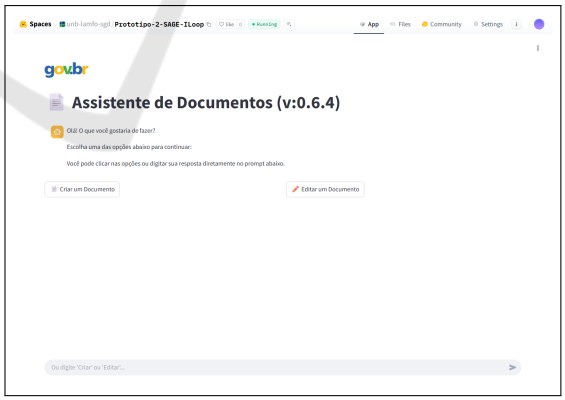


Figure 3: Application home screen with options to create or edit documents.

From a software perspective, the prototype was built with Python and Streamlit for the front-end interface, ChromaDB for semantic retrieval and the OpenAI API for text generation and evaluation. ChromaDB was chosen for its lightweight, open-source design and efficient semantic search across thousands

Figure 4: Similar documents suggested on the basis of the entered theme.

of records. The OpenAI API ensures high-quality Portuguese text rewriting, and the use of small models with short prompts keeps operational costs low and predictable, suitable for public-sector use. The application was deployed as a Hugging Face Space, allowing public access[1] for demonstration and feedback collection. Figure 3 shows the initial interface, which lets users create or edit service descriptions and enter the document theme. Figure 4 presents the list of documents retrieved by similarity. Figure 5 depicts the live Markdown editor used for rewriting the selected description.

Figure 6 summarises the evaluation workflow for a text submitted by the user. After the initial assessment, the interface displays the rewritten versions generated by the simulated agents, each configured with different parameters such as model variant, rhetorical viewpoint, specialisation, top-p sampling and temperature. In the deliberative stage, evaluator agents compare the alternatives and select the most appropriate version according to the defined linguistic and legal criteria. The system then permits resubmission for additional refinement and records user ratings, which can inform future improvements in model behaviour and system performance.

This phase culminated in a robust, operational system capable of ingesting, processing and rewriting real service descriptions from *Gov.br* while preserving full traceability of agent decisions and model outputs. Owing to its modular architecture, the solution can scale to the entire corpus of *Gov.br* texts, supporting large-scale updates and systematic standardisation efforts.

Finally, to better illustrate the practical impact of the proposed approach, Figure 7 presents a side-by-side comparison between an example of the original description and its rewritten version generated by the platform (translated into English). This visual representation highlights how the rewriting process improves clarity, consistency, and adherence to the intended style guidelines.

[1] https://huggingface.co/spaces/unb-lamfo-sgd/Prototipo-2-SAGE-ILoop

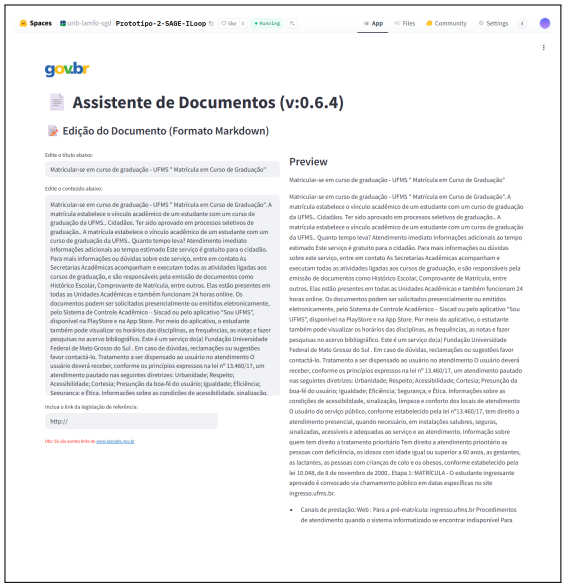


Figure 5: Markdown editor with real-time rewrite preview.

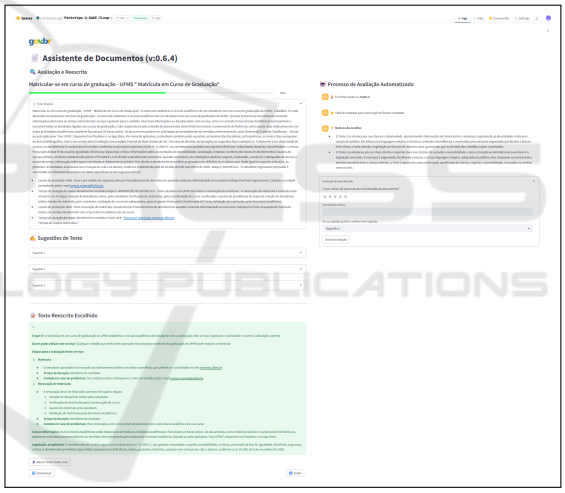
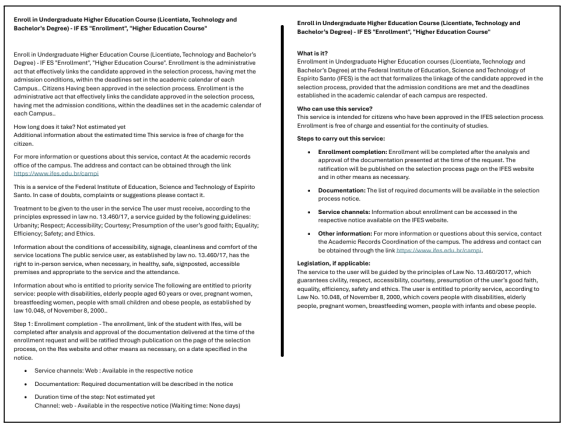


Figure 6: Result of automatic evaluation and rewriting, highlighting the selected best version.

4.2 Validation

Validation activities combine automated metrics with human-in-the-loop feedback to assess the functional adequacy and qualitative performance of the system when rewriting official texts. No quantitative results are reported at this stage; numerical indicators will be added once formal studies are completed.

Internal Compliance Checks. For each candidate rewrite the system applies a rule-based checklist de-



Enroll in Undergraduate Higher Education Course (Licentiate, Technology and Bachelor's Degree) - IF ES "Enrollment", "Higher Education Course"

Enroll in Undergraduate Higher Education Course (Licentiate, Technology and Bachelor's Degree) - IF ES "Enrollment", "Higher Education Course". Enrollment is the administrative act that effectively links the candidate approved in the selection process, having met the admission conditions, within the deadlines set in the academic calendar of each Campus.. Citizens Having been approved in the selection process. Enrollment is the administrative act that effectively links the candidate approved in the selection process, having met the admission conditions, within the deadlines set in the academic calendar of each Campus..

How long does it take? Not estimated yet
Additional information about the estimated time This service is free of charge for the citizen.

For more information or questions about this service, contact At the academic records office of the campus. The address and contact can be obtained through the link https://www.ifes.edu.br/campi

This is a service of the Federal Institute of Education, Science and Technology of Espírito Santo. In case of doubts, complaints or suggestions please contact it.

Treatment to be given to the user in the service The user must receive, according to the principles expressed in law no. 13.460/17, a service guided by the following guidelines: Urbanity; Respect; Accessibility; Courtesy; Presumption of the user's good faith; Equality; Efficiency; Safety; and Ethics.

Information about the conditions of accessibility, signage, cleanliness and comfort of the service locations The public service user, as established by law no. 13.460/17, has the right to in-person service, when necessary, in healthy, safe, signposted, accessible premises and appropriate to the service and the attendance.

Information about who is entitled to priority service The following are entitled to priority service: people with disabilities, elderly people aged 60 years or over, pregnant women, breastfeeding women, people with small children and obese people, as established by law 10.048, of November 8, 2000..

Step 1: Enrollment completion - The enrollment, link of the student with Ifes, will be completed after analysis and approval of the documentation delivered at the time of the enrollment request and will be ratified through publication on the page of the selection process, on the Ifes website and other means as necessary, on a date specified in the notice.

- Service channels: Web : Available in the respective notice
- Documentation: Required documentation will be described in the notice
- Duration time of the step: Not estimated yet
  Channel: web - Available in the respective notice (Waiting time: None days)

Enroll in Undergraduate Higher Education Course (Licentiate, Technology and Bachelor's Degree) - IF ES "Enrollment", "Higher Education Course"

What is it?
Enrollment in Undergraduate Higher Education courses (Licentiate, Technology and Bachelor's Degree) at the Federal Institute of Education, Science and Technology of Espírito Santo (IFES) is the act that formalizes the linkage of the candidate approved in the selection process, provided that the admission conditions are met and the deadlines established in the academic calendar of each campus are respected.

Who can use this service?
This service is intended for citizens who have been approved in the IFES selection process. Enrollment is free of charge and essential for the continuity of studies.

Steps to carry out this service:

- **Enrollment completion:** Enrollment will be completed after the analysis and approval of the documentation presented at the time of the request. The ratification will be published on the selection process page on the IFES website and in other means as necessary.
- **Documentation:** The list of required documents will be available in the selection process notice.
- **Service channels:** Information about enrollment can be accessed in the respective notice available on the IFES website.
- **Other information:** For more information or questions about this service, contact the Academic Records Coordination of the campus. The address and contact can be obtained through the link https://www.ifes.edu.br/campi.

Legislation, if applicable:
The service to the user will be guided by the principles of Law No. 13.460/2017, which guarantees civility, respect, accessibility, courtesy, presumption of the user's good faith, equality, efficiency, safety and ethics. The user is entitled to priority service, according to Law No. 10.048, of November 8, 2000, which covers people with disabilities, elderly people, pregnant women, breastfeeding women, people with infants and obese people.

Figure 7: Left: Original description. Right: Rewritten description generated by the platform (translated into English). This comparison illustrates the improvements in clarity, consistency, and style.

rived from the plain-language guidelines and legal requirements summarised in Table 2. The checklist returns a Boolean trace for every rule, allowing automatic exclusion of drafts that violate mandatory constraints. All traces are stored for audit and reproducibility.

Deliberative Selection and Justification. Drafts that satisfy the checklist enter a voting round conducted by evaluator agents, which produces a ranked list of alternatives and an explanatory note articulating the reasons for the chosen winner. This artefact preserves the rationale for every decision without exposing model internals.

User-Centred Feedback Loop. A web interface built with Streamlit allows end-users to submit topics, inspect the selected rewrite and provide structured feedback via a star rating and optional comment. All interactions are registered in a local `SQLite` database for subsequent qualitative analysis and iterative refinement.

In prototype runs, the average processing time per document — from retrieval to rewrite selection — was approximately a few minutes, depending on document length. This includes parallel rewriting by three agents and consensus evaluation.

Ongoing and Future Validation. Formal usability testing with representative citizen groups is scheduled for the next development cycle. Preliminary pilot sessions with administrative professionals have indicated that the system simplifies bureaucratic communication and enhances user engagement with government services. Quantitative indicators of clarity, comprehension and processing latency will be published once these studies are concluded.

These activities uphold factual validation, transparency, accessibility and legal conformity while deliberately postponing numerical claims until rigorous empirical data become available.

5 DISCUSSION

The results obtained thus far confirm the technical feasibility and institutional relevance of applying Large Language Models (LLMs) to public service communication within the highly structured environment of the *Gov.br* platform. By orchestrating a network of specialised agents, configured as a Mixture of Experts (MoE) within a single LLM backbone, the proposed system ingests, evaluates and rewrites real service descriptions while preserving alignment with plain-language principles and statutory constraints.

Scalability and Adaptability. The integration of semantic retrieval via ChromaDB, prompt-engineered evaluation routines and agent-based rewriting has proven compatible with the heterogeneous corpus of *Gov.br* services. Because all transformations are mediated by prompt logic rather than model fine-tuning, the framework remains adaptable to evolving normative guidelines or domain extensions without retraining overhead.

Diversity Through Rhetorical Specialisation. The distributed specialisation adopted for the agents, varying in sampling parameters, domain focus and rhetorical stance (technical, critical, lay), mirrors the MoE paradigm and yields complementary textual alternatives. This diversity ensures that the final output balances clarity, legal precision and citizen accessibility, thereby addressing the varied literacy and cultural backgrounds of Brazilian users.

Transparency and Collective Deliberation. The structured voting stage, accompanied by mandatory justifications, operationalises simulated collective intelligence. By exposing ranked alternatives and rationales, the system provides an auditable trail that supports accountability and facilitates human oversight, a critical requirement for generative AI in the public sector.

Institutional Impact. Deployment on Hugging Face Spaces enabled preliminary validation of the in-

teraction flow under realistic usage conditions. Early feedback from administrative staff suggests that the solution can help establish a unified communication standard across federal agencies, reducing terminological drift and improving the consistency of user experience across more than 5,000 digital services.

Citizen-Centric Benefits. From the citizen perspective, clearer and more standardised descriptions reduce cognitive load and enhance comprehension, promoting equitable access to information. The inclusion of multiple rhetorical viewpoints ensures that final texts remain legally accurate yet approachable by individuals with diverse educational profiles.

In preliminary tests, the end-to-end pipeline processed dozens of real Gov.br descriptions within a few minutes per document, even when executing three parallel rewrites and evaluation cycles. Because the architecture relies on prompt-engineering and a single shared LLM instance, computational cost grows linearly with the number of documents, making the approach scalable for large-scale deployments. Future work will report detailed latency distributions and throughput metrics.

In sum, the architecture advances beyond conventional text correction: it embeds transparency, inclusiveness and accountability into the fabric of digital public communication. As such, it constitutes a transferable model for other governmental domains seeking to leverage collective, LLM-driven intelligence to standardise and democratise institutional language while remaining fully compliant with applicable norms.

6 CONCLUSION AND FUTURE WORK

This paper introduced a novel multi-agent system that leverages Large Language Models (LLMs) to enhance the clarity, accessibility and legal compliance of public service descriptions published on Brazil's federal Gov.br platform. By simulating collective deliberation among specialised rhetorical agents within a Mixture of Experts (MoE) architecture, implemented via prompt-based conditioning, the proposed framework successfully generates high-quality rewritten texts that balance technical accuracy with linguistic simplicity.

The results reinforce the viability of applying generative AI to institutional communication, demonstrating the system's scalability, adaptability and legal compliance. The distributed agent model, combined with a structured evaluation and voting mechanism, enabled the generation of complementary textual alternatives and the transparent selection of optimal versions. Prototype deployment further confirmed the system's potential for integration into production pipelines for digital public services.

Beyond technical efficacy, the solution establishes a universal communication standard across federal entities, fostering consistency and cohesion in user experiences. It reduces cognitive load, enhances inclusiveness and democratizes access to information, particularly for citizens with diverse cultural and educational backgrounds. By aligning cutting-edge AI with principles of public interest design, this architecture transcends conventional text simplification and contributes meaningfully to transparency, inclusion and institutional accountability.

Future developments will focus on formal usability studies with end-users to refine interaction paradigms and measure real-world impact. Additionally, the system architecture supports up to five iterative refinement cycles, which may be activated in scenarios involving evaluator disagreement or low confidence, further enhancing robustness and output precision. This work sets a precedent for deliberative LLM systems integration into public governance and offers a replicable model for institutional communication in diverse governmental contexts.

REFERENCES

Action, P. L. and Network, I. (2021). Plain writing act guidelines. https://www.plainlanguage.gov. Accessed: 01 jun. 2025.

AI, D. (2024). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Bai, Y. e. a. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bommasani, R., Liang, P., and Lee, T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Brasil (2017). Lei nº 13.460, de 26 de junho de 2017. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13460.htm. Provides for the participation, protection, and defense of the rights of users of public administration services. Accessed: 01 jun. 2025.

Brasil (2023). Iso 24495-1:2023 - plain language — part 1: Governing principles and guidelines. First edition.

Center for Public Sector AI (2024). The state of ai in govtech 2024. Technical report, Center for Public Sector AI.

Chen, M. e. a. (2023). Self-refine: Iterative refinement with self-critique. *arXiv preprint arXiv:2303.17651*.

Chroma Inc. (2025). Chroma: An open-source embedding database. https://www.trychroma.com/. Accessed: 2025-09-10.

De Melo, M. K., dos Reis, S. A., Di Oliveira, V., Faria, A. V. A., de Lima, R., Weigang, L., Salm Junior, J., de Moraes Souza, J. G., Freitas, V., Brom, P. C., et al. (2024). Implementing ai for enhanced public services gov. br: A methodology for the brazilian federal government. In *Proceedings of the 20th International Conference on Web Information Systems and Technologies*, pages 90–101.

Devaraj, S. and Li, M. (2023). Leveraging large language models for government communication. *Digital Government: Research and Practice*.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Fischer, H. (2022). *Método Comunica Simples: Como usar linguagem simples para transformar o relacionamento com o cidadão*. Comunicado Simples, Rio de Janeiro. ISBN 9786589652202.

Government Digital Service (2025). Government digital service style guide — updated 18 july 2025. https://www.gov.uk/guidance/style-guide/. Accessed: 18 jul. 2025.

Guo, T., Chen, X., Wang, Y., et al. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of IJCAI 2024*.

Guo, Y. and Zhang, T. (2023). Text simplification with large language models: A study on legal and administrative texts. *Transactions of the Association for Computational Linguistics*.

Han, J., Ning, Y., and Yuan, Z. t. (2025). Large language model powered intelligent urban agents: Concepts, capabilities, and applications. *arXiv preprint arXiv:2507.00914*.

Hendrycks, D. e. a. (2023). Aligning language models to follow legal and ethical norms. *NeurIPS*.

Jiang, A. Q., Sablayrolles, A., and et al. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Liang, P. et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y. e. a. (2023). Chain-of-thought hub: Voting and deliberation with llms. *ACL Findings*.

LMArena (2025). Lm arena leaderboard. https://lmarena.ai/leaderboard. Evaluation through anonymous, crowdsourced pairwise comparisons of LLM tools.

Lo, K. M., Huang, Z., Qiu, Z., Wang, Z., and Fu, J. (2025). A closer look into mixture-of-experts in large language models. In *Findings of NAACL 2025*, pages 4427–4447.

Madaan, A., Gupta, S., and et al. (2023a). Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023b). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Melo, R. and Castro, A. (2023). Gov.br simplification protocols using ai. Whitepaper, Ministério da Gestão e da Inovação, Brasil.

Muhoberac, M., Parikh, A., and Vakharia, N. t. (2025). State and memory is all you need for robust and reliable ai agents. *arXiv preprint arXiv:2507.00081*.

Oliveira, V. D., Bezerra, Y. F., Weigang, L., Brom, P. C., and Celestino, V. R. R. (2024). Slim-raft: A novel fine-tuning approach to improve cross-linguistic performance for mercosur common nomenclature.

OpenAI (2024). Hello gpt-4o. https://openai.com/index/hello-gpt-4o. Accessed: 1 jun. 2025.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Park, J. e. a. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Reflection.AI (2025). Introducing asimov: the code research agent for engineering teams. https://reflection.ai/blog/introducing-asimov. Accessed: 18 jul. 2025.

Sallam, M. and Farouk, H. (2023). A review of large language models in public sector applications. *AI and Society*.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Sellam, T. e. a. (2020). Bleurt: Learning robust metrics for text generation. *ACL*.

Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., Chung, H. W., Zoph, B., Fedus, W., Chen, X., Vu, T., Wu, Y., Chen, W., Webson, A., Li, Y., Zhao, V., Yu, H., Keutzer, K., Darrell, T., and Zhou, D. (2023). Mixture-of-experts meets instruction tuning:a winning combination for large language models.

Sloan, K. (2025). California court system adopts rule on ai use. *Reuters*.

Stanford Human–Centered AI Institute (2025). The 2025 ai index report. Technical report, Stanford Human–Centered AI Institute.

Streamlit Inc. (2025). Streamlit: The fastest way to build data apps in python. https://streamlit.io/. Accessed: 2025-09-10.

Weigang, L. and Brom, P. C. (2025). Llm-bt: Back-translation as a framework for terminology standardization and dynamic semantic embedding. *arXiv preprint arXiv:2506.08174*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Zhang, T. e. a. (2020). Bertscore: Evaluating text generation with bert. *ICLR*.

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., and Fedus, W. (2022). St-moe: Designing stable and transferable sparse expert models.