Centre Énergie Matériaux Télécommunications

ENHANCING ROBUSTNESS AND ASSESSING IDEOLOGICAL INFLUENCE IN RETRIEVAL-AUGMENTED GENERATION SYSTEMS

AMÉLIORATION DE LA ROBUSTESSE ET ÉVALUATION DE L'INFLUENCE IDÉOLOGIQUE DANS LES SYSTÈMES DE GÉNÉRATION AUGMENTÉE PAR RÉCUPÉRATION

Par

Hazem Amamou

Mémoire présenté pour l'obtention du grade de

Maître ès Sciences (M.Sc.) en télécommunications

Jury d'évaluation

Président du jury et examinateur interne	Shervin Vakili Professeur à l'INRS-EMT
Examinateur externe	Richard Khoury Professeur à l'Université Laval
Directeur de recherche	Anderson R. Avila Professeur à INRS-EMT
Codirecteur de recherche	Alan Davoust Professeur à l'UQO

© Droits réservés de Hazem Amamou, Avril 2026

Acknowledgements

I would like to thank my supervisor, Prof. Anderson Avila, for all the support and guidance throughout my academic journey and for the expertise and insights that have been instrumental in shaping the direction of this thesis. I would also like to acknowledge Prof. Alan Davoust, whose insightful discussions have played a crucial role in developing this project.

I would like to express my deepest gratitude to my family, who always supports and encourages me. Their love and dedication have been a constant source of inspiration, and for that, I am truly thankful.

I thank my colleagues, whose collaboration has been a cornerstone of this academic pursuit. A special mention goes to Elmira Salari, whose dedication and collaborative approach have left an indelible mark on this project.

To my circle of friends, both in and outside of academia, I am grateful for the unwavering support, understanding, and moments of levity that have lightened the challenges along the way. Your friendship has been a cornerstone of my personal and academic life.

Résumé

La Génération Augmentée par Récupération (RAG) améliore les modèles de langage (LLM) en intégrant des sources de connaissances externes. Cependant, cette approche présente deux vulnérabilités critiques : un manque de robustesse face aux informations incohérentes ou bruitées, et une sensibilité aux discours idéologiques contenus dans les documents récupérés. Cette thèse aborde ces deux problématiques pour contribuer au développement de systèmes RAG plus fiables et dignes de confiance.

Dans un premier volet, nous évaluons et améliorons la robustesse des systèmes RAG face à des informations externes défaillantes. En nous appuyant sur le benchmark RGB (Retrieval-Augmented Generation Benchmark), nous démontrons qu'une structuration des connaissances sous forme de graphes de connaissances (KG), via le framework GraphRAG, atténue significativement l'impact des documents bruités, contradictoires ou insuffisants. Nos customisations de GraphRAG, notamment l'utilisation combinée des connaissances encodées du LLM et des connaissances externes structurées (GRcomb), améliorent la détection et la correction d'erreurs, particulièrement pour les modèles moins performants comme GPT-3.5.

Dans un second volet, nous investiguons un risque négligé : l'influence idéologique. En appliquant l'Analyse Lexicale Multidimensionnelle (LMDA) à un corpus de débats présidentiels américains, nous identifions six dimensions discursives idéologiques. Nos expériences révèlent que les réponses des LLMs, lorsqu'elles sont générées via un pipeline RAG, s'alignent significativement sur les discours idéologiques présents dans le contexte externe fourni. Ce phénomène est plus marqué pour certains modèles (GPT-3.5) et dépend fortement de la configuration des connaissances (interne vs. externe) spécifiée dans l'incitation.

En synthèse, cette thèse établit que la fiabilité des systèmes RAG repose sur la résolution conjointe des défis de robustesse factuelle et d'influence discursive. Elle propose des pistes pour un cadre unifié intégrant une structuration des connaissances pour la robustesse et une conscience discursive pour l'atténuation des biais, visant des systèmes d'IA augmentée plus résilients et équitables.

Mots-clés Génération Augmentée par Récupération (RAG) ; Robustesse ; Hallucination ; Graphe de Connaissances ; Biais Idéologique ; Analyse de Discours ; Modèles de Langage ; Évaluation ; Fiabilité de l'IA ; Traitement du Langage Naturel

Abstract

Large Language Models (LLMs) are improved by Retrieval-Augmented Generation (RAG), which grounds them in outside knowledge sources. Nevertheless, this paradigm has two serious flaws : it is not robust against noisy or inconsistent data, and it is vulnerable to the ideological discourses found in retrieved documents. This thesis addresses these dual challenges to advance the development of more reliable and trustworthy RAG systems.

In its first part, the research evaluates and enhances the robustness of RAG systems when faced with flawed external information. Using the Retrieval-Augmented Generation Benchmark (RGB), we demonstrate that structuring knowledge via Knowledge Graphs (KGs) using the GraphRAG framework significantly mitigates the impact of noisy, contradictory, or insufficient documents. Our customizations of GraphRAG, particularly the combined use of the LLM's encoded knowledge and structured external knowledge (GR-comb), improve error detection and correction, especially for less capable models like GPT-3.5.

In its second part, the thesis investigates a largely overlooked risk : ideological influence. By applying Lexical Multidimensional Analysis (LMDA) to a corpus of US presidential debates, we identify six ideological discursive dimensions. Our experiments reveal that LLM responses generated through a RAG pipeline show significant alignment with the ideological discourses present in the provided external context. This phenomenon is more pronounced for certain models (e.g., GPT-3.5) and heavily depends on the knowledge configuration (internal vs. external) specified in the prompt.

In synthesis, this work establishes that the reliability of RAG systems hinges on jointly addressing the challenges of factual robustness and discursive influence. It proposes pathways for a unified framework integrating knowledge structuring for robustness and discourse-awareness for bias mitigation, aiming towards more resilient and equitable augmented AI systems.

Keywords Retrieval-Augmented Generation (RAG) ; Robustness ; Hallucination ; Knowledge Graph ; Ideological Bias ; Discourse Analysis ; Large Language Models ; Evaluation ; AI Reliability ; Natural Language Processing

Table des matières

Acknowledgements **3**

Résumé **5**

Abstract **7**

Table des matières **9**

List of Figures **13**

Liste des tableaux **15**

List of Acronyms and Abbreviations **17**

Liste des abréviations **19**

Synthèse **1**
0.1 Introduction . . . 1
0.2 Revue de Littérature et Fondements Théoriques . . . 3
0.2.1 Génération Augmentée par Récupération (RAG) . . . 3
0.2.2 Graphes de Connaissances pour un RAG Amélioré . . . 4
0.2.3 Biais, Équité et Influence Idéologique . . . 4
0.2.4 Analyse du Discours en Linguistique Computationnelle . . . 5
0.2.5 Évaluation de la Robustesse dans les Systèmes RAG . . . 6
0.2.6 Synthèse et Identification du Vide de Recherche . . . 6
0.3 Robustesse dans les Systèmes RAG Augmentés par des Connaissances . . . 7
0.3.1 Méthodologie : Améliorer le RAG avec des Connaissances Structurées . . . 7
0.3.2 Résultats Expérimentaux et Analyse . . . 8
0.3.3 Discussion des Résultats . . . 10
0.4 Influence Idéologique dans les Systèmes RAG . . . 11
0.4.1 Méthodologie : Mesurer l'Alignement Discursif . . . 11
0.4.2 Résultats Expérimentaux et Analyse . . . 12
0.4.3 Discussion des Résultats . . . 13
0.5 Conclusions et Travaux Futurs . . . 14

1 INTRODUCTION **19**
1.1 Context and Motivation . . . 19
1.2 Problem Statement . . . 20
1.3 Research Objectives and Questions . . . 21
1.4 Contributions of the Thesis . . . 22

1.5 List of Publications . . . 23
1.6 Document Structure . . . 24

2 BACKGROUND AND FUNDAMENTAL CONCEPTS 25
2.1 Large Language Models (LLMs) : Architecture and Knowledge . . . 25
2.1.1 Transformer Architecture & Autoregressive Generation . . . 25
2.1.2 Parametric vs. Contextual Knowledge . . . 26
2.2 Retrieval-Augmented Generation (RAG) : A Standard Framework . . . 27
2.2.1 Core Architecture . . . 27
2.2.2 Fundamental Limitations . . . 28
2.3 GraphRAG : Structured Knowledge for Enhanced RAG . . . 28
2.3.1 Knowledge Graphs : Foundations for Structured Retrieval . . . 28
2.3.2 Architecture : A Two-Phase Pre-processing Pipeline . . . 29
2.3.3 Altered Retrieval and Query Answering . . . 29

3 LITERATURE REVIEW AND RESEARCH GAP 31
3.1 Factual Robustness in Retrieval-Augmented Generation . . . 31
3.1.1 Defining and Benchmarking Factual Robustness . . . 31
3.1.2 Architectural Advances for Robustness . . . 32
3.1.3 Graph-Based Retrieval and Knowledge Structuring . . . 33
3.2 Ideological Influence in LLMs and RAG Systems . . . 33
3.2.1 Measuring Social and Demographic Bias in LLMs . . . 33
3.2.2 Analyzing Political and Ideological Bias . . . 34
3.2.3 The Gap : Contextual Bias and the Need for New Measures . . . 34
3.2.4 Lexical Multidimensional Analysis (LMDA) as a Methodological Bridge . . . 35
3.3 Synthesis and Thesis Positioning . . . 35
3.3.1 Integrated Research Gap . . . 35
3.3.2 Thesis Contribution and Positioning . . . 35

4 ROBUSTNESS IN KNOWLEDGE-AUGMENTED RAG SYSTEMS 37
4.1 Introduction . . . 37
4.2 Methodology : Enhancing Rag With Structured Knowledge . . . 38
4.2.1 GraphRAG Framework and Customizations . . . 38
4.2.2 Experimental Setup: RGB Benchmark . . . 40
4.3 Experimental Results and Analysis . . . 41
4.3.1 Noise Robustness Evaluation . . . 41
4.3.2 Counterfactual Robustness Performance . . . 42
4.3.3 Information Integration Capability . . . 43
4.3.4 Negative Rejection Analysis . . . 44
4.4 Discussion of Findings . . . 45

5 IDEOLOGICAL INFLUENCE IN RAG SYSTEMS 47
5.1 Introduction . . . 47
5.2 Methodology: Measuring Discourse Alignment . . . 48
5.2.1 Theoretical Foundation: Lexical Multidimensional Analysis (LMDA) . . . 48
5.2.2 Measurement: Semantic Similarity and Cosine Distance . . . 50
5.2.3 Corpus Construction and LMDA Analysis . . . 50
5.2.4 Experimental Design and Configurations . . . 53

5.3 Experimental Results and Analysis . . . 54
5.3.1 LMDA Dimensions and Dataset . . . 54
5.3.2 Alignment of LLM Responses with Ideological Discourses . . . 55
5.3.3 Interaction Between Internal and External Knowledge . . . 56
5.3.4 GraphRAG Configuration Performance on Ideological Influence . . . 57
5.4 Discussion of Findings . . . 59

6 Conclusions and Future Work **61**
6.1 Summary of Contributions . . . 61
6.2 Derived Publications . . . 62
6.3 Limitations . . . 63
6.4 Implications for Research and Practice . . . 64
6.5 Future Research Directions . . . 64

Bibliography **67**

Appendix **71**
6.6 Complete Prompt Templates for GraphRAG Configurations . . . 71
6.6.1 GR_{RGB} Prompt . . . 71
6.6.2 GR_{ext} Prompt . . . 72
6.6.3 GR_{def} Prompt . . . 72
6.6.4 GR_{comb} Prompt . . . 72
6.7 RGB Benchmark Experimental Parameters . . . 74
6.8 Prompt Templates . . . 74

List of Figures

4.1 Comparative Framework of RGB Benchmark, GraphRAG, and RobustGraphRAG for Evaluating Robustness in RAG. The diagram depicts the main scenarios explored in this work, where the information retrieval comprises positive (relevant/correct) and negative (irrelevant/noisy/incorrect) documents. The custom solution requires not just some level of customization of GraphRAG, but also the use of an enhanced prompt. Results generally show that our approaches mitigate the impact of inconsistency in the retrieved context. 39
4.2 Comparative analysis for the noise robustness task measured by accuracy (%) for GPT-3.5 and GPT-4o-mini under varying noise ratios. 42
4.3 Visual comparison of counterfactual robustness performance across configurations for GPT-3.5 and GPT-4o-mini, showing ACC_{doc}, error detection (ED), and correction rates (CR). . . 43
4.4 Information integration performance across noise ratios for GPT-3.5 and GPT-4o-mini under different configurations. 44
4.5 Negative Claim Rejection Rates for GPT-4o-mini and GPT-3.5: Comparison of Baseline (RGB) and GraphRAG-Augmented Methods (GRdef, GRext, GRcomb) 45

5.1 Retrieval-based question answering where the LLM is provided with a context underlining different discourses for a given topic. 51
5.2 An overview of the pipeline to extract LLM responses under five configurations. The external knowledge is designed to contain documents with a specific ideological discourse. 53
5.3 Overall similarity scores for LLMs responses based on their encoded knowledge and LLM outputs based on RAG. 56
5.4 Overall semantic similarity between responses based on LLMs' encoded and extensive knowledge (LLM+) and responses based on RAG configurations. 57

6.1 GR_{RGB} prompt presented to the LLM, based on the RGB system combined with structured knowledge. 71
6.2 GR_{ext} External-Only GraphRAG Prompt for Robust, Noise-Aware Answering 72
6.3 GR_{def} GraphRAG's Default Prompt for Accurate, Evidence-Based Generation 73
6.4 GR_{comb} Hybrid Prompt for Noise-Robust Question Answering Using Documents and Prior Knowledge 73
6.5 Basic prompt presented to the LLM, characterizing responses based on encoded knowledge only. 74
6.6 Basic prompt presented to the LLM, characterizing responses based on encoded and extensive knowledge. 75
6.7 Contextual prompt used to retrieve LLM responses based solely on the external knowledge. . 75
6.8 Contextual prompt presented to the LLM, characterizing responses based on both encoded and external knowledge. 76
6.9 Contextual prompt presented to the LLM, characterizing responses based on encoded, extensive and external knowledge. 76

Liste des tableaux

1 Configurations de modèles évaluées 8
2 Configurations expérimentales des sources de connaissances 12

4.1 Descriptions of the structured knowledge based approaches assessed in our experiments. . . 40

5.1 US Presidential Debates corpus composition. 51
5.2 Dimension labels and description for the discourses within the U.S. presidential debates corpus. 52
5.3 Descriptions of the types of knowledge interaction used by LLMs to generate responses. . . 54
5.4 Semantic similarity between ideology description and LLMs' responses based on five configurations. 55
5.5 One-Way ANOVA results assessing the effect of configurations on the identified dimensions. 56
5.6 Similarity scores between LLMs' responses (*LLM*) and responses based on three RAG configurations. For *RAG+* and *RAG++* the differences between their results and the result of *RAG* is provided. 56
5.7 Semantic similarity scores for standard RAG and GraphRAG configurations across ideological dimensions. The adaptive GraphRAG version (G-RAG AD) consistently achieves higher alignment with ideological discourses compared to both standard RAG and default GraphRAG. 58
5.8 Lexical similarity (TF-IDF) for standard RAG and GraphRAG configurations across ideological dimensions. The adaptive GraphRAG version (G-RAG AD) consistently achieves the highest lexical alignment, indicating adoption of discourse-specific vocabulary. 59

6.1 Hyperparameter settings for Knowledge Graph construction 73
6.2 Model-specific parameters for answer generation 74
6.3 Question topics used for ideological alignment analysis, with one question per topic and polarity. 77

List of Acronyms and Abbreviations

AI Artificial Intelligence

ANOVA Analysis of Variance

BERT Bidirectional Encoder Representations from Transformers

CR Correction Rate

CRAG Comprehensive Retrieval-Augmented Generation Benchmark

ED Error Detection Rate

FA Factor Analysis

GenAI Generative Artificial Intelligence

GNN Graph Neural Network

GPT Generative Pre-trained Transformer

GR GraphRAG (with various subscripts : RGB, def, ext, comb)

KG Knowledge Graph

LMDA Lexical Multidimensional Analysis

LLM Large Language Model

NLP Natural Language Processing

PDTB Penn Discourse TreeBank

QA Question Answering

RAG Retrieval-Augmented Generation

RAG+ RAG with encoded knowledge

RAG++ RAG with encoded and extensive knowledge

RGB Retrieval-Augmented Generation Benchmark

RST Rhetorical Structure Theory

SPARQL SPARQL Protocol and RDF Query Language

U.S. United States

WER Word Error Rate

Liste des abréviations

ACV Analyse de la Covariance

ANOVA Analyse de la Variance

BERT Représentations d'Encodeur Bidirectionnel à partir de Transformeurs

CR Taux de Correction

CRAG Référentiel Complet de Génération Augmentée par Récupération

ED Taux de Détection d'Erreur

FA Analyse Factorielle

GPT Transformateur Pré-entraîné Génératif

GRA Réseau de Neurones sur Graphes

GR GraphRAG (avec divers indices : RGB, def, ext, comb)

IA Intelligence Artificielle

IAg Intelligence Artificielle Génératrice

KG Graphe de Connaissances

LLM Modèle de Langage à Grande Échelle

LMDA Analyse Lexicale Multidimensionnelle

PDTB Penn Discourse TreeBank

RAG Génération Augmentée par Récupération

RAG+ RAG avec connaissance encodée

RAG++ RAG avec connaissance encodée et extensive

RGB Référentiel de Génération Augmentée par Récupération

RST Théorie de la Structure Rhétorique

SPARQL Protocole SPARQL et Langage de Requête RDF

TAL Traitement Automatique des Langues

TRA Question-Réponse

É.-U. États-Unis

SYNTHÈSE :RENFORCER LA ROBUSTESSE ET ÉVALUER L'INFLUENCE IDÉOLOGIQUE DANS LES SYSTÈMES DE GÉNÉRATION AUGMENTÉE PAR RECUPÉRATION

0.1 Introduction

L'arrivée des grands modèles de langage a bouleversé l'intelligence artificielle. Grâce à eux, les machines comprennent désormais le langage humain, créent des textes et automatisent des tâches d'une complexité inédite, et ce, à une échelle gigantesque. Ils sont le moteur derrière nos assistants numériques et chatbots, qu'ils nourrissent à partir de milliards d'échanges. Actuellement, ils orientent nos choix dans divers domaines tels que la création artistique et l'assistance technique, devenant progressivement des alliés de tous les jours. Toutefois, leurs compétences exceptionnelles sont entravées par des restrictions intrinsèques, y compris les hallucinations factuelles (Ji *et al.*, 2023), la dépendance à des données d'apprentissage potentiellement dépassées (Wang *et al.*, 2023) et l'absence de compétence spécialisée dans certains domaines (Ling *et al.*, 2023). Ces lacunes posent des risques significatifs dans des applications à fort enjeu, érodant la confiance des utilisateurs et entravant le déploiement sécurisé des systèmes basés sur les LLM.

La Génération Augmentée par Récupération (RAG) a émergé comme une solution centrale à ces défis en augmentant dynamiquement les LLMs avec des sources de connaissances externes (Lewis *et al.*, 2020). Ce

paradigme ancre les réponses dans des informations récupérées, pertinentes et souvent actuelles, réduisant la dépendance à la connaissance paramétrique statique et atténuant les hallucinations (Gao *et al.*, 2023). En transformant les LLM de dépôts de connaissances fermés en systèmes de raisonnement basés sur des preuves, le RAG promet de fournir des assistants IA plus précis, fiables et transparents, adaptés à un déploiement dans le monde réel. Sa capacité à accéder et à citer des informations à jour en fait une technologie clé pour construire des applications IA dignes de confiance.

Malgré sa promesse conceptuelle, le cadre RAG introduit de nouvelles vulnérabilités complexes qui menacent sa fiabilité et son déploiement éthique. Ces vulnérabilités se manifestent selon deux dimensions critiques et interdépendantes. La première est la robustesse opérationnelle : les systèmes RAG sont sensibles aux informations bruitées, contradictoires ou insuffisantes récupérées, ce qui peut conduire à des réponses confiantes mais incorrectes (Chen *et al.*, 2024). Les pipelines RAG standards peinent avec le raisonnement multi-sauts (Lin *et al.*, 2021), la résolution de conflits et le refus approprié de répondre lorsque les preuves manquent. La seconde dimension, plus subtile, est l'influence idéologique et discursive : les sources de connaissances externes ne sont pas neutres ; elles véhiculent intrinsèquement des perspectives, des cadrages et des biais (Bender *et al.*, 2021). Lorsqu'un système RAG récupère des textes idéologiquement chargés, comme des débats politiques ou des articles d'opinion il risque d'aligner ses sorties sur ces discours sous-jacents, orientant potentiellement l'opinion des utilisateurs, renforçant des stéréotypes ou propageant des points de vue unilatéraux sous une apparence d'objectivité (Buçinca *et al.*, 2021). Si la recherche a abordé dans une certaine mesure la robustesse factuelle, l'investigation systématique de l'influence discursive au sein du RAG reste gravement sous-développée (Chen *et al.*, 2024; Yang *et al.*, 2025).

Cette thèse adresse donc un problème de recherche double : Comment pouvons-nous développer des systèmes RAG qui soient simultanément robustes aux incohérences factuelles et résilients à un alignement non désiré avec les discours idéologiques présents dans les connaissances externes ? N'aborder qu'un seul aspect laisse les systèmes vulnérables sur l'autre, compromettant l'objectif global d'une augmentation IA digne de confiance. Pour résoudre ce problème, la recherche établit trois objectifs principaux : (1) évaluer comment la structuration des connaissances externes via les Graphes de Connaissances peut améliorer la robustesse face à des informations défectueuses ; (2) développer une méthodologie novatrice pour identifier et mesurer l'influence idéologique dans les systèmes RAG ; et (3) synthétiser ces perspectives vers des principes pour un cadre RAG plus holistique et fiable. Ce travail est guidé par trois questions de recherche centrales qui examinent la performance des personnalisations de GraphRAG (Edge *et al.*, 2024) sur des références de robustesse (Chen *et al.*, 2024), quantifient l'alignement des sorties des LLM avec des discours

idéologiques externes (Berber Sardinha & Fitzsimmons-Doolan, 2025), et explorent les tensions et synergies entre l'amélioration de la robustesse factuelle et l'atténuation de l'influence idéologique. À travers cette double investigation, la thèse apporte des contributions à l'analyse empirique de la robustesse, à l'innovation méthodologique pour la mesure des biais, à la création d'un ensemble de données idéologiques annotées et à la synthèse d'une proposition de cadre pour des systèmes RAG dignes de confiance.

0.2 Revue de Littérature et Fondements Théoriques

Cette section synthétise les travaux fondateurs établissant les deux défis de la robustesse factuelle et de l'influence discursive dans les systèmes de Génération Augmentée par Récupération (RAG). Elle passe en revue l'évolution des architectures RAG, l'incorporation de connaissances structurées via les Graphes de Connaissances (KG), la nature multiforme des biais dans les modèles de langage, et les méthodes computationnelles d'analyse du discours, pour aboutir à l'identification du vide de recherche spécifique que cette thèse adresse.

0.2.1 Génération Augmentée par Récupération (RAG)

Le paradigme de la Génération Augmentée par Récupération, formalisé par (Lewis *et al.*, 2020), a été introduit pour surmonter les limitations des connaissances paramétriques statiques des LLM. L'architecture centrale hybride un récupérateur neuronal dense, qui extrait des documents pertinents d'un corpus externe, avec un transformateur génératif qui conditionne sa sortie sur ce contexte récupéré. Cette conception transforme le rôle du LLM d'un simple stockage de connaissances fermé en un moteur de raisonnement ouvert, capable de citer des preuves à jour et de réduire les hallucinations factuelles ((Lewis *et al.*, 2020) ; (Izacard & Grave, 2021)).

Malgré sa promesse, les systèmes RAG standards font face à des limitations critiques qui compromettent leur fiabilité. Leur performance est très sensible à la qualité de la récupération ; des documents bruités ou non pertinents se propagent directement vers des sorties erronées. Une faiblesse majeure est l'incapacité à effectuer un raisonnement multi-sauts robuste – synthétisant des informations dispersées sur plusieurs documents – en raison de la simple concaténation des passages récupérés ((Chen *et al.*, 2024)). De plus, ces systèmes manquent de mécanismes explicites pour résoudre les contradictions au sein du contexte ou pour refuser de répondre de manière fiable lorsque l'information est insuffisante, un problème exacerbé

par la tendance des LLMs à *l'arrogance générative* (Kadavath *et al.*, 2022). Ces défis soulignent que le RAG standard est une étape nécessaire mais insuffisante vers une IA digne de confiance et ancrée sur des connaissances, nécessitant des innovations architecturales.

0.2.2 Graphes de Connaissances pour un RAG Amélioré

Les Graphes de Connaissances offrent une alternative structurée aux corpus de texte non structuré, représentant l'information comme des réseaux d'entités et de relations. Cette structuration explicite permet une récupération plus précise et facilite le raisonnement logique multi-sauts par le parcours de graphes ((Hogan *et al.*, 2021)). Dans les approches RAG neuro-symboliques, les KG peuvent être utilisés pour récupérer des sous-graphes pertinents, fournissant au générateur un contexte relationnel cohérent qui réduit le bruit et clarifie les connexions ((Sun *et al.*, 2023)).

Une avancée significative dans cette direction est GraphRAG, un framework qui automatise la construction d'un KG à partir d'un corpus de documents ((Edge *et al.*, 2024)). Son pipeline implique l'extraction d'entités et de relations pour construire un graphe, l'application d'un algorithme de détection de communautés pour identifier des clusters thématiques, et la génération de résumés hiérarchiques de ces communautés. Pour une requête, GraphRAG récupère ces résumés synthétisés au lieu du texte brut, visant à fournir un contexte mieux intégré pour des questions complexes. D'autres frameworks explorent la construction de graphes spécifiques à la requête ou l'intégration de KG publics à grande échelle. Le thème unificateur est un passage de la récupération de documents à la récupération de connaissances structurées. Si ces méthodes sont prometteuses pour améliorer la robustesse factuelle, leur performance dans des conditions adverses et leur potentiel à hériter ou même amplifier les biais des textes sources restent des questions ouvertes – un domaine clé d'investigation pour cette thèse.

0.2.3 Biais, Équité et Influence Idéologique

Les biais dans les LLM sont un phénomène bien documenté, découlant de données d'entraînement déséquilibrées, de la conception algorithmique et de boucles de rétroaction interactives (Bender *et al.*, 2021) ; (Weidinger *et al.*, 2021). Ils se manifestent par des préjudices représentationnels, des stéréotypes et des traitements inégaux selon les groupes démographiques. La recherche sur le biais idéologique et politique examine spécifiquement comment les modèles reflètent les orientations présentes dans leurs corpus d'entraînement,

montrant souvent un alignement mesurable avec certaines positions politiques (Feng *et al.*, 2023). Les méthodologies pour détecter ces biais incluent l'analyse de sentiment, le prompting via des questionnaires et des tests de similarité d'embeddings.

De façon cruciale, les travaux antérieurs ont principalement analysé les biais au sein de la connaissance paramétrique du modèle – les biais encodés dans ses poids pendant l'entraînement. La dynamique introduite par le RAG crée un nouveau vecteur d'influence : le biais discursif externe. Lorsqu'un système RAG récupère des textes encadrés idéologiquement, il introduit une nouvelle source de biais, contrôlable, qui peut interagir avec les penchants internes du modèle. L'étude systématique de cette interaction – comment les discours externes façonnent les sorties dans un contexte d'augmentation par récupération – est largement absente de la littérature, représentant un vide critique.

0.2.4 Analyse du Discours en Linguistique Computationnelle

Comprendre l'influence idéologique nécessite d'aller au-delà de l'analyse au niveau de la phrase pour examiner le discours l'utilisation systématique du langage pour mettre en œuvre et promouvoir des visions du monde. En linguistique computationnelle, l'analyse traditionnelle du discours s'est concentrée sur les relations de cohérence et la structure rhétorique. Le concept de discours idéologique, cependant, concerne des schémas d'ordre supérieur de choix lexicaux, de métaphores et de cadrages narratifs qui transmettent implicitement des valeurs et des perspectives (Berber Sardinha & Fitzsimmons-Doolan, 2025).

L'Analyse Lexicale Multidimensionnelle (LMDA) fournit une méthode robuste, basée sur des corpus, pour identifier ces discours idéologiques (Berber Sardinha, 2019) ; (Fitzsimmons-Doolan, 2024). En appliquant une analyse factorielle sur les modèles de co-occurrence lexicale à travers un grand corpus, la LMDA extrait des dimensions discursives latentes (par exemple, une dimension de "responsabilité sociale" versus une dimension "traditionaliste"). Cette méthode a été utilisée en sociolinguistique pour analyser des textes politiques mais n'a pas été exploitée pour évaluer les systèmes d'IA. Cette thèse comble cette lacune méthodologique, en employant la LMDA non seulement comme un outil d'analyse mais comme un moteur pour créer un environnement de test contrôlé pour mesurer comment les sorties des LLM influencées par le RAG s'alignent avec des discours externes identifiables.

0.2.5 Évaluation de la Robustesse dans les Systèmes RAG

Évaluer les systèmes RAG nécessite des références (benchmarks) qui ciblent leurs modes d'échec spécifiques. Les références générales de questions réponses sont insuffisantes. Le RGB par Chen *et al.* (2024) aborde ceci en définissant quatre tâches : la Robustesse au Bruit, la Robustesse aux Contre-Faits, l'Intégration d'Information et le Rejet Négatif. De même, le Comprehensive RAG Benchmark (CRAG) par Yang *et al.* (2025) simule divers environnements de récupération.

Alors que ces références fournissent une base solide pour évaluer la robustesse factuelle, elles opèrent uniquement au niveau de l'exactitude factuelle. Elles manquent du cadre nécessaire pour évaluer la susceptibilité d'un système à l'influence discursive ou idéologique – des scénarios où l'information récupérée est factuellement exacte mais cadrée dans un récit persuasif et biaisé. Ceci crée un angle mort dangereux : un système pourrait obtenir des scores élevés sur les références de robustesse factuelle tout en étant très malléable à un pilotage idéologique. L'absence de protocoles d'évaluation pour cette dimension de fiabilité constitue une lacune significative dans le domaine.

0.2.6 Synthèse et Identification du Vide de Recherche

La littérature révèle deux trajectoires de recherche parallèles mais largement déconnectées. Un premier courant se concentre sur l'amélioration de la robustesse factuelle des systèmes de Retrieval-Augmented Generation (RAG) à travers des innovations architecturales, notamment l'intégration de Graphes de Connaissances. Un second courant s'intéresse aux biais et à l'idéologie dans les grands modèles de langage (LLMs), principalement en tant que propriétés issues de leur connaissance paramétrique interne apprise lors de l'entraînement.

L'intersection critique entre ces deux axes demeure cependant largement inexplorée. En particulier, il reste à comprendre comment les améliorations structurelles conçues pour renforcer la robustesse factuelle se comportent lorsque le « bruit » ou le « contexte » injecté par le mécanisme de récupération est de nature idéologique. Réciproquement, il est encore incertain si des méthodes d'analyse idéologique issues de la linguistique de corpus, telles que la Lexical Multidimensional Analysis (LMDA), peuvent être opérationnalisées pour mesurer l'influence discursive dans le cadre dynamique, contrôlé et contextuel des systèmes RAG.

Cette thèse vise ainsi à combler un triple vide de recherche :

— **Vide empirique** : l'absence d'évaluations systématiques des systèmes RAG enrichis par des Graphes de Connaissances (par exemple, les personnalisations de GraphRAG) dans des scénarios explicitement conçus pour tester l'influence idéologique.
— **Vide méthodologique** : l'absence de lien entre les méthodes d'analyse de discours dérivées de la linguistique de corpus (comme la LMDA) et l'évaluation quantitative et expérimentale des systèmes RAG.
— **Synthèse Manquante** : Le besoin d'une compréhension globale des compromis et des interactions entre le renforcement de la solidité factuelle et la réduction de l'impact idéologique dans l'élaboration et l'évaluation des systèmes RAG.

En menant des investigations parallèles sur la robustesse factuelle et l'influence discursive, puis en en synthétisant les résultats, cette thèse ambitionne de fournir une base conceptuelle et expérimentale plus holistique pour la conception de systèmes de Génération Augmentée par Récupération à la fois fiables, transparents et dignes de confiance.

0.3 Robustesse dans les Systèmes RAG Augmentés par des Connaissances

Ce chapitre présente la première investigation expérimentale centrale, se concentrant sur l'amélioration de la robustesse factuelle des systèmes de Génération Augmentée par Récupération en intégrant des connaissances structurées via les Graphes de Connaissances (KG). L'objectif principal est d'évaluer si, et dans quelle mesure, la personnalisation du framework GraphRAG peut atténuer les vulnérabilités classiques du RAG lorsqu'il traite des informations récupérées bruitées, contradictoires ou insuffisantes.

0.3.1 Méthodologie : Améliorer le RAG avec des Connaissances Structurées

Notre approche est basée sur le framework GraphRAG, qui transforme des documents non structurés en un Graphe de Connaissances structuré pour augmenter la génération ((Edge *et al.*, 2024)). Le pipeline standard comporte deux phases : (1) la Dérivation du Graphe de Connaissances, où un LLM extrait les entités, relations et affirmations de segments de documents pour construire un graphe ; et (2) la Synthèse par Communautés, où l'algorithme de Leiden identifie des clusters d'entités densément connectées et génère des

résumés textuels hiérarchiques pour chaque communauté. À l'inférence, GraphRAG récupère les résumés de communautés pertinents plutôt que les passages de documents bruts, fournissant ainsi au générateur un contexte synthétisé et thématiquement cohérent.

Pour adapter GraphRAG à l'évaluation de la robustesse, nous avons introduit plusieurs personnalisations au niveau des prompts qui créent des configurations expérimentales distinctes, résumé dans Tableau 1. La configuration GR_{RGB} adapte le prompt standard du benchmark RGB pour opérer sur les résumés dérivés du KG. La configuration GR_{def} correspond au prompt original et générique de GraphRAG. Nos contributions méthodologiques principales sont la configuration GR_{ext}, qui impose strictement l'utilisation du contexte externe du KG tout en supprimant la connaissance interne du modèle, et la configuration GR_{comb}, qui demande explicitement au modèle d'intégrer sa connaissance encodée avec le contexte externe du KG.

TABLEAU 1 – Configurations de modèles évaluées

Modèle	Description optimisée
RGB	Modèle RAG de référence tel que défini dans le benchmark RGB.
GR_{RGB}	GraphRAG utilisant le prompt contextuel adapté du benchmark RGB.
GR_{def}	GraphRAG avec son prompt par défaut pour le raisonnement général.
GR_{ext}	GraphRAG limité exclusivement aux connaissances externes issues du graphe de connaissances.
GR_{comb}	GraphRAG combinant connaissances internes encodées et externes du graphe de connaissances.

Toutes les configurations ont été évaluées en utilisant le benchmark Retrieval-Augmented Generation (RGB), qui cible quatre modes d'échec clés : la Robustesse au Bruit, la Robustesse aux Contre-Faits, l'Intégration d'Information et le Rejet Négatif ((Chen *et al.*, 2024)). Pour chaque requête, le Graphe de Connaissances a été construit dynamiquement à partir des mêmes cinq documents récupérés par le système de référence, assurant ainsi une comparaison contrôlée. Les expériences ont été menées avec deux modèles génératifs, GPT-3.5-turbo et GPT-4o-mini, pour analyser la dépendance des gains de robustesse aux capacités du modèle.

0.3.2 Résultats Expérimentaux et Analyse

Robustesse au Bruit

La tâche de robustesse au bruit, qui injecte des documents non pertinents à des ratios de 0,2 à 0,8, a révélé un effet fortement dépendant du modèle. Pour le GPT-4o-mini, plus performant, la référence RGB a montré des performances quasi maximales, maintenant une précision d'environ 98% à 100% même avec un

bruit élevé. Dans ce régime, toutes les configurations GraphRAG ont légèrement moins performé, démarrant autour de 90–95% de précision à faible bruit et dégradant plus rapidement à mesure que le bruit augmentait. Cela suggère que, pour les modèles ayant de fortes capacités de raisonnement et de filtrage internes, l'obligation de s'appuyer sur un contexte externe structuré mais imparfait peut introduire une pénalité de performance modérée.

À l'inverse, l'effet sur GPT-3.5 a été dramatique. La référence RGB s'est effondrée, passant d'une précision de 85% à 0,2 de bruit à environ 10% à 0,8 de bruit. Toutes les configurations GraphRAG ont atténué de manière substantielle cette dégradation, avec GR_{comb} et GR_{def} performant le mieux et maintenant une précision au-dessus de 40% même au plus haut niveau de bruit. Ces résultats démontrent que la synthèse contextuelle basée sur les KG fournit un biais inductif crucial pour les modèles moins capables, leur permettant d'isoler le signal d'un bruit écrasant.

Robustesse aux Contre-Faits

Cette tâche évalue la capacité d'un modèle à détecter et ignorer des documents factuellement incorrects. Pour GPT-4o-mini, la référence RGB était déjà forte, atteignant 71% de précision en présence de contre-faits. Cependant, la configuration GR_{comb} a atteint un taux de détection d'erreur parfait de 100%, bien qu'au prix d'une précision réduite des réponses (35,35%). Cela met en lumière un compromis : le contexte structuré du KG peut augmenter la sensibilité aux incohérences, conduisant parfois à un comportement de rejet trop conservateur.

Pour GPT-3.5, l'amélioration par KG a été transformative. La référence a presque complètement échoué, avec seulement 9% de précision et 7% de détection d'erreurs. En contraste, GR_{comb} a atteint un taux de détection d'erreur de 94,94% et un taux de correction de 95,74%. Cette amélioration d'un ordre de grandeur indique que la structure relationnelle explicite rend les contradictions beaucoup plus saillantes, permettant une vérification de cohérence efficace dans les modèles plus faibles.

Intégration d'Information

Cette tâche nécessite de synthétiser des preuves distribuées sur plusieurs documents. Pour les deux modèles et à tous les niveaux de bruit testés (0%, 20%, 40%), chaque configuration GraphRAG a constamment et substantiellement surpassé la référence RGB. Par exemple, avec GPT-3.5, la précision de référence est

passée de 55% à 34% avec l'augmentation du bruit, tandis que GR_{ext} a démarré à 86,86% et n'est descendue qu'à 78%. Ces résultats confirment une hypothèse centrale : les résumés de communautés de GraphRAG pré-intègrent efficacement l'information dispersée, réduisant la charge cognitive du générateur et facilitant le raisonnement multi-sauts.

Rejet Négatif

Le rejet négatif, où le modèle doit s'abstenir de répondre en raison d'une information insuffisante, s'est avéré être la tâche la plus difficile. La référence RGB a atteint des taux de rejet modérés (environ 35% pour GPT-4o-mini et 25% pour GPT-3.5). La plupart des configurations GraphRAG ont performé en deçà de cette référence, avec une exception notable : GR_{ext}. En imposant une dépendance exclusive au contexte externe, GR_{ext} a atteint les taux de rejet les plus élevés (42,66% pour GPT-4o-mini et 33% pour GPT-3.5).

Ce résultat souligne l'importance de contraindre la connaissance interne. Les configurations qui permettaient le mélange des connaissances internes (GR_{def} et GR_{comb}) ont exhibé des taux de rejet nettement inférieurs (environ 12–20%), reflétant la tendance bien documentée à l'"arrogance générative", où les LLMs hallucinent des réponses plausibles plutôt que de suivre l'instruction de s'abstenir.

0.3.3 Discussion des Résultats

Plusieurs idées clés émergent de ces expériences. Premièrement, l'efficacité du RAG amélioré par KG est fortement dépendante du modèle : elle procure des gains de robustesse substantiels pour les modèles moins performants, tout en offrant des bénéfices plus nuancés et spécifiques aux tâches pour les modèles avancés. Deuxièmement, la stratégie d'intégration des connaissances introduit un compromis fondamental. Les configurations "externes seulement" maximisent la fidélité aux preuves et améliorent le comportement de rejet, tandis que les configurations à "connaissances combinées" excellent dans la synthèse et la correction d'erreurs. Troisièmement, GraphRAG est particulièrement bien adapté aux tâches d'intégration d'information et de détection de contre-faits, mais ne résout pas intrinsèquement le problème du rejet négatif, qui reste très sensible à la conception du prompt et à la conscience de ses propres limites de connaissance.

Finalement, ces résultats établissent une fondation cruciale pour l'investigation ultérieure de l'influence idéologique. Ils démontrent que la représentation structurelle et les contraintes des prompts peuvent puissamment façonner la façon dont les modèles traitent et priorisent l'information factuelle. Le chapitre suivant

s'appuie directement sur cette idée, en se demandant si les mêmes mécanismes qui améliorent la robustesse factuelle modulent, ou potentiellement amplifient, la susceptibilité au cadrage idéologique intégré au sein de la connaissance structurée.

0.4 Influence Idéologique dans les Systèmes RAG

Ce chapitre présente la seconde investigation expérimentale centrale, déplaçant le focus de la robustesse factuelle vers la dimension plus subtile de l'influence idéologique. L'objectif est de mesurer systématiquement comment les discours idéologiques identifiables, intégrés dans les documents récupérés, façonnent les sorties des LLMs au sein d'un cadre de Génération Augmentée par Récupération (RAG).

0.4.1 Méthodologie : Mesurer l'Alignement Discursif

Le fondement de cette étude est un nouveau corpus idéologiquement annoté, dérivé de 45 débats présidentiels américains (1960–2020). Le corpus contient 3428 prises de parole et plus de 575 000 mots. Pour découvrir la structure idéologique latente de ce matériel, nous avons appliqué l'Analyse Lexicale Multidimensionnelle (LMDA), une méthodologie de corpus-linguistique qui utilise l'analyse factorielle pour identifier des dimensions discursives à partir de motifs systématiques de co-occurrence lexicale (Berber Sardinha & Fitzsimmons-Doolan, 2025).

La procédure LMDA a identifié six dimensions idéologiques distinctes, chacune définie par un pôle positif et un pôle négatif représentant des orientations discursives opposées. Par exemple, la Dimension 1 oppose un discours d'"Équité Prospère par la Responsabilité Sociale" (par ex., économie, classe moyenne, soutien, juste) à un discours de "Conservatisme Traditionaliste" (par ex., moral, vie, tradition, valeurs). Collectivement, ces dimensions couvrent les domaines centraux du débat politique, incluant la politique économique, les valeurs sociales, la gouvernance et la sécurité nationale. Elles jouent un double rôle : fournir un corpus étiqueté pour une récupération contrôlée, et fournir des descriptions textuelles concises qui servent de références sémantiques pour mesurer l'alignement idéologique.

Pour quantifier l'influence idéologique, nous avons conçu une expérience contrôlée de type RAG où les sources de connaissances disponibles pour le LLM étaient systématiquement manipulées. Pour chaque dimension idéologique, nous avons formulé des prompts thématiques alignés avec son focus (par exemple,

l'équité économique pour la Dimension 1). Nous avons ensuite évalué cinq configurations de connaissances distinctes, résumées dans le tableau 2.

TABLEAU 2 – Configurations expérimentales des sources de connaissances

Configuration	Source de connaissance
LLM	Connaissances encodées uniquement (pré-entraînement)
LLM+	Connaissances encodées + connaissances étendues (post-entraînement)
RAG	Contexte idéologique externe uniquement
RAG+	Connaissances encodées + connaissances externes
RAG++	Connaissances encodées + étendues + externes

La composante de récupérateur consistait en une recherche sémantique sur une base de données vectorielle contenant uniquement les textes associés à la dimension idéologique cible, assurant que le contexte fourni était discursivement homogène. Quatre LLMs GPT-3.5, GPT-4o-mini, Gemini et Claude – ont été évalués à travers toutes les configurations et dimensions. La métrique d'évaluation principale était la similarité sémantique, calculée comme la similarité cosinus entre les embeddings (basés sur BERT) de la réponse agrégée du modèle et la description officielle dérivée de la LMDA de la dimension idéologique correspondante.

0.4.2 Résultats Expérimentaux et Analyse

Alignement avec les Discours Idéologiques

Les résultats révèlent un schéma clair et constant : les configurations RAG augmentent substantiellement l'alignement des réponses des LLMs avec les discours idéologiques ciblés, par rapport aux bases internes. À travers presque tous les modèles et dimensions, les scores de similarité les plus élevés ont été observés sous les configurations basées sur le RAG (RAG, RAG+, RAG++), plutôt que sous LLM ou LLM+.

Une susceptibilité dépendante du modèle était prononcée. GPT-3.5 est apparu comme le modèle le plus influencé, atteignant les scores de similarité absolus les plus élevés sur plusieurs dimensions, particulièrement sous la configuration RAG+ (par exemple, une similarité de 0,57 pour la Dimension 4). GPT-4o-mini, en revanche, a exhibé son plus fort alignement sous la configuration strictement contrainte RAG (externe seulement), suggérant que sa base de connaissances interne plus riche agit comme une force stabilisatrice à moins d'être explicitement supprimée. Gemini et Claude ont montré un comportement intermédiaire. Une

analyse de variance (ANOVA) a confirmé que le choix de la configuration avait un effet statistiquement significatif sur les scores d'alignement pour cinq des six dimensions.

Interaction entre Connaissances Internes et Externes

Pour mieux comprendre comment le discours externe interagit avec la connaissance interne du modèle, nous avons analysé la similarité entre les réponses générées en utilisant uniquement la connaissance interne (LLM) et celles générées sous l'influence du RAG.

Pour GPT-4o-mini, les configurations qui combinaient les sources de connaissances (RAG+ et RAG++) produisaient des réponses qui restaient plus proches de la base LLM, avec des augmentations de similarité allant de +0,08 à +0,23 à travers les dimensions. Ce schéma indique que GPT-4o-mini a tendance à ancrer ses sorties dans sa connaissance interne, incorporant sélectivement le discours externe plutôt que de le laisser dominer.

À l'inverse, GPT-3.5 a exhibé une dynamique nettement différente. La similarité entre sa base et ses sorties influencées par le RAG était déjà élevée, et le changement incrémental lors de la combinaison des sources de connaissances était minimal (+0,02 à +0,10). Cela suggère que les représentations internes de GPT-3.5 sont soit moins stables, soit plus facilement écrasées, permettant au discours externe de dominer le processus de fusion. Par conséquent, ses réponses sont plus plastiques et plus susceptibles d'être orientées idéologiquement via le contexte récupéré.

0.4.3 Discussion des Résultats

Plusieurs implications importantes découlent de ces résultats. Premièrement, le RAG agit comme un puissant conduit pour l'influence idéologique : fournir un contexte récupéré n'est pas une augmentation neutre mais un acte de cadrage discursif qui façonne de manière mesurable les sorties des modèles. Deuxièmement, la vulnérabilité à cette influence est fortement dépendante du modèle, soulevant des inquiétudes que les modèles moins performants ou plus compliants puissent être disproportionnellement susceptibles à un pilotage idéologique. Troisièmement, l'interaction entre robustesse et influence révèle une tension de conception critique. La même configuration "externe seulement" (RAG) qui maximise l'alignement discursif s'est aussi avérée dans le Chapitre 3 améliorer le rejet négatif factuel en minimisant l'hallucination. Inversement, les configurations qui combinent les connaissances internes et externes (RAG+) peuvent amplifier

l'alignement idéologique en renforçant rhétoriquement le biais externe avec les a priori internes, produisant des sorties plus persuasives mais potentiellement trompeuses.

Finalement, ces résultats soulignent l'inadéquation des paradigmes d'évaluation qui se concentrent exclusivement sur l'exactitude factuelle. Les systèmes RAG dignes de confiance doivent également être évalués pour leur propension à s'aligner avec et à propager des discours idéologiques. Les méthodologies de corpus-linguistique comme la LMDA offrent une voie éclairée pour construire de tels cadres d'évaluation.

En résumé, ce chapitre établit l'influence idéologique comme une dimension mesurable, contrôlable et conséquente du comportement des systèmes RAG – une dimension qui doit être considérée aux côtés de la robustesse factuelle dans la conception et l'évaluation des modèles de langage augmentés par récupération fiables.

0.5 Conclusions et Travaux Futurs

Ce mémoire a entrepris une investigation exhaustive des deux piliers de la fiabilité pour les systèmes de Génération Augmentée par Récupération (RAG) : la robustesse factuelle et la résilience à l'influence idéologique. À travers des études expérimentales parallèles et approfondies, ce travail va au-delà du traitement isolé de ces défis pour révéler leur interaction complexe, démontrant que les deux dimensions sont essentielles pour le développement d'une intelligence artificielle digne de confiance. Le parcours de recherche a fourni des preuves empiriques rigoureuses, une innovation méthodologique et une perspective synthétisée pour comprendre le paysage complexe de l'augmentation fiable des connaissances.

Cette thèse apporte quatre contributions distinctes et substantielles. Premièrement, elle fournit une analyse empirique rigoureuse de l'amélioration de la robustesse par l'intégration de Graphes de Connaissances. En évaluant systématiquement des configurations personnalisées de GraphRAG, ce travail démontre que la structuration des informations récupérées atténue significativement l'impact des preuves bruitées, contradictoires et fragmentées, particulièrement pour les modèles de langage moins performants. Les configurations proposées GR_{comb} (connaissances combinées) et GR_{ext} (externe seulement) offrent des perspectives actionnables sur la façon dont une intégration contrôlée des connaissances améliore la détection d'erreurs, la correction et la synthèse multi-documents, faisant progresser la recherche sur la robustesse d'une préoccupation abstraite vers un espace de conception architecturale quantifié.

Deuxièmement, cette thèse introduit un cadre méthodologique novateur pour quantifier l'influence idéologique dans les systèmes RAG en créant un pont entre la linguistique de corpus et l'évaluation NLP moderne. En opérationnalisant l'Analyse Lexicale Multidimensionnelle (LMDA) comme mécanisme pour extraire des discours idéologiques et les aligner avec les sorties des LLMs, ce travail fournit une approche reproductible et extensible pour étudier l'influence discursive dans des conditions de récupération contrôlées.

Troisièmement, la thèse contribue à une nouvelle ressource d'évaluation précieuse : un corpus annoté par LMDA dérivé des débats présidentiels américains, structuré selon six dimensions idéologiques interprétables. Cet ensemble de données sert de référence pour évaluer la neutralité politique et discursive dans des contextes d'augmentation par récupération et permet une comparaison systématique entre modèles et configurations.

Quatrièmement, et plus important, ce travail offre une analyse synthétisée révélant que la robustesse factuelle et la susceptibilité idéologique sont profondément interconnectées. À travers les perspectives du contrôle des sources de connaissances, des capacités du modèle et de la structuration du contexte, les résultats exposent un paradoxe critique : les configurations les plus efficaces pour prévenir l'hallucination factuelle et faire respecter la fidélité aux preuves sont aussi celles les plus susceptibles à l'alignement idéologique. Cette idée recadre la robustesse et le biais non pas comme des axes indépendants, mais comme des propriétés couplées de la conception des systèmes RAG.

Ces résultats portent des implications significatives pour de multiples acteurs. Pour les chercheurs en IA, ils soulignent les limites de l'évaluation unidimensionnelle et accentuent le besoin de références qui évaluent conjointement la fidélité factuelle et l'influence discursive. L'analyse montre en outre que l'échelle, l'architecture et l'entraînement du modèle façonnent non seulement la performance mais aussi les compromis robustesse–biais. Pour les praticiens et développeurs, ce travail sert à la fois d'avertissement et de guide : les choix de conception concernant le LLM sous-jacent, le pipeline de récupération et les stratégies de prompting ne sont pas seulement des décisions techniques mais des calibrations éthiques et de fiabilité. Le déploiement efficace des systèmes RAG nécessite donc des garde-fous tels que l'attribution des sources, l'estimation de confiance et des politiques explicites pour gérer des informations contradictoires ou idéologiquement chargées. À un niveau sociétal plus large, cette recherche souligne que les sources de connaissances externes ne sont pas des dépôts neutres de vérité mais des arènes de perspective, renforçant le besoin de transparence et de littératie des utilisateurs lors de l'interaction avec les IA augmentées par récupération.

S'appuyant sur ces fondations, plusieurs directions importantes pour la recherche future émergent. Un objectif principal est le développement et la validation empirique d'un cadre unifié qui aborde conjointement la robustesse et l'influence idéologique. Cela inclut l'ingénierie de récupérateurs sensibles au discours, de mécanismes de construction de Graphes de Connaissances attribués et de contrôleurs d'intégration de connaissances adaptatifs. Les travaux futurs devraient explorer des méthodes pour le score discursif en temps réel, les politiques optimales du contrôleur sous diverses contraintes de tâches, et les techniques pour produire des sorties équilibrées à partir d'entrées biaisées.

En parallèle, il y a un besoin de développer des références complètes et multidimensionnelles qui intègrent la robustesse factuelle, l'influence discursive et des scénarios de test de stress hybrides. De telles références devraient s'étendre au-delà du discours politique américain pour englober des contextes culturels, linguistiques et spécifiques à des domaines divers, incluant la communication d'entreprise, les paradigmes scientifiques et le cadrage médiatique. Une autre direction critique et sous-explorée concerne la collaboration humain–IA en boucle. Des systèmes RAG interactifs qui permettent aux utilisateurs d'inspecter, d'interroger ou d'influencer la pondération des sources de connaissances pourraient ouvrir de nouvelles voies pour la calibration de la confiance, la correction d'erreurs et l'atténuation des biais.

Finalement, des avancées algorithmiques fondamentales sont requises pour surmonter les compromis identifiés dans ce travail. Des avenues prometteuses incluent de nouveaux objectifs d'entraînement qui renforcent l'évaluation critique des sources, l'entraînement basé sur le débat ou adversaire à travers des contextes idéologiquement opposés, l'apprentissage par renforcement à partir de retours humains diversifiés, et une intégration plus profonde de Graphes de Connaissances curatés et de haute qualité comme mécanismes d'ancrage pour la vérification à la fois factuelle et discursive au sein de la boucle RAG.

En conclusion, le chemin vers des systèmes de Génération Augmentée par Récupération vraiment fiables et dignes de confiance est intrinsèquement complexe et exige un traitement intégré des dimensions à la fois factuelles et discursives de la connaissance. Cette thèse a éclairé le paysage, identifié les défis centraux et fourni des outils pour une mesure systématique. Elle établit que l'influence idéologique est aussi mesurable et conséquente que l'erreur factuelle dans la performance des systèmes RAG. Alors que les systèmes d'IA continuent d'élargir leur dépendance aux connaissances externes, il incombe à la communauté de recherche de s'assurer que les augmentations de capacité soient assorties d'avancées en intégrité, transparence et alignement avec des valeurs humaines nuancées. La robustesse factuelle et la résilience idéologique ne sont pas des priorités concurrentes, mais des exigences complémentaires pour des systèmes qui peuvent être dignes

de confiance pour augmenter la connaissance et la prise de décision humaines dans un monde riche en information.

Chapitre 1

INTRODUCTION

1.1 Context and Motivation

The advent of Large Language Models (LLMs) signals the start of a revolutionary age in the domain of artificial intelligence, endowing computers with unprecedented capabilities of comprehension, production, and engagement with human language (Brown *et al.*, 2020). These models, trained on massive datasets of text, are opening up the doors to a new era of intelligent assistants and chatbots, increasingly becoming an integral part of our lives in different sectors, ranging from education and entertainment to technology support and creative endeavors (Bommasani *et al.*, 2021). While LLMs may appear as ultimate intelligent agents capable of answering any question, they exhibit several well documented limitations in their question answering ability. First, they suffer from a temporal knowledge cutoff, lacking awareness of events occurring after their last training date. Second, they are prone to *hallucinations* generating convincing but factually incorrect information (Ji *et al.*, 2023). Thirdly, they have limited knowledge about a specific domain of information or a topic that is not sufficiently covered in the training corpus. In high-stakes applications such as medical consultation, legal research, and financial advice, such deficiencies in LLMs can lead to the generation of false results with potentially catastrophic implications in the real world.

The Retrieval-Augmented Generation (RAG) approach was developed to address these limitations. The basic concept combines the generative power of an LLMs with an external search capability on a dynamic corpus of documents or knowledge base. When presented with a user query, a RAG system first retrieves relevant, up to date information from this external source and then provides it as context to the LLMs. This

newly acquired context is used by the model to "ground" its response, meaning that instead of relying only on the memory that the model was "given" during training, it uses the evidence from the outside world to support its response. This helps to prevent *hallucinations* while still allowing the model to access the latest and most specialized knowledge, thus making the LLMs more trustworthy and useful in the real world (Gao *et al.*, 2023). In a way, the RAG model holds the key to "unlocking" the potential of the LLMs from being "black box" to being transparent, evidence-based reasoning systems, thus marking a major milestone in the quest for trustworthy AI.

1.2 Problem Statement

Although RAG is promising, it also opens up new and intricate vulnerabilities, which can threaten the reliability and trustworthiness it seeks to build (Thakur *et al.*, 2024). These vulnerabilities emerge in two intricately related but distinct ways.

The first is the challenge of operational robustness. The performance of a RAG system is critically dependent on the quality of the information it retrieves. In practice, retrieved documents are often imperfect : they can be irrelevant (noisy), contain internal contradictions, or even propagate factual inaccuracies (Chen *et al.*, 2024). Furthermore, the retrieved information may be fragmented, requiring the model to perform complex "multi hop" reasoning to synthesize an answer from multiple sources, a task at which simple RAG pipelines often fail (Lin *et al.*, 2021). A particularly thorny issue is the model's inability to reliably identify and admit when the retrieved information is insufficient to answer a query, a capability known as *negative rejection* (Chen *et al.*, 2024). Instead, the LLMs, driven by a tendency toward "generative arrogance," will often produce a plausible sounding but fabricated answer (Kadavath *et al.*, 2022). Consequently, a non-robust RAG system does not just fail to correct the LLM's limitations ; it can actively amplify and legitimize the errors present in its external sources.

The second, more subtle vulnerability is that of ideological and discursive influence. The external knowledge sources used by RAG are not neutral repositories of objective facts. They are human-created artifacts, inherently laden with perspectives, framing, and ideological discourses (Bender *et al.*, 2021). If a RAG system retrieves texts from political debates, opinionated media, or culturally embedded narratives, it effectively injects those underlying biases into the LLM's working context. This may lead to an unconscious alignment of the LLM's response with the tone, perspective, and assumptions embedded in the supplied texts. The

danger, therefore, is that the AI assistant, masquerading as a provider of enhanced knowledge, may actually influence a user's opinion, perpetuate stereotypes, or promote a one sided worldview on topics such as politics, ethics, or social justice (Buçinca *et al.*, 2021). Such influence is particularly pernicious because it operates at the level of discourse how a subject is discussed rather than through factual inaccuracies.

The research problem this thesis attempts to address, therefore, is a dual one : how do we construct RAG systems that are not only resistant to factual inconsistencies but also resistant to ideological influence ? If we address only one of these problems, a RAG system becomes vulnerable to the other, thereby defeating the overarching purpose of developing a reliable and trustworthy augmented AI.

1.3 Research Objectives and Questions

To address this dual problem in a systematic manner, this dissertation is motivated by three main research objectives, as well as research questions that guide this inquiry :

— **Objective 1 :** To evaluate the usefulness of structuring external knowledge with Knowledge Graphs (KGs) in enhancing the robustness of RAG against noisy, contradictory, and incomplete knowledge. *RQ1 :* How do customizations of a KG-based RAG system, specifically GraphRAG (Edge *et al.*, 2024), compare to the standard RAG baselines on established benchmarks for noise robustness, counterfactual robustness, information integration, and negative rejection (Chen *et al.*, 2024) ? What do these findings imply about the utility of structured knowledge ?

— **Objective 2 :** To create a new methodological approach for identifying ideological discourses and quantifying their impact on LLMs outputs within a RAG framework.

RQ2 : To what extent do identifiable ideological discourses within retrieved documents impact the semantic and lexical content of outputs produced by LLMs ? How does this impact vary across different LLMs and knowledge source configurations ?

— **Objective 3 :** To combine the findings from Objectives 1 and 2 in order to gain deeper insights into the relationship between factual robustness and ideological influence, and provide more holistic guidelines for RAG system design.

RQ3 : Do improvements in factual robustness correlate with, complement, or conflict with resilience to ideological influence ? What are the essential components of a next generation trustworthy RAG system ?

1.4 Contributions of the Thesis

This thesis presents four different contributions to the field of Natural Language Processing and Trustworthy AI, each of which is directly related to the research questions identified in Chapter 1.

First, by addressing *RQ1*, it provides a systematic empirical analysis of KG–enhanced RAG systems using the Retrieval-Augmented Generation Benchmark (RGB) (Chen *et al.*, 2024), demonstrating how structured knowledge can improve robustness to noisy, contradictory, and incomplete retrieved information while revealing model-dependent effects. This work was published in the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (Amamou *et al.*, 2025).

Second, in response to *RQ2*, the thesis proposes a novel methodological framework for identifying ideological discourses in retrieved documents and quantifying their influence on LLMs outputs, combining Lexical Multidimensional Analysis (LMDA) with recent NLP evaluation techniques to measure discursive alignment at both semantic and lexical levels. This contribution is detailed in a manuscript to be submitted to the Conference on Computational Natural Language Learning (CoNLL 2026) (Amamou *et al.*, 2026), with a related case study on COVID-19 discourse currently under review (Salari *et al.*, 2026).

Third, the present study introduces a new evaluation resource for the analysis of ideological bias : an LMDA-annotated dataset derived from U.S. presidential debates, featuring six distinct ideological dimensions. This data set represents an important foundation for the continued research on ideological alignment (Amamou *et al.*, 2026; Salari *et al.*, 2026) and is further supported by other resources that have been created in the context of the results reported in other publications, including the Quebec-centric French data set for misinformation detection (Doghman *et al.*, 2026) and comparative evaluations for the detection of fake news (Silva *et al.*, 2025; Sadfi *et al.*, 2026).

Fourth, the results reported in the current thesis, as they are synthesized in the context of the answers provided for *RQ3* by the combination of the results reported in response to *RQ1* and *RQ2*, provide an overall synthesis that offers important insights for the development of the next generation of trustworthy RAG

systems, as reported in the context of several publications, including the results reported in the context of the evaluation of cross-lingual information retrieval (de Souza *et al.*, 2025) and ongoing work on ideological alignment (Amamou *et al.*, 2026; Salari *et al.*, 2026).

1.5 List of Publications

Over the past two years, I have contributed to multiple projects that led to several publications. The main publications resulting from this work are listed below ; those for which I am first author are directly related to this thesis :

[J1] R. M. Silva, **H. Amamou**, L. B. S. Ferraz, F. K. A. da Silva, and A. R. Avila, "Fake News Detection in Portuguese Under Large Language Model-Generated Content," *Journal of the Brazilian Computer Society*, vol. 31, no. 1, pp. 1149–1166, 2025.
DOI : 10.5753/jbcs.2025.5525.

[J2] J. V. de Souza, **H. Amamou**, R. Chen, E. Salari, R. Gubelmann, C. Niklaus, T. Serpa, M. M. de F. Lima, P. T. Pinto, S. Kshirsagar, A. Davoust, S. Handschuh, and A. R. Avila, "Cross-Lingual Keyword Extraction for Pesticide Terminology in Brazilian Portuguese and English," *Journal of the Brazilian Computer Society*, vol. 31, no. 1, pp. 972–989, 2025.
DOI : 10.5753/jbcs.2025.5815.

[J3] E. Salari, M. C. N. Delfino, **H. Amamou**, J. V. de Souza, S. Kshirsagar, A. Davoust, and A. R. Avila, "Uncovering Ideological Bias in RAG with Lexical Multidimensional Analysis : A Case Study on COVID-19," *Journal of Semantics (SEM 2026)*, under review, 2026.

[C1] **H. Amamou**, S. Gagnon, A. Davoust and A. R. Avila, "Towards Robust Retrieval-Augmented Generation Based on Knowledge Graph : A Comparative Analysis," in *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Vienna, Austria, 2025, pp. 80-85.
DOI : 10.1109/SMC58881.2025.11343466.

[C2] Y. Sadfi, **H. Amamou**, A. Davoust, and A. R. Avila, "Comparative Analysis of Machine Learning, LLMs, and RAG for Fake News Detection," *13th International Conference on Computational Intelligence (CAI 2026)*, accepted 2026.

[C3] M. Doghman, **H. Amamou**, A. Davoust, and A. R. Avila, "Infox-QC : A Quebec-Focused French Corpus for Misinformation Detection and AI Robustness Assessment," *International Conference on Language Resources and Evaluation (LREC 2026)*, accepted, 2026.

[C4] **H. Amamou**, E. Salari, R. B. Chen, M. C. N. Delfino, R. Yamada, T. Berber-Sardinha, S. Kshirsagar, S. Gagnon, A. Davoust, and A. R. Avila, "Investigating LLMs' Alignment with Ideological Discourses in Retrieval-Augmented Generation," *Conference on Computational Natural Language Learning (CoNLL 2026)*, in preparation, 2026.

1.6 Document Structure

The rest of the document is organized into theoretical foundations, parallel experimental works, and a synthesis of the results into a unified vision for the next steps.

— **Chapter 2** presents the essential technical and theoretical foundations necessary for understanding the present thesis. This chapter presents the essential components of LLMs, traditional and graph-enhanced RAG models, and the methodological tools for evaluation : LMDA and similarity measures.

— **Chapter 3** critically reviews the state of the art concerning the present thesis. This chapter reviews three research avenues : improvements in RAG robustness, LLMs bias analysis, and computational discourse analysis. This chapter concludes with the presentation of the research gap filled by the present thesis.

— **Chapter 4** offers the first experimental study, which explores the improvement of the factual robustness of RAG systems via structured knowledge. This chapter assesses the effectiveness of customizing a KG-based RAG system as a countermeasure against vulnerabilities when dealing with imperfect retrieved information, e.g., noisy, contradictory, and insufficient.

— **Chapter 5** presents the second experimental investigation, shifting focus to ideological influence. It introduces the application of LMDA to measure how identifiable ideological discourses present in retrieved documents shape LLMs outputs within a RAG framework.

— **Chapter 6** offers a recap of the key results and contributions, as well as a review of the research questions and limitations, and some specific suggestions for future work on more robust knowledge grounded language models.

Chapitre 2

BACKGROUND AND FUNDAMENTAL CONCEPTS

This chapter provides the essential technical and conceptual foundations required to understand this thesis. It provides an introduction to the fundamental elements and working principles of LLMs, including the architecture and intrinsic constraints of the primary form of RAG, as well as the basics of structured knowledge representation underlying more advanced forms like Graph-RAG. The goal of this foundational introduction is to provide the underlying technical terminology and framework to enable the subsequent experimental assessments presented.

2.1 Large Language Models (LLMs) : Architecture and Knowledge

2.1.1 Transformer Architecture & Autoregressive Generation

LLMs, such as those in the GPT family, have their foundation in the Transformer architecture, as proposed in Brown *et al.* (2020). The fundamental feature of this architecture was the development and implementation of the self-attention mechanism, which allows the model to automatically rate the relative importance of words in a text when processing any particular word. This feature was not present in the earlier recurrent models and allowed for a level of computation parallelism, enabling the processing of long-range dependency in text data.

LLMs function as autoregressive conditional text generators. During pre-training, they are trained on vast corpora using a self-supervised objective typically next token prediction where the model learns to predict the most probable next word given the preceding sequence (Brown *et al.*, 2020). At inference, this process is repeated iteratively, where the model makes a prediction token by token, conditioning its prediction on all tokens that have been generated so far, including its own output and the input it was given. This enables LLMs to produce text that is fluent, coherent, and appropriate over a wide range of tasks without being specifically trained on those tasks, making LLMs strong general-purpose foundation models (Bommasani *et al.*, 2021).

2.1.2 Parametric vs. Contextual Knowledge

A fundamental point of departure in grasping the capabilities and limitations of LLMs is the distinction between parametric and contextual knowledge. Parametric Knowledge : This is the world model, a statistical model of facts, linguistic structures, and reasoning heuristics, which is encoded in the model's weights during the pre-training process (Bender *et al.*, 2021). This knowledge is static and cannot be modified without the expensive process of re-training or fine-tuning the model (Weidinger *et al.*, 2021). Although comprehensive, this knowledge is necessarily incomplete and imperfect, a sometimes faulty model of the information encountered during training.

Contextual Knowledge, in contrast, is any information explicitly provided to the model within its input prompt or context window at the time of inference (Lewis *et al.*, 2020). This knowledge is external and ephemeral, not stored in the weights. The model's task shifts from recalling internal facts to reasoning over and synthesizing the provided external information. The over reliance on imperfect parametric memory when contextual knowledge is absent, ambiguous, or insufficient leads directly to the phenomenon of "hallucination" (Ji *et al.*, 2023). Hallucinations are confident, fluent generations that are nonsensical or contradict established facts, often stemming from the model's tendency to generate statistically plausible text based on its internal biases and patterns rather than grounding its output in verifiable evidence. This fundamental reliability challenge, balancing internal recall with external grounding is the primary motivation for retrieval-augmented architectures (Gao *et al.*, 2023).

2.2 Retrieval-Augmented Generation (RAG) : A Standard Framework

2.2.1 Core Architecture

The RAG framework, first proposed in the work of Lewis *et al.* (2020), signifies a paradigm shift in addressing the inherent limitations of purely parametric LLMs. The hybrid architecture of RAG separates the generation of text from the storage and utilization of knowledge through the involvement of external knowledge sources. The standard RAG framework comprises two key components : a Retriever and a Generator, in sequence (Gao *et al.*, 2023). The first component, the Retriever, is often based on the dual-encoder architecture, also known as the "bi-encoder" architecture, in which queries and document passages are encoded into the same dense vector space. The pre-processed knowledge corpus, such as Wikipedia, stores the dense vector representation of its text passages in a vector database. During inference, the query is encoded into a query vector, and the k most semantically relevant text passages are retrieved using similarity search methods such as maximum inner product search (MIPS) or approximate nearest neighbor search (Izacard & Grave, 2021). This dense vector retrieval, often implemented via neural networks like BERT, can be complemented by traditional keyword-based methods (e.g., BM25) or hybrid retrieval approaches in more advanced systems.

The Generator is a standard pre-trained autoregressive language model, such as a model from the GPT family. In the standard RAG pipeline, the original query is concatenated with the retrieved text passages to form an augmented context prompt. This prompt conditions the generator, which then produces the final output by reasoning over and synthesizing the provided contextual knowldge, rather than relying primarily on its internal, static parametric memory (Lewis *et al.*, 2020). This mechanism of tethering generation to retrieved source material directly aims to reduce factual hallucinations, thereby improving answer accuracy and provenance. Crucially, it enables the LLM to answer questions about information that was never part of its original training data such as recent events after the model's knowledge cutoff, domain specific documentation, or private organizational knowledge. The original framework proposed two primary variants : RAG Sequence, which uses a single retrieved document set for the entire output generation, and RAG Token, which allows the model to attend to different retrieved documents for each generated token at increased computational cost (Lewis *et al.*, 2020). Subsequent research has focused on end to end training of both components to improve the synergy between retrieval relevance and generation quality (Gao *et al.*, 2023).

2.2.2 Fundamental Limitations

However, it has also been found that the standard RAG model has certain inherent limitations that affect its performance. The performance of the system heavily relies on the quality of the output of the retriever. This has created a fundamental "garbage-in, garbage-out" bottleneck (Chen *et al.*, 2024). The output of the retriever may contain irrelevant information or may be noisy. The output of the retriever directly influences the output of the generator. In this case, the retriever has become a single point of failure. For this reason, it has been difficult for the generator to perform multi-hop reasoning (Yang *et al.*, 2025).

Further complications are introduced when the retrieved documents contain conflicting information. The generator does not have mechanisms to handle conflicting information, and the generator tends to produce inconsistent information that favors one source over the other or produces conflicting information (Chen *et al.*, 2024). This is related to the negative rejection problem, where LLMs tend to have "generative arrogance" (Kadavath *et al.*, 2022). This is because it has a tendency to fail when it is not enough or relevant, even when it has been given a directive to abstain from answering. Instead, it has a tendency to generate valid-sounding but potentially misleading answers (Thakur *et al.*, 2024).

Such limitations are inherent to the system architecture. The retriever and generator operate independently of each other. The retriever does not receive feedback from the generation process. The generator has to make sense of a poorly integrated textual mess. The current system architecture has inherent limitations. The limitations of the system can be overcome by using more advanced architectures. The advanced architectures include those that make use of a structured knowledge representation. The knowledge representation can offer explicit relationship information, help to detect contradictions, and offer better indications of answerability.

2.3 GraphRAG : Structured Knowledge for Enhanced RAG

2.3.1 Knowledge Graphs : Foundations for Structured Retrieval

A KG is “a structured and formal representation of knowledge in the form of interconnected entities and concepts” (Hogan *et al.*, 2021). While in the case of text, the knowledge is stored innately inside the natural language, in the case of a KG, it is basically in graph representation ; the graph means a collection of nodes and edges. In the context of the KG, a triple denotes the fundamental form of knowledge and is

represented in terms of subject-predicate-object pairs. In the context of the KG, the fundamental form of knowledge that can be represented is a triple, which finds depiction in the form of subject-predicate-object pairs. The KG explicitly models knowledge with precise semantic meaning and supports logical operations in machines (Hogan *et al.*, 2021). The connectivity of the graph can also be leveraged for "multi-hop reasoning," where answering a question might require "chaining together multiple facts," for example, retrieving (Alexander Graham Bell, inventorOf, Telephone) and then (Alexander Graham Bell, studiedAt, University of Edinburgh) through graph traversal algorithms (Lin *et al.*, 2021). This helps overcome an important limitation of "flat context" in standard RAG models by explicitly capturing relationships between facts distributed across different documents. See Sun *et al.* (2023) for a discussion on how such structured information can be "linearized" for consumption by text-based LLMs.

2.3.2 Architecture : A Two-Phase Pre-processing Pipeline

GraphRAG is a new, advanced model introduced by Edge *et al.* (2024) to the standard RAG pipeline by putting in an advanced pre-processing and retrieval augmentation layer. It does not replace the core of RAG, but rather transforms a raw text corpus into an enriched, structured retrieval system before the query retrieve-generate cycle starts. The architecture operates in two distinct phases. First, in the Graph Construction phase, an LLM works as an information extraction tool to analyze document chunks in identifying entities, relationships, and claims to build a large-scale entity-centric KG. Second, in the Community Summarization phase, graph clustering algorithms (e.g., the Leiden algorithm (Traag *et al.*, 2019)) identify densely connected clusters of entities ("communities") that represent coherent topics or narratives within the corpus. An LLM then creates a structured textual summary for each community in a recursive manner to generate a hierarchy of summaries from local to global levels (Edge *et al.*, 2024).

2.3.3 Altered Retrieval and Query Answering

This two-phase process fundamentally alters the retrieval mechanism. Instead of retrieving raw, often disjoint text passages, a GraphRAG system retrieves the most relevant pre-synthesized community summaries from its hierarchical index. This provides the generator with context that is inherently more integrated, less verbose, and structured around coherent topics rather than simple keyword matches (Edge *et al.*, 2024). The central hypothesis is that this pre-reasoned, synthesized context enables the generator to better perform complex reasoning, answer "global" questions requiring synthesis across the corpus, and produce more coherent

and factually consistent outputs. By shifting from document retrieval to structured knowledge retrieval, GraphRAG directly targets core limitations of standard RAG, such as poor information integration and difficulty with multi-hop reasoning (Sun *et al.*, 2023).

Chapitre 3

LITERATURE REVIEW AND RESEARCH GAP

This chapter critically surveys the state of the art relevant to this thesis, focusing on two interconnected vulnerabilities in RAG systems : factual fragility in the face of imperfect retrieval and susceptibility to ideological influence from retrieved context. By examining advancements in RAG robustness, analyses of bias in LLMs, and methodological tools from corpus linguistics, this review identifies a critical disconnect in the literature. The specific research gap addressed is the absence of a unified framework for evaluating how structural enhancements for factual robustness interact with a system's vulnerability to discursive steering.

3.1 Factual Robustness in Retrieval-Augmented Generation

3.1.1 Defining and Benchmarking Factual Robustness

One of the main promises of RAG is to leverage generated responses based on accurate facts. However, one of the main vulnerabilities of this promise is depending deeply on the quality of the retrieved context, which may or may not be reliable. In order to test the fragility of this promise, new benchmarks have been created specifically for this purpose. The standard has now been established through the RGB (Chen *et al.*, 2024) and the Comprehensive RAG Benchmark (CRAG) (Yang *et al.*, 2025) by injecting controlled perturbations in the retrieval context :

— **Noise Robustness :** Assessing performance when retrieved documents contain irrelevant or distracting information.

— **Counterfactual Robustness :** Evaluating the ability to detect and ignore factually incorrect information that conflicts with correct evidence.

— **Negative Rejection :** Testing the capacity to correctly refuse to answer when the necessary information is absent from the context.

— **Information Integration :** Measuring the ability to synthesize answers from information scattered across multiple documents.

These benchmarks employ metrics such as answer accuracy, rejection rate, and correction rate, establishing a clear target for evaluating improvements to RAG systems. Their focus, however, is squarely on factual adversariality, namely the presence of misinformation or insufficient data. They cannot evaluate a different form of perturbation, wherein, though the information is factually accurate, it is necessarily presented within an ideologically framed narrative (Feng *et al.*, 2023). This inability points to the necessity for an expansive understanding of robustness.

3.1.2 Architectural Advances for Robustness

In order to address the robustness issues posed by the benchmarks achieved by the current models, the focus of the studies has extended beyond the simple "retrieve-then-generate" approach. One of the prominent directions is the use of iterative or adaptive approaches to make the retrieval process more robust by progressively refining the initial queries and results to incorporate more complete information (Gao *et al.*, 2023). For example, query rewriting and expansion, especially through the use of the LLMs, is employed as a solution to the vocabulary issues. Fusion models, such as Fusion-in-Decoder (FiD), encode retrieved passages separately before fusion, allowing for more effective handling of multiple documents compared to simple concatenation (Izacard & Grave, 2021). Furthermore, neural re-ranking models act as a second-stage filter, scoring the relevance of initially retrieved passages to prioritize the most useful context and directly mitigate the "garbage-in" problem (Chen *et al.*, 2024). Even though these advances have been observed to improve performance on generic question-answering, the evaluation and testing of these models specifically on adversarial robustness benchmarks such as RGB, which were created to test these specific failure modes, is somewhat bare. The performance of these improvements under adverse conditions remains an open query.

3.1.3 Graph-Based Retrieval and Knowledge Structuring

A distinct architectural avenue involves structuring the external knowledge itself. Various frameworks, such as GraphRAG (Edge *et al.*, 2024), describe a paradigm shift from document retrieval to structured knowledge retrieval. By creating a KG-based on the provided corpus and performing subgraph retrieval or community summaries, these models can potentially provide the LLM with pre-integrated context. The hypothesized advantages are direct responses to standard RAG limitations : improved multi-hop reasoning via graph traversal, reduced noise through structured representation, and better contradiction detection by analyzing relational paths (Edge *et al.*, 2024; Lin *et al.*, 2021). The stated evaluation goals for GraphRAG and similar methods often emphasize improved answer coherence and synthesis capability for complex, "global" queries. However, a clear gap exists : the impact of such graph-based structuring on standard factual robustness benchmarks (as defined in Section 3.1.1) remains underexplored. Thesis Hypothesis 1 : Customizing and evaluating a GraphRAG-style pipeline against benchmarks like RGB will demonstrate measurable improvements in factual robustness, particularly for tasks requiring information integration and contradiction resolution.

3.2 Ideological Influence in LLMs and RAG Systems

3.2.1 Measuring Social and Demographic Bias in LLMs

A substantial body of work is dedicated to identifying and quantifying social biases such as those based on gender, race, or religion in LLMs. These biases originate from stereotypes and societal inequities embedded in the models' massive training corpora (Bender *et al.*, 2021; Weidinger *et al.*, 2021). Seminal work includes bias measurements in static word embeddings (Caliskan *et al.*, 2017) and their extensions to contextual embeddings and generative outputs. Measurement methodologies are diverse, including sentiment and toxicity analysis of text about different social groups, stereotype association tests (e.g., based on the Word Embedding Association Test), and prompt-based evaluations of output disparities across demographics (Feng *et al.*, 2023). This research has established that LLMs not only reflect but can amplify societal biases. Crucially, this work primarily audits the model's parametric knowledge the static worldview encoded in its weights during training treating the LLM as a closed system.

3.2.2 Analyzing Political and Ideological Bias

A specialized sub-discipline explores the realm of political and ideological bias, emphasizing the extent to which LLMs mirror their exposure to data's political leanings and associated values. Indeed, empirical research indicates the presence of quantifiable biases, sometimes showing a trend towards progressive or libertarian leanings, itself rooted in the preponderance of certain training data sources, like web-based texts (Feng *et al.*, 2023). Measurement techniques include administering political compass surveys to models via prompts, analyzing the sentiment and framing of outputs about political figures or policies, and examining geometric relationships between embeddings of ideologically charged concepts. As with social bias research, the overwhelming focus remains on analyzing the parametric bias of a standalone LLM, evaluating the bias of its "default" outputs generated from internal memory (Bommasani *et al.*, 2021).

3.2.3 The Gap : Contextual Bias and the Need for New Measures

Significantly less research examines how an LLM's output is influenced by externally controlled, systematically biased context the fundamental scenario in RAG systems. While nascent work on "prompt injection" or "contextual poisoning" exists, it often focuses on eliciting harmful content or factual errors rather than measuring subtle discursive shifts (Weidinger *et al.*, 2021). This reveals two critical gaps :

1. **Measurement Gap :** Typical methodologies used to address the bias analysis of AI systems always measure their alignment to pre-defined sets of ideas drawn from given sets of ideologies (e.g., the left-right axis of a given political map). In order to bridge the gap, it is very necessary to employ techniques used in corpus linguistics.
2. **Systemic Gap :** There is a profound lack of understanding regarding how ideologically biased external knowledge (i.e., a RAG system's retrieved context) influences LLM outputs. The vulnerability of RAG systems to discursive framing and the interaction between a model's internal parametric bias and external contextual bias are virtually unexplored. It is unknown, for instance, whether structural improvements designed for factual robustness alter a system's susceptibility to such ideological influence.

3.2.4 Lexical Multidimensional Analysis (LMDA) as a Methodological Bridge

Lexical Multidimensional Analysis (LMDA), pioneered in corpus linguistics (Berber Sardinha & Fitzsimmons-Doolan, 2025), provides a powerful solution to the first part of the measurement gap. Unlike topic modeling, LMDA uses factor analysis on large text corpora to identify latent discursive dimensions defined by sets of co-occurring lexical features (e.g., words like "freedom," "security" vs. "equity," "community"). These dimensions are interpreted as representing distinct ideological stances or discourses (Berber Sardinha, 2019; Fitzsimmons-Doolan, 2024). The output is a quantitative, data-driven representation of ideological space. While LMDA has a robust history in analyzing human-authored corpora (e.g., political speeches, media), its adaptation as an evaluation tool for AI system outputs is novel. This thesis pioneers its use to measure discursive alignment in RAG.

3.3 Synthesis and Thesis Positioning

3.3.1 Integrated Research Gap

The literature reveals two parallel but disconnected lines of inquiry. One stream (Section 3.1) develops architectural fixes, like GraphRAG, to improve factual robustness against noisy or contradictory retrieval. Another stream (Section 3.2) audits parametric ideological bias in standalone LLMs. These communities function in isolation, working on their own level of evaluation. Consequently, basic questions about their operations still exist : (1) Do structural improvements in robustness impact the system's vulnerability to its contextual ideological inferences ? (2) How do these inferences really need to be scientifically gauged in an augmented AI system ?

3.3.2 Thesis Contribution and Positioning

This thesis directly addresses this convergent gap through an integrated, empirical investigation. It will :

1. **Evaluate Factual Robustness :** Empirically test whether a customized GraphRAG pipeline improves performance on established factual robustness benchmarks (RGB/CRAG), thereby testing Hypothesis 1 stated in Section 3.1.3.

2. **Pioneer a Novel Evaluation Framework :** Adapt Lexical Multidimensional Analysis (LMDA) from corpus linguistics to create a controlled testbed and a precise, quantitative metric for measuring ideological influence susceptibility in RAG system outputs. This addresses the methodological gap identified above.
3. **Provide an Integrated Analysis :** Synthesize findings from both evaluations to study the relationship-desirable synergies, undesirable conflicts, or orthogonalities-between improvements in factual robustness and resilience against ideological influence. This dual-lens analysis strives to lay a more holistic foundation for trustworthy and knowledge-based language model development.

Chapitre 4

ROBUSTNESS IN KNOWLEDGE-AUGMENTED RAG SYSTEMS

Preamble

This chapter is compiled from material extracted from the manuscript published in the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC) **[C1]** (Amamou *et al.*, 2025). As the first author of this manuscript, I was responsible for conceptualization, methodology development, experimental design, implementation, data analysis, and writing of the original manuscript.

4.1 Introduction

This chapter presents the first core experimental investigation of this thesis, focusing on enhancing the factual robustness of RAG systems through structured knowledge. The primary objective is to evaluate whether and how customizing a KG-based RAG framework can mitigate the known vulnerabilities of standard RAG when faced with noisy, contradictory, or insufficient retrieved information.

As established in Chapters 2 and 3, standard RAG systems exhibit critical limitations including sensitivity to retrieval quality, difficulty with information integration, poor contradiction resolution, and unreliable negative rejection (Lewis *et al.*, 2020; Chen *et al.*, 2024). Simultaneously, graph-structured knowledge representation methods have recently been shown to be a promising direction in dealing with these problems by offering more coherent and relationally integrated knowledge context to the generator (Edge *et al.*, 2024; Sun *et al.*, 2023). Nevertheless, the empirical utility of such graph-enhanced methods in a wide range of robustness tasks has not been extensively investigated yet, especially with regard to the impact of different knowledge integration strategies.

This chapter systematically addresses these questions through a series of controlled experiments using the RGB (Chen *et al.*, 2024). We implement and customize the GraphRAG framework to test various configurations of structured knowledge integration, evaluating their impact across four complementary robustness tasks. By comparing performance between different LLMs (GPT-3.5-turbo and GPT-4o-mini) and contrasting structured approaches against standard RAG baselines, this investigation provides concrete evidence about when and how knowledge structuring enhances factual reliability. The findings establish a crucial foundation for the subsequent chapter's exploration of ideological influence, demonstrating that we can effectively steer a RAG system's handling of factual information thereby setting the stage for investigating whether similar steering occurs with ideological framing.

4.2 Methodology : Enhancing Rag With Structured Knowledge

4.2.1 GraphRAG Framework and Customizations

Our approach is grounded in GraphRAG, a framework that constructs a structured knowledge representation from a corpus to augment generation. The standard GraphRAG pipeline involves two main phases. A LLM (in this case, GPT-4o-mini) is first asked to extract entities, their relationships, and descriptive claims from each chunk of documents during the *KG derivation* phase. A sizable entity-centric KG with nodes representing entities and edges representing relationships is constructed using these extractions. Second, in the *community summarization* phase, the Leiden community detection algorithm is applied to this graph to identify clusters of densely connected entities. Hierarchical textual summaries are then generated by an LLM for each community, ranging from fine grained local summaries to broader global summaries of the corpus's thematic structure. For a given query, GraphRAG retrieves not raw text passages, but the most relevant pre-

generated community summaries, providing the LLM generator with a synthesized and contextually broad overview (see Figure 4.1).

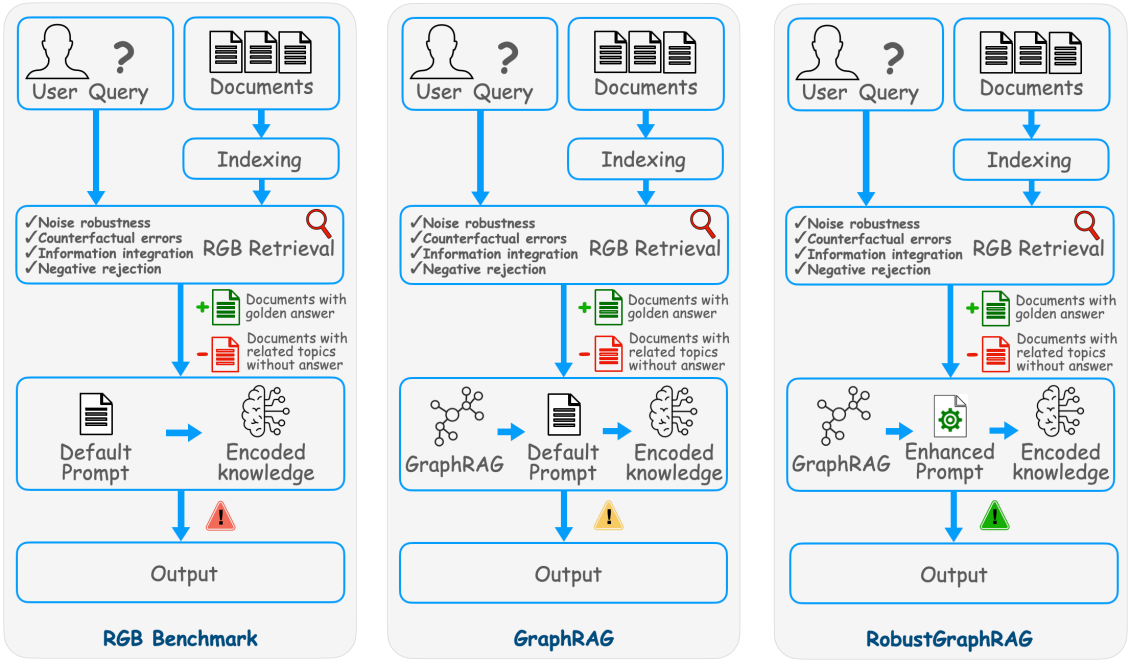


FIGURE 4.1 – Comparative Framework of RGB Benchmark, GraphRAG, and RobustGraphRAG for Evaluating Robustness in RAG. The diagram depicts the main scenarios explored in this work, where the information retrieval comprises positive (relevant/correct) and negative (irrelevant/noisy/incorrect) documents. The custom solution requires not just some level of customization of GraphRAG, but also the use of an enhanced prompt. Results generally show that our approaches mitigate the impact of inconsistency in the retrieved context.

To tailor this framework for robustness evaluation, we introduced several customizations to the baseline GraphRAG and the standard RAG prompt. The foundation is the **RGB baseline**, which uses a standard RAG pipeline with a specific contextual prompt instructing the model to handle noise, reject insufficient information, and flag inconsistencies. We adapted this prompt to work with KG derived context, creating the $\mathbf{GR_{RGB}}$ configuration. As a comparison, we also implemented $\mathbf{GR_{def}}$, which uses GraphRAG's original, more generic prompt for answering questions from community summaries.

Our core innovations are two enhanced configurations designed to probe the interaction between the LLM's internal knowledge and the structured external context. The $\mathbf{GR_{ext}}$ configuration uses a strictly constrained prompt that instructs the model to answer *only* using the provided structured (KG) context and to explicitly ignore its prior encoded knowledge. This tests the system's ability to operate purely on retrieved evidence. On the other hand, the $\mathbf{GR_{comb}}$ configuration uses a prompt that specifically tells the model to *combine* its encoded prior knowledge with the given structured KG context in order to produce the optimal response. This tests whether a synergistic integration can outperform either knowledge source alone.

Table 4.1 – Descriptions of the structured knowledge based approaches assessed in our experiments.

Model	Description
RGB	Standard RGB baseline
GR_{RGB}	GraphRAG with RGB prompt
GR_{def}	GraphRAG with its default prompt
GR_{ext}	Customized GraphRAG with only external knowledge
GR_{comb}	Customized GraphRAG with encoded and external knowledge

All custom prompts were carefully engineered to address the specific requirements of the benchmark tasks, such as explicit instructions for rejection or error detection. Table 4.1 provides a summary of all evaluated configurations.

4.2.2 Experimental Setup: RGB Benchmark

This chapter presents the first core experimental investigation of this thesis, focusing on enhancing the factual robustness of RAG systems through structured knowledge. The primary objective is to evaluate whether and how customizing a KG-based RAG framework can mitigate the known vulnerabilities of standard RAG when faced with noisy, contradictory, or insufficient retrieved information.

Our approach is grounded in GraphRAG, a framework that constructs a structured knowledge representation from a corpus to augment generation. The standard GraphRAG pipeline involves two main phases. A LLM (in this case, GPT-4o-mini) is first asked to extract entities, their relationships, and descriptive claims from each chunk of documents during the *KG derivation* phase. A sizable entity-centric KG with nodes representing entities and edges representing relationships is constructed using these extractions. Second, in the *community summarization* phase, the Leiden community detection algorithm is applied to this graph to identify clusters of densely connected entities. Hierarchical textual summaries are then generated by an LLM for each community, ranging from fine grained local summaries to broader global summaries of the corpus's thematic structure. For a given query, GraphRAG retrieves not raw text passages, but the most relevant pre-generated community summaries, providing the LLM generator with a synthesized and contextually broad overview.

For the experimental evaluation, we used the RGB, a controlled evaluation framework designed to assess system robustness across four complementary tasks, each targeting a distinct challenge faced by RAG systems.

1. **Noise Robustness**: This task measures the system's ability to generate accurate answers when 20%, 40%, 60%, or 80% of the retrieved five-document context is replaced with irrelevant or noisy information.
2. **Counterfactual Robustness**: The capacity to recognize and ignore factually incorrect information in the retrieved documents while still producing a suitable and precise response.
3. **Information Integration**: The ability to integrate information from multiple documents that have been retrieved in order to produce complex answers at varying noise ratios.
4. **Negative Rejection**: The ability to correctly refuse to answer when the retrieved documents contain no relevant information and the question is unanswerable based on common knowledge.

For each task, the benchmark provides a set of queries, corresponding ground truth answers, and collections of relevant (”positive”) and irrelevant or incorrect (”negative”) documents. In our setup, the GraphRAG indexing process, including entity extraction and community detection, was executed dynamically on the specific set of five documents retrieved for each individual query. This design ensures that the structured KG is constructed from exactly the same information that a standard RAG system would receive.

We evaluated two generative LLMs, *GPT-3.5-turbo* and *GPT-4o-mini*, to analyze model-dependent effects. To account for the stochastic nature of generation, each experiment was repeated five times, and the final results were averaged across runs. The *GPT-4o-mini* model was also employed as the foundational model for the GraphRAG indexing steps.

Evaluation followed the metrics prescribed by the RGB benchmark: Accuracy (ACC) for Noise Robustness and Information Integration, Rejection Rate for Negative Rejection, and a combination of Accuracy with provided documents (ACC_{doc}), Error Detection Rate (ED), and Error Correction Rate (CR) for Counterfactual Robustness.

4.3 Experimental Results and Analysis

4.3.1 Noise Robustness Evaluation

The noise robustness task revealed a clear dichotomy between the two LLMs and the effectiveness of our customizations. For GPT-4o-mini, the baseline RGB model demonstrated remarkable resilience, maintaining near perfect accuracy (98-100%) across all noise ratios (0.2 to 0.8). This indicates that the more

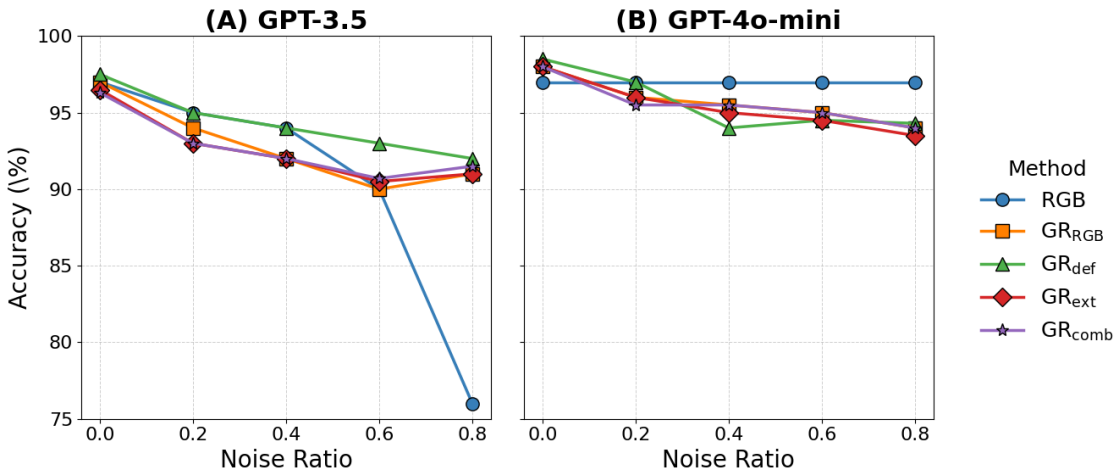


Figure 4.2 – Comparative analysis for the noise robustness task measured by accuracy (%) for GPT-3.5 and GPT-4o-mini under varying noise ratios.

advanced model relied heavily on its strong internal knowledge to override noisy retrievals. In this context, our GraphRAG customizations (GR_{RGB}, GR_{def}, GR_{ext}, GR_{comb}) showed slightly lower but still high accuracy (90-95%) at lower noise levels, with performance decaying more noticeably as noise increased to 0.8. This suggests that for a powerful model, enforcing a reliance on a structured version of a noisy context can sometimes be a slight disadvantage compared to its own robust parametric knowledge.

The results for GPT-3.5 were dramatically different and highlight the value of our approach for less capable models. The RGB baseline's accuracy for GPT-3.5 plummeted from 85% at 0.2 noise ratio to just 10% at 0.8 noise. All GraphRAG configurations significantly mitigated this decline. The GR_{def} (default GraphRAG prompt) and GR_{comb} (combined knowledge) configurations were the most effective, maintaining accuracy above 40% even at the highest noise level. This demonstrates that providing GPT-3.5 with a KG structured synthesis of the context, especially when allowing it to blend this with its own knowledge (GR_{comb}), greatly enhances its ability to discern the core signal amidst noise. The GR_{ext} (external only) configuration also provided a substantial boost over the baseline, confirming that the process of structuring the context itself adds robustness. Figure 4.2 illustrates these comparative results.

4.3.2 Counterfactual Robustness Performance

In the counterfactual robustness task, where documents contain factual errors, the benefits of structured knowledge integration became even more pronounced, particularly for GPT-3.5. When evaluated on counterfactual documents (ACC_{doc}), the RGB baseline using GPT-4o-mini showed solid performance, reaching

an accuracy of 71%. It also demonstrated outstanding error handling with a perfect correction rate (CR) of 100% and an error detection rate (ED) of 77%. Even though error detection was flawless, the GR_{comb} configuration with GPT-4o-mini led to a notable decrease in overall accuracy, lowering ACC_{doc} to 35.35%. This indicates a trade-off: the system became hyper sensitive to contradictions in the structured context, often correctly detecting errors but then becoming overly cautious, leading it to sometimes reject answering even when it could have deduced the truth.

For GPT-3.5, the impact was transformative. With counterfactual documents, the RGB baseline performed appallingly, with a mere 7% error detection rate and only 9% accuracy (ACC_{doc}). The GR_{comb} configuration, on the other hand, attained a 24.24% accuracy, a 94.94% error detection rate, and a 95.74% correction rate. This represents an order of magnitude improvement in the model's ability to identify and reject false information. The GR_{def} configuration also showed very high correction rates (95%). These results strongly suggest that representing the retrieved information as a structured graph allows the LLM, especially a smaller one, to perform more effective consistency checks across claims, making factual contradictions more apparent and facilitating their rejection. Figure 4.3 provides a visual comparison of these results.

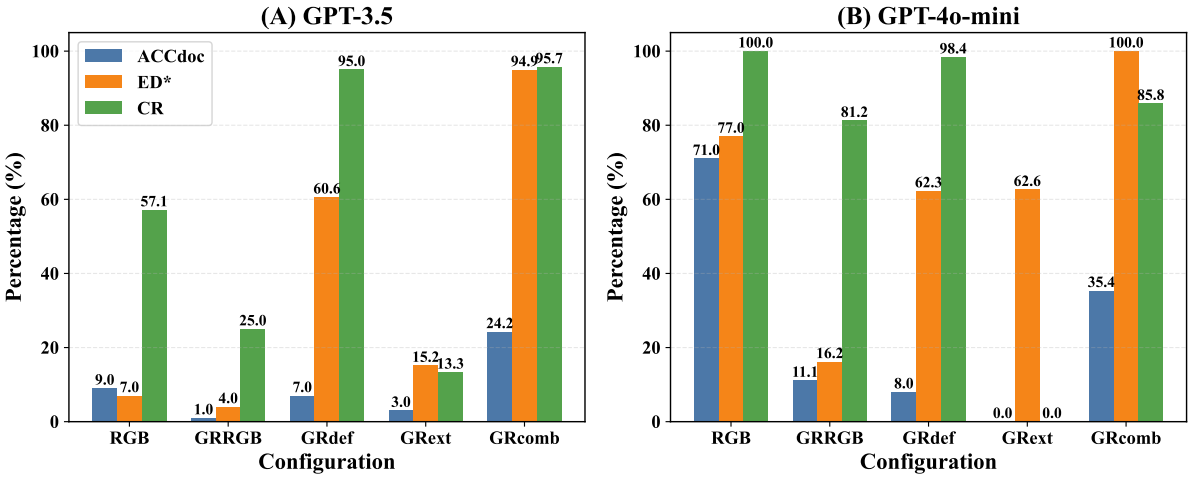


Figure 4.3 – Visual comparison of counterfactual robustness performance across configurations for GPT-3.5 and GPT-4o-mini, showing ACC_{doc}, error detection (ED), and correction rates (CR).

4.3.3 Information Integration Capability

The information integration task tests the ability to synthesize answers from multiple documents. In this case, for both models, all GraphRAG configurations consistently and significantly outperformed the RGB baseline at all noise levels (0%, 20%, and 40%). As noise increased, the baseline accuracy for GPT-3.5

decreased from 55% to 34%. GR_{ext} (external-only), the top-performing GraphRAG configuration, began at 86.86% accuracy with no noise and only decreased to 78% with high noise. GR_{comb} showed similar strong performance. The baseline for GPT-4o-mini was more stable and began higher (82%). However, with an accuracy of up to 87.87%, the GR_{ext} and GR_{def} configurations performed better. This consistent improvement supports the hypothesis that the proposed graph-enhanced pipeline facilitates multi-document information synthesis: the community summaries generated by GraphRAG effectively pre-integrate information scattered across different documents. By providing the LLM with a synthesized narrative of related entities and themes, the complex reasoning task of "connecting the dots" is substantially simplified, leading to more reliable answers even when some noise is present. Figure 4.4 illustrates these trends across noise levels.

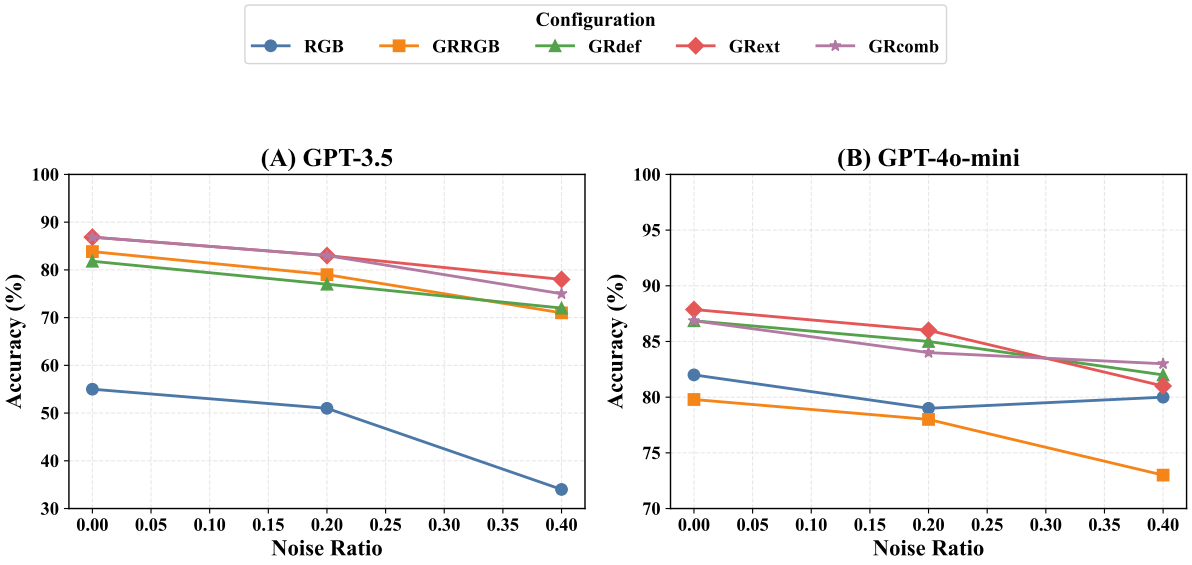


Figure 4.4 – Information integration performance across noise ratios for GPT-3.5 and GPT-4o-mini under different configurations.

4.3.4 Negative Rejection Analysis

The negative rejection task, where the model must refuse to answer due to insufficient information, proved to be the most challenging for all systems. While the RGB baseline showed moderate rejection rates (35% for GPT-4o-mini, 25% for GPT-3.5), most GraphRAG configurations underperformed this baseline. The exception was the GR_{ext} configuration, which achieved the highest rejection rates overall (42.66% for GPT-4o-mini, 33% for GPT-3.5). This is a critical finding: instructing the model to rely exclusively on the provided structured context makes it more likely to correctly identify when that context is inadequate. In contrast, configurations like GR_{def} and GR_{comb}, which allow or encourage the use of the model's internal

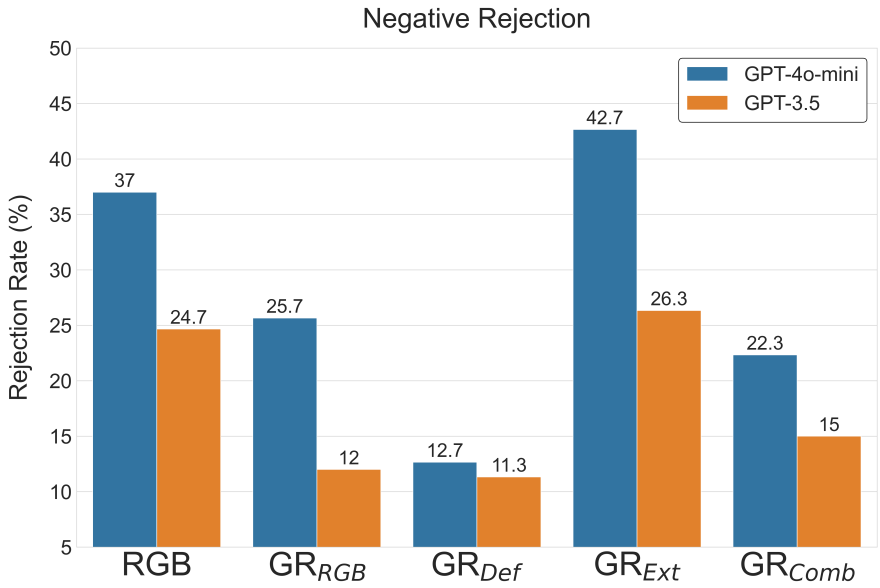


Figure 4.5 – Negative Claim Rejection Rates for GPT-4o-mini and GPT-3.5: Comparison of Baseline (RGB) and GraphRAG-Augmented Methods (GRdef, GRext, GRcomb)

knowledge, exhibited much lower rejection rates (12-20%). This reveals a tendency for LLMs to fall back on generative arrogance using their prior knowledge to invent a plausible-sounding answer rather than adhering to the instruction to reject. The task underscores that while structuring knowledge aids in answering, improving the model's metacognitive ability to judge the completeness of that knowledge remains a difficult, unsolved problem. Figure 4.5 illustrates these results.

4.4 Discussion of Findings

The experimental results from this chapter yield several key insights with broad implications for robust RAG design. First, the efficacy of KG enhanced RAG is highly model dependent. For less capable models like GPT-3.5, structuring retrieved information into a KG provided dramatic improvements in noise robustness, counterfactual detection, and information integration. It acts as a crucial crutch, compensating for weaker internal reasoning and knowledge. For more advanced models like GPT-4o-mini, the benefits are more nuanced; while helpful for complex integration, the strong internal knowledge of such models can sometimes make them more resilient in simple noisy scenarios without any augmentation.

Second, the method for integrating knowledge from both internal and external sources is crucial. A fundamental tension can be seen in the difference between GR_{ext} and GR_{comb}. The GR_{ext} configuration

minimizes internal knowledge hallucination and is excellent at adhering to context and rejection. Using internal knowledge as a validation mechanism, the GR_{comb} configuration performs exceptionally well in robust synthesis and error correction. The ideal configuration depends on the particular task and the intended trade-off between faithfulness to retrieved evidence and utilizing the model's general capabilities; there is no one best approach.

Third, different robustness challenges require different solutions. GraphRAG's community summarization is exceptionally well suited for the information integration task. Its structured representation is highly effective for counterfactual detection. However, it does not inherently solve the negative rejection problem, which is more related to prompt design and the model's instruction following propensity regarding its own knowledge boundaries.

A limitation of our experimental design lies in the dual use of GPT-4o-mini, which was employed both during the GraphRAG preprocessing stages (i.e., entity and relation extraction, as well as community summarization) and as one of the downstream answering models. This introduces a potential representational compatibility bias, where the observed performance gains may partially stem from an inherent alignment between the structure of the generated KG and the model's own internal representations, rather than solely from the effectiveness of the graph-enhanced retrieval pipeline. Consequently, some improvements attributed to GraphRAG should be interpreted with caution, as they may reflect model-to-model affinity in addition to methodological advantages.

Finally, these findings lay a crucial foundation for the subsequent chapter. They demonstrate that we can significantly alter and improve a RAG system's behavior towards factual information through structuring and prompt engineering. This sets the stage for our next investigation: if we can steer a model's handling of facts so effectively, to what degree can we also measure and understand its steering by the ideological framing present within that same structured knowledge? The tools and configurations validated here will be applied to this novel question in the following chapter.

Chapter 5

IDEOLOGICAL INFLUENCE IN RAG SYSTEMS

Preamble

This chapter is compiled from material extracted from a manuscript to be submitted to the Conference on Computational Natural Language Learning **[C4]** (Amamou *et al.*, 2026). As the first author of this manuscript, I was responsible for conceptualization, methodology development, experimental design, implementation, data analysis, and writing of the original manuscript.

5.1 Introduction

This chapter describes the second basic experimental investigation of this thesis, which changes from information robustness to ideological influence in RAG. While Chapter 4 dealt with whether using information can increase robustness to incorrect or contradictory information, this chapter deals with a different but also subtle problem. What are the effects of current information being given from a practical point of view but having a different ideological orientation? While research on LLM bias predominantly analyzes parametric bias embedded during training, little work examines how externally supplied knowledge in RAG systems interacts with or amplifies this bias, creating a pathway for contextual ideological influence.

This investigation pioneers the application of Lexical Multidimensional Analysis (LMDA) a corpus-linguistic method for quantifying ideological discourse to evaluate AI systems. We construct a controlled experimental framework using U.S. presidential debates as ideologically rich source material. After extracting six distinct ideological dimensions via LMDA, we systematically expose different RAG configurations and LLMs (GPT-4o-mini, GPT-3.5-turbo, Gemini 1.5 Pro, Claude 3 Sonnet) to these discourses. By measuring the semantic and lexical alignment between generated responses and predefined ideological descriptions, we quantify a system's susceptibility to discursive steering.

The findings demonstrate that ideological influence is a measurable and configurable property of RAG systems, with significant variation across models and configurations. This work establishes LMDA as a novel methodology for assessing discursive alignment and provides a crucial complementary perspective to purely factual robustness benchmarks, enabling a more holistic evaluation of RAG reliability.

5.2 Methodology: Measuring Discourse Alignment

5.2.1 Theoretical Foundation: Lexical Multidimensional Analysis (LMDA)

Ideological positions are not merely abstract sets of beliefs; they are expressed and reproduced through systematic patterns of language use. In relation to this, one important component of sociolinguistics and discourse analysis is that ideologies have distinguishable characteristics that, when used, fingerprint a certain discursive or communicative approach, called an Ideological Lexicon (Berber Sardinha & Fitzsimmons-Doolan, 2025). This chapter is the second key experimental study of this thesis, this time shifting the spotlight from retrieved factual robustness to the possible influence of ideology on RAG systems. While Chapter 4 aimed to assess whether KG-based systems are beneficial to the retrieval of information under noisy and conflicting information, this chapter will seek to assess the possibility of encountering ideologically different information when such information is, in fact, factually accurate. While research on LLM bias predominantly analyzes parametric bias embedded during training, little work examines how externally supplied knowledge in RAG systems interacts with or amplifies this bias, creating a pathway for contextual ideological influence.

This investigation pioneers the application of Lexical Multidimensional Analysis (LMDA)—a corpus-linguistic method for quantifying ideological discourse—to evaluate AI systems. We construct a controlled experimental framework using U.S. presidential debates as ideologically rich source material. After extract-

ing six distinct ideological dimensions via LMDA, we systematically expose different RAG configurations and LLMs (GPT-4o-mini, GPT-3.5-turbo, Gemini 1.5 Pro, Claude 3 Sonnet) to these discourses. By measuring the semantic and lexical alignment between generated responses and predefined ideological descriptions, we quantify a system's susceptibility to discursive steering.

The findings demonstrate that ideological influence is a measurable and configurable property of RAG systems, with significant variation across models and configurations. This work establishes LMDA as a novel methodology for assessing discursive alignment and provides a crucial complementary perspective to purely factual robustness benchmarks, enabling a more holistic evaluation of RAG reliability. Use of this involves repetition, such as when certain terms, phrases, metaphors, and schemes are used subtly to inscribe or encode a certain worldview or belief system. To illustrate this, one could use terms such as "free market," "individual liberty," or "limited government" when talking about classical liberalism, while terms such as "systemic inequality," "equity," or "collective action" may dominate when talking about social justice (Fitzsimmons-Doolan, 2024).

Lexical Multidimensional Analysis (LMDA), pioneered by Berber Sardinha & Fitzsimmons-Doolan (2025), is a corpus-based, factor-analytic technique designed to operationalize and measure such discourses. The methodology begins with the compilation of large, representative text corpora for different discursive domains. Statistical factor analysis (or Principal Component Analysis) is then applied to a matrix of lexical frequencies to identify latent **dimensions** underlying factors defined by sets of words that consistently co-occur across texts. Each dimension is interpreted as representing a distinct discursive pattern or ideological stance (Berber Sardinha, 2019). The output is a set of **discourse vectors**: quantitative representations in a high-dimensional lexical space where each dimension's coordinates are defined by the factor loadings (weights) of the characteristic words. A text's position in this space provides a precise, data-driven measure of its alignment with the extracted discourses (Berber Sardinha & Fitzsimmons-Doolan, 2025).

In this work, LMDA is leveraged not merely as a descriptive tool for human-authored texts but as a novel evaluation framework for AI systems. We apply LMDA to curated corpora to construct stable **Discourse Prototypes** vector representations of distinct ideological positions. Any new text generated by a RAG system can then be analyzed using the same lexical feature set and projected into this pre-defined discursive space.

5.2.2 Measurement: Semantic Similarity and Cosine Distance

The quantitative analysis of text alignment in this thesis relies on the concept of **text embeddings**: dense, fixed-dimensional vector representations where semantically similar texts are mapped to proximate points in a high-dimensional space. These embeddings are typically generated by neural models (e.g., sentence-transformers) that encode the semantic content of a text passage into a numerical vector (Brown *et al.*, 2020).

For finding the relationship between these two vectors, say, a response generated to a discourse prototype, **cosine similarity** metric will be used. Cosine similarity is a function that calculates the cosine of the angle between any given pairs of vectors, which can be defined by the formula involving the dot product of the vectors. It can be written as follows:

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \tag{5.1}$$

The result ranges from -1 (perfect opposition) to +1 (perfect alignment), with 0 indicating orthogonality. Crucially, cosine similarity measures the **orientation** of vectors, not their magnitude, making it ideal for comparing semantic direction in embedding spaces where vector length may be influenced by factors like text length rather than meaning (Bommasani *et al.*, 2021). In our framework, the cosine similarity between a generated text's embedding and a Discourse Prototype serves as the primary, quantitative metric of ideological alignment or influence.

5.2.3 Corpus Construction and LMDA Analysis

The foundation of this study is a carefully constructed corpus designed to contain explicit and analyzable ideological discourses. 45 U.S. presidential debate transcripts from 1960 to 2020 were used. Table 5.1 describes the corpus, which is over 575,000 words long and comprises 3,428 different texts, each of which represents a candidate's speaking turn. A multitude of real-world instances of politically charged language that deliberately and overtly expresses opposing ideologies can be found in this dataset.

To extract the underlying ideological structures from this large text collection, we employed **Lexical Multidimensional Analysis (LMDA)**, a corpus-linguistic method developed by Berber Sardinha & Fitzsimmons-Doolan (2025). Unlike topic modeling approaches, which cluster words based on their co-

Table 5.1 – US Presidential Debates corpus composition.

Criterion	Description
Definition	Televised US presidential debates
Time frame	1960 – 2020
Debates	45
Words	575,535
Texts	3428
Mean text length	167.89
SD	144.40
Min.	7
Max.	1507

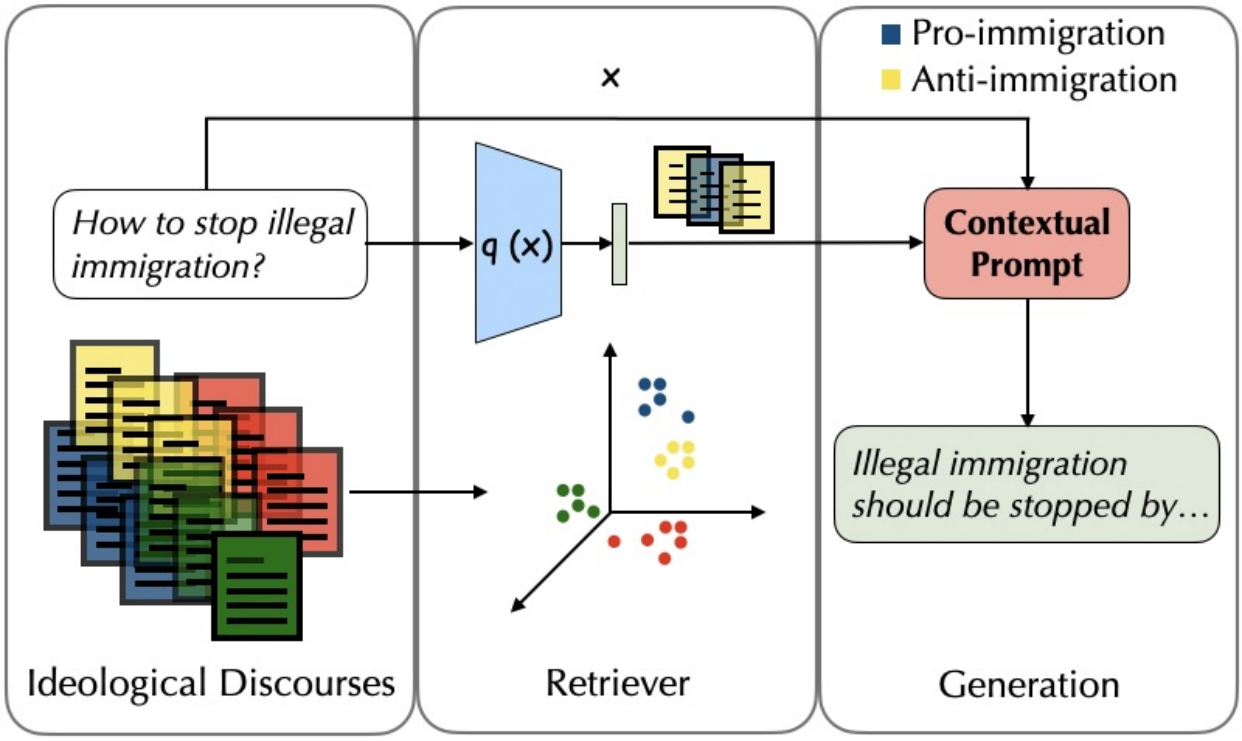


Figure 5.1 – Retrieval-based question answering where the LLM is provided with a context underlining different discourses for a given topic.

occurrence within documents, LMDA uses factor analysis to identify latent discursive dimensions based on the systematic co-occurrence of lexical features (nouns, verbs, adjectives, and adverbs) across the entire corpus (see Figure 5.1).

The methodology consisted of several stages. Firstly, the corpus was lemmatized and then the results were subjected to part-of-speech tagging, whereafter content words were obtained. Secondly, factor analysis was used for the frequency matrix representing the 1,000 most frequent lemmas. From the figure obtained for the scattered plot, it was apparent that the corpus required six factors to be significant. Next, the Promax

rotation was carried out on the corpus and resulted in the emergence of six factors, each comprising both positive and negative poles.

Each dimension was then subjected to a rigorous qualitative interpretation process. This was conducted by an expert in LMDA and corpus linguistics, who analyzed the highest-loading keywords for each pole and examined representative texts scoring highly on each dimension. Relying on domain expertise in the field of political discourse analysis, the expert provided descriptive terms and conceptual definitions that appropriately characterized the underlying ideological contrasts that each dimension's lexical profile represented. For example, Dimension 1 was given the following label: *"Equitable Prosperity through Social Responsibility vs. Traditionalist Conservatism"*, emphasizing the contrast between the keywords associated with economic fairness and moral absolutism, respectively. This expert-led interpretation ensures that the resulting dimension labels are not merely statistical artifacts but carry meaningful interpretive validity grounded in established discourse analysis practice. The six identified dimensions are described in Table 5.2.

Table 5.2 – Dimension labels and description for the discourses within the U.S. presidential debates corpus.

Dim.	Short Labels	Dimension Description (short)
1	Equitable Prosperity through Social Responsibility vs Traditionalist Conservatism	**Economic Focus:** Fiscal responsibility, fair taxation, middle-class support vs **Moral Focus:** Moral absolutism, sanctity of life, preservation of traditional values
2	Guardianship of Freedom: America's Global Mission vs Sovereign Libertarian Conservatism	**Global Leadership:** U.S. as global defender of democracy and freedom; moral responsibility to counter authoritarianism. Security vs **Limited Government:** Emphasis on personal responsibility, limited government, law enforcement, borders, and strict security controls
3	American Leadership through Strength and Strategic Alliances vs Constitutional Traditionalism and Federalist Exceptionalism	**Nuclear Security:** Urgent action against nuclear proliferation and terrorism; credible, multilateral foreign policy. Constitutional vs **Traditional Values:** Limited federal power, state sovereignty, individual responsibility, and traditional family/moral values
4	Community-Driven Law-and-Order Conservatism with Civic Accountability vs Cold War Geopolitical Dominance and Militarized Exceptionalism	**Law & Public Safety:** Support for tough law enforcement, community-based safety, and gun control to protect society and uphold moral responsibility vs **Defense & Global Competition:** Focus on military strength, defending allies, preserving U.S. influence, and proactive crisis response against global threats
5	Strategic Energy Independence with Sustainable Innovation and Leadership vs Sovereign Statism with Authoritarian National Security	**Balanced Energy Independence:** Support for traditional and renewable energy, with a focus on national independence, responsible resource use, and job creation vs **Military Sovereignty & Protectionism:** Emphasis on military strength, sovereignty, protection of freedoms, and defense-oriented, protectionist policies
6	Transformative Educational Equity: Accountability, Access, and Opportunity vs Conservative Governance: Fiscal Discipline, Institutional Stability, and Incremental Reform	**Education & Equal Opportunity:** Emphasis on quality education, accountability, and fairness for all children as keys to democracy and economic growth vs **Fiscal Restraint & Caution:** Focus on fiscal conservatism, cautious reforms, limited government spending, and measured approach to governance and foreign policy

The six identified dimensions span a range of politically sensitive topics, including economic policy, global affairs, national security, and social values. Crucially, this process yielded two key resources for the present study:

1. A labeled dataset of ideologically loaded texts, in which each text is associated with its primary ideological dimension(s) based on factor scores; and
2. A precise textual description of each ideological dimension, authored by the LMDA expert, which serves as a semantic reference point for measuring ideological alignment.

For this study, we focused exclusively on texts loading on the positive pole of each dimension, thereby constructing six distinct ideological sub-corpora that served as controlled external knowledge sources for our RAG experiments.

5.2.4 Experimental Design and Configurations

To measure ideological influence, we designed a controlled RAG style experimental framework where the retrieved context is both relevant and noiseless, yet deliberately ideologically charged. For each of the six ideological dimensions, we formulated two questions directly related to the dominant themes of that dimension (see Table 5.2 for examples). Figure 5.2 shows an overview of the pipeline used to extract LLM responses under the five configurations. The core experimental design systematically manipulates the

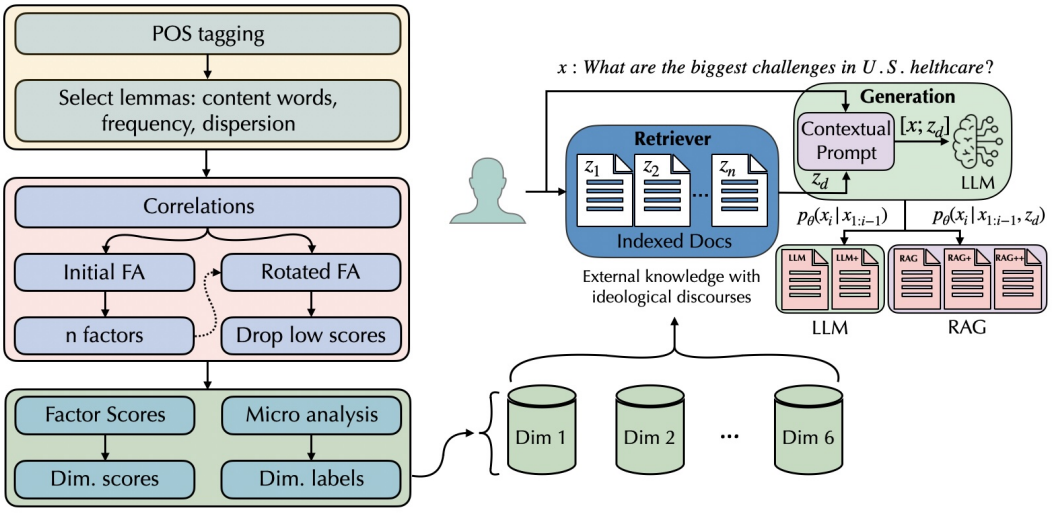


Figure 5.2 – An overview of the pipeline to extract LLM responses under five configurations. The external knowledge is designed to contain documents with a specific ideological discourse.

knowledge sources available to the LLM. We defined five configurations used in this study are described in Table 5.3:

For all RAG based configurations, the retriever was implemented as a semantic similarity search over a vector database containing exclusively texts from the targeted ideological dimension. This ensured that the

Table 5.3 – Descriptions of the types of knowledge interaction used by LLMs to generate responses.

Model	Responses based on
LLM	Encoded knowledge only
LLM+	Encoded & extensive knowledge
RAG	External knowledge only
RAG+	Encoded & external knowledge
RAG++	Encoded & extensive & external knowledge

retrieved context remained fully discourse-consistent. The prompts associated with each configuration were carefully engineered to explicitly constrain the allowable knowledge sources.

Four cutting-edge LLMs were assessed: Claude 3 Sonnet, Gemini 1.5 Pro, GPT-3.5-turbo, and GPT-4o-mini. These models were selected to ensure diversity in providers and architectures, as well as to include state-of-the-art systems with strong generative and reasoning capabilities, thereby improving the robustness and generalizability of our evaluation. Each model was evaluated across all five configurations and six ideological dimensions. For each query, the model generated five independent responses (temperature = 0.7); semantic similarity was computed for each response individually, and the final score corresponds to the average across the five runs.

The primary evaluation metric was **semantic similarity** (cosine similarity) between the embedding of the aggregated model response and the embedding of the official textual description of the corresponding ideological dimension, both computed using a pre-trained BERT model. To enhance interpretability, we complement the cosine similarity analysis with qualitative examples and a linguistic examination of model outputs, illustrating how different similarity levels translate into concrete ideological framing in generated responses.

5.3 Experimental Results and Analysis

5.3.1 LMDA Dimensions and Dataset

Six distinct ideological dimensions were successfully identified by the LMDA analysis, each with interpretable themes and coherent lexical patterns (see Table 5.2). This process produced a novel, annotated dataset where 3,428 debate texts are categorized by their primary discursive loading. For this study, we fo-

Table 5.4 – Semantic similarity between ideology description and LLMs' responses based on five configurations.

Model	Config	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
GPT-4omini	*LLM*	0.27	0.30	0.45	0.55	0.46	0.39
	LLM+	0.31	0.31	0.45	0.55	0.47	0.40
	RAG	**0.36**	**0.44**	0.41	0.50	**0.48**	**0.45**
	RAG+	0.32	0.36	**0.48**	**0.56**	0.47	0.41
	RAG++	0.32	0.35	**0.48**	**0.56**	**0.48**	0.40
GPT-3.5	*LLM*	0.30	0.36	0.48	0.56	0.50	0.38
	LLM+	0.30	0.34	0.48	0.55	0.49	0.40
	RAG	0.36	0.39	0.45	0.53	**0.54**	**0.44**
	RAG+	**0.40**	**0.41**	**0.53**	**0.57**	0.51	0.43
	RAG++	0.36	0.40	0.51	0.54	0.50	0.42
Gemini	*LLM*	0.26	0.34	**0.46**	0.54	0.47	0.38
	LLM+	0.26	0.32	0.44	0.53	0.47	0.39
	RAG	0.35	0.35	0.40	0.45	0.42	0.38
	RAG+	**0.36**	**0.36**	**0.46**	0.52	0.45	0.40
	RAG++	0.33	0.35	0.45	**0.55**	**0.49**	**0.41**
Claude	*LLM*	0.31	0.32	**0.45**	0.52	0.47	0.36
	LLM+	0.32	0.31	0.44	**0.55**	0.46	0.39
	RAG	**0.35**	0.36	0.41	0.49	0.46	0.41
	RAG+	0.32	**0.39**	0.44	**0.55**	0.45	**0.44**
	RAG++	0.32	0.35	**0.45**	0.54	**0.50**	**0.44**

cused on the positive pole of each dimension, creating six ideologically homogeneous vector databases that served as controlled external knowledge sources.

5.3.2 Alignment of LLM Responses with Ideological Discourses

The central finding, presented in Table 5.4, is clear: **RAG configurations significantly increase the alignment of LLM responses with provided ideological discourses compared to configurations relying solely on internal knowledge**. Across almost all dimensions and models, the highest similarity scores were recorded under RAG based configurations.

GPT-3.5 emerged as the most susceptible model, achieving the highest absolute similarity scores across multiple dimensions under RAG+ configuration. GPT-4o-mini showed strongest alignment under the external-only RAG configuration for most dimensions, suggesting its robust internal knowledge acts as a buffer unless explicitly suppressed.

Configuration type had a statistically significant impact on alignment scores for five of the six dimensions, according to an ANOVA test (Table 5.5).

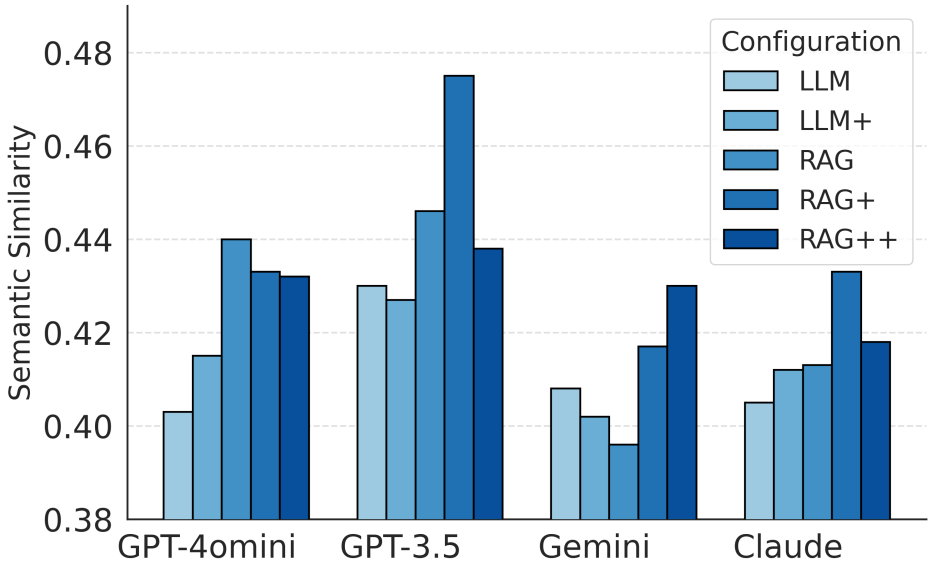


Figure 5.3 – Overall similarity scores for LLMs responses based on their encoded knowledge and LLM outputs based on RAG.

Table 5.5 – One-Way ANOVA results assessing the effect of configurations on the identified dimensions.

Dimension	F-value	p-value	($p < 0.05$)
Dim 1	6.3088	0.0035	✓
Dim 2	4.6040	0.0126	✓
Dim 3	3.3342	0.0384	✓
Dim 4	5.7367	0.0052	✓
Dim 5	0.4123	0.7971	✗
Dim 6	4.0115	0.0208	✓

Table 5.6 – Similarity scores between LLMs' responses (*LLM*) and responses based on three RAG configurations. For *RAG+* and *RAG++* the differences between their results and the result of *RAG* is provided.

	Dim 1			Dim 2			Dim 3			Dim 4			Dim 5			Dim 6		
Model	*RAG*	*RAG+*	*RAG++*	*RAG*	*RAG+*	*RAG++*	*RAG*	*RAG+*	*RAG++*	*RAG*	*RAG+*	*RAG++*	*RAG*	*RAG+*	*RAG++*	*RAG*	*RAG+*	*RAG++*
GPT-4omini	0.83	+0.08	+0.11	0.71	+0.21	+0.20	0.73	+0.23	+0.22	0.80	+0.15	+0.16	0.81	+0.10	+0.12	0.86	+0.09	+0.08
GPT-3.5	0.89	+0.02	+0.06	0.84	+0.07	+0.07	0.88	+0.05	+0.06	0.87	+0.09	+0.10	0.86	+0.03	+0.05	0.85	+0.10	+0.09
Gemini	0.77	+0.08	+0.11	0.71	+0.10	+0.10	0.66	+0.22	+0.22	0.73	+0.14	+0.17	0.69	+0.13	+0.15	0.75	+0.13	+0.12
Claude	0.82	+0.10	+0.11	0.77	+0.10	+0.14	0.76	+0.15	+0.17	0.81	+0.11	+0.13	0.79	+0.11	+0.12	0.81	+0.09	+0.11

5.3.3 Interaction Between Internal and External Knowledge

Analysis of similarity between LLM only outputs and RAG influenced responses reveals critical knowledge interaction dynamics (Table 5.6). For GPT-4o-mini, similarity scores increased substantially (+0.08 to +0.23) when allowed to combine knowledge (RAG+, RAG++), indicating it anchors responses in internal knowledge. GPT-3.5 showed minimal increases (+0.02 to +0.10), suggesting its responses diverge more from baseline and external discourse dominates the fusion.

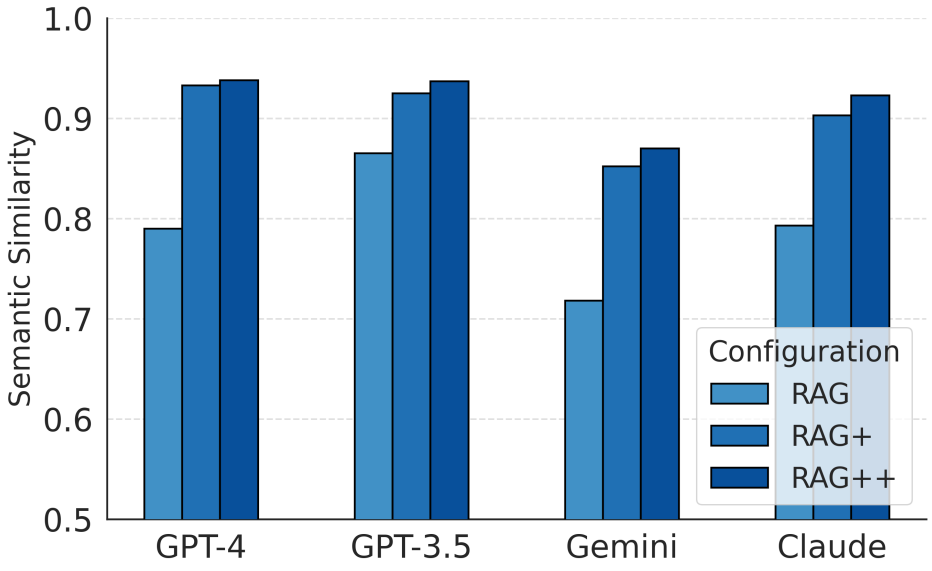


Figure 5.4 – Overall semantic similarity between responses based on LLMs' encoded and extensive knowledge (LLM+) and responses based on RAG configurations.

In contrast, GPT-3.5 exhibited a fundamentally different interaction pattern. Similarity scores between its LLM baseline and its RAG based outputs were already high under the RAG configuration (0.84–0.89), with only marginal increases observed under RAG+ and RAG++ (+0.02 to +0.10). This behavior indicates that GPT-3.5's internal knowledge is either less firmly anchored or more easily overridden. Its outputs diverge more significantly from its baseline position when exposed to external ideological discourse, and the external discourse tends to dominate the fusion process when knowledge sources are allowed to be combined. GPT-3.5 is therefore more malleable and more vulnerable to ideological reshaping.

5.3.4 GraphRAG Configuration Performance on Ideological Influence

Building upon the findings from Chapter 4 regarding GraphRAG's effectiveness for factual robustness, we conducted an additional experiment to evaluate whether structured knowledge representation affects a system's susceptibility to ideological influence. Specifically, we compared three configurations: standard RAG as a baseline, G-RAG using the default GraphRAG pipeline with generic prompts for knowledge extraction and response generation, and G-RAG AD, an adaptive version where both the indexing prompts (for entity and relationship extraction) and the response generation prompts were specifically engineered to enhance discourse awareness and ideological alignment.

The results, presented in Table 5.7 and Table 5.8, reveal important insights about structured knowledge and ideological influence. When measuring semantic similarity (Table 5.7), we observe that structured knowledge representation does not inherently reduce ideological influence and can, with targeted prompting, significantly amplify it. The similarity score for the baseline RAG configuration for the GPT-4o-mini model ranged between 0.36 and 0.50. It was noticed that the basic GraphRAG (G-RAG) model performed erratically, sometimes better in certain aspects, such as in Dim 1, 0.42 compared to 0.36 in the basic model, while at other times it performed worse, such as in Dim 3, 0.19 compared to 0.41 in the basic model. This suggests that naive graph structuring does not necessarily counter discursive alignment. However, the adaptive configuration (G-RAG AD) showed the highest semantic similarity scores on five out of six ideological dimensions for both models, with GPT-4o-mini showing an improvement of 0.07-0.13 compared to G-RAG.

Table 5.7 – Semantic similarity scores for standard RAG and GraphRAG configurations across ideological dimensions. The adaptive GraphRAG version (G-RAG AD) consistently achieves higher alignment with ideological discourses compared to both standard RAG and default GraphRAG.

Model	Config.	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
GPT-4o.mini	RAG	0.36	0.44	**0.41**	**0.50**	**0.48**	0.45
	G-RAG	0.42	0.44	0.19	0.17	0.32	0.60
	G-RAG AD	**0.49**	**0.51**	0.26	0.24	0.45	**0.62**
GPT-3.5	RAG	0.36	0.39	**0.45**	**0.53**	**0.54**	0.44
	G-RAG	0.37	0.37	0.16	0.17	0.28	0.53
	G-RAG AD	**0.41**	**0.45**	0.18	0.20	0.34	**0.58**

The lexical similarity results using TF-IDF (Table 5.8) reveal an even stronger and more consistent pattern. Across both models and all dimensions, G-RAG AD achieves the highest lexical alignment scores. For GPT-4o, the adaptive configuration shows improvements of 0.02-0.11 points over standard RAG and 0.02-0.08 points over default G-RAG across dimensions. This lexical analysis demonstrates that the adaptive prompts cause the models to adopt not just the semantic concepts but the specific vocabulary and terminology characteristic of the target ideological discourses.

Several critical patterns are also observable. Firstly, Dimension 6, which concerns Educational Equity vs. Fiscal Conservatism, consistently has the highest degree of alignment with all configurations, with GPT-4o attaining a peak lexical similarity score of 0.64 with G-RAG AD. Secondly, GPT-4o consistently outperforms GPT-3.5 in semantic as well as lexical alignment with all configurations, suggesting that stronger models are capable of generating more coherent, richly-terminological discourse-aligned texts. Most significantly, however, G-RAG AD's strong performance in both semantic and lexical alignment suggests that discourse-

Table 5.8 – Lexical similarity (TF-IDF) for standard RAG and GraphRAG configurations across ideological dimensions. The adaptive GraphRAG version (G-RAG AD) consistently achieves the highest lexical alignment, indicating adoption of discourse-specific vocabulary.

Model	Config.	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
GPT-4o.mini	RAG	0.53	0.51	0.47	0.47	0.46	0.56
	G-RAG	0.54	0.52	0.49	0.50	0.45	0.57
	G-RAG AD	**0.58**	**0.56**	**0.51**	**0.52**	**0.56**	**0.64**
GPT-3.5	RAG	0.48	0.43	0.44	0.41	0.38	0.48
	G-RAG	0.50	0.49	0.48	0.49	0.41	0.55
	G-RAG AD	**0.50**	**0.48**	**0.48**	**0.47**	**0.43**	**0.56**

aware prompt engineering at both the indexing and generation levels is capable of systematically enhancing levels of ideological alignment. What is perhaps most interesting, however, is that while Chapter 4 discussed how optimizations for robustness with structured knowledge are important, they also inadvertently create systems that are more prone to ideological manipulation unless specifically designed to counter such a threat.

5.4 Discussion of Findings

The experimental results presented in this chapter yield several important implications for the reliability and trustworthiness of RAG systems. First, they provide empirical evidence that RAG pipelines can act as a pathway for ideological influence, as evidenced by the measurable alignment shifts observed under controlled experimental conditions. Supplying an LLM with externally retrieved context is not a neutral enhancement of factual capability; rather, it acts as a form of discursive framing. It can be observed that the generated outputs have a clear correlation and alignment with the ideological orientation and tone of the provided sources, as evidenced through semantic and lexical analysis. This risk does not only cover factual inaccuracies but also encompasses the nuanced dissemination of perspective under the guise of authority as an AI assistant.

Second, the findings expose a critical model-dependent vulnerability. The substantial variation in susceptibility between models such as GPT-3.5 and GPT-4o indicates that resilience to discursive influence is not an inherent property of RAG systems. More capable models generally maintained higher alignment scores, producing outputs that were not only more coherent but also adopted the specific vocabulary (as shown in Table 5.8) characteristic of the target discourses. This suggests that while advanced models may appear more "knowledgeable," they can also produce more persuasive and terminology-rich ideologically aligned content when influenced by biased context.

Third, the analysis of structured knowledge systems reveals a crucial and potentially counterintuitive dimension: while GraphRAG was shown in Chapter 4 to improve factual robustness, our results here demonstrate that this structural enhancement does not inherently confer resilience against ideological influence. In fact, the adaptive GraphRAG configuration (G-RAG AD) achieved the highest alignment scores across both semantic and lexical metrics, revealing that the same prompt engineering techniques used to boost factual performance can be co-opted to systematically amplify discursive steering at multiple linguistic levels.

The interaction analysis highlights a fundamental tension in RAG design that creates complex trade-offs. Such configurations as RAG (external-only), which minimize confabulation on factual questions (as demonstrated in Chapter 4 on negative rejection) also maximize susceptibility to ideological framing when retrieved context is biased. On the other hand, configurations that allow internal knowledge integration may actually use that knowledge to amplify instead of counteract external bias, making them both persuasive and ideologically slanted.

One significant limitation of this study is the scope of linguistic feature analysis employed. The research primarily utilizes semantic similarity metrics to assess how closely the models' outputs align with ideological dimensions and external contexts. While these metrics are useful for capturing overarching similarities, they may overlook finer-grained linguistic nuances. Expanding the analysis to include a broader range of linguistic features provides a more comprehensive understanding of how LLMs generate content under different knowledge configurations. By examining these additional features, future studies might uncover subtle shifts in language use or rhetorical strategies that are influenced by the interplay between the models' prior knowledge and the injected external information.

Finally, these results also point to the necessity of a paradigm shift in the evaluation of RAG systems. In order to properly evaluate RAG systems, a new set of benchmarks is required that goes beyond factual accuracy. In fact, as demonstrated here using Lexical Multidimensional Analysis (LMDA) and both semantic and lexical similarity measures, ideological bias is a measurable, controllable, and significant factor in RAG system performance. A new set of benchmarks is required that can evaluate a system's capacity for discursive alignment across different ideological contexts, balancing the twin requirements of factual robustness and ideological resilience.

Chapter 6

Conclusions and Future Work

In this dissertation, a comprehensive examination of the two pillars of reliability, i.e., factual robustness and immunity to ideological influence, in RAG systems was conducted. Following a comprehensive parallel experimental examination, this journey has not only explored individual aspects but has extended to uncovering their interlinked nature by recognizing how both domains contribute to the development of trustworthy AI. In this final chapter, it entails an outline of the entire experience, an articulation of its broader meanings, and paving a new way forward to the important work to be done.

6.1 Summary of Contributions

This thesis makes four distinct and substantive contributions to the fields of NLP and Trustworthy AI. First, it provides a rigorous empirical analysis of KG-enhanced robustness in RAG. Through systematic experimentation with the RGB benchmark, we demonstrated that structuring retrieved information via a GraphRAG framework significantly mitigates the impact of noisy, contradictory, and multi-document information, particularly for less capable models such as GPT-3.5. Our customizations notably the GR_{comb} (combined knowledge) and GR_{ext} (external only) configurations yield actionable insights into how controlled knowledge integration can dramatically improve error detection, correction, and synthesis, moving the discourse on RAG robustness from abstract concern to a quantified architectural solution space.

Second, this thesis introduces a novel methodological framework for quantifying ideological influence in RAG systems. By bridging corpus linguistics specifically Lexical Multidimensional Analysis (LMDA)—with

modern NLP evaluation techniques, we developed a replicable process to extract ideological discourses from text corpora and measure their alignment with LLM outputs. This innovation provides researchers with a powerful new lens to examine bias, shifting from broad fairness notions toward precise, quantifiable measurement of discursive influence. As part of this contribution, we also present a new resource for bias evaluation: an LMDA-annotated dataset derived from U.S. presidential debates, categorized along six identifiable ideological dimensions. This dataset fills a notable gap in existing evaluation suites by serving as a benchmark for assessing the political and discursive neutrality of AI systems operating in retrieval-augmented contexts.

Third, and most significantly, this thesis offers a synthesized analysis of the RAG reliability landscape. It establishes that improvements in factual robustness and susceptibility to ideological influence are not independent but are deeply interconnected through mechanisms of knowledge source control, model capability, and context structuring. The paradox indicated that the setting that performs best regarding the avoidance of factual hallucinations (GR_{ext}/RAG) is in fact the setting that is the most prone to ideological alignment. The above is an extremely important issue both for researchers and practitioners. It raises the need for a holistic approach to the RAG system design, focusing on the dual requirements of fidelity and discursiveness as non-competing requirements.

6.2 Derived Publications

The work presented in this thesis has led to several publications in international journals and conference proceedings. Chapter 4 is primarily derived from Amamou *et al.* (2025), published in the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC), which presents the experimental framework and results for KG–enhanced factual robustness in RAG systems. Chapter 5 draws from a manuscript to be submitted to the Conference on Computational Natural Language Learning (CoNLL 2026) (Amamou *et al.*, 2026), detailing the application of Lexical Multidimensional Analysis to measure ideological influence, with complementary methodological validation provided by a related case study on COVID-19 discourse currently under review (Salari *et al.*, 2026). Additional publications have informed the broader research context: studies on fake news detection in Portuguese (Silva *et al.*, 2025) and comparative analyses of machine learning and RAG approaches (Sadfi *et al.*, 2026) explore complementary dimensions of misinformation and AI robustness, cross-lingual keyword extraction research (de Souza *et al.*, 2025) informs considerations of multilingual RAG deployment, and the development of a Quebec-focused French corpus for misinformation

detection (Doghman *et al.*, 2026) extends the thesis's methodological contributions to new linguistic and cultural contexts.

6.3 Limitations

However, despite the several critical contributions made by this thesis, it remains necessary to recognize its limitations. Firstly, despite our evaluation on factual robustness being presented with reference to the RGB benchmark alone, it remains to be explored to what extent the robustness of factual retrieval could be evaluated on additional benchmarks. RGB offers systematic perturbations which prove useful for specific failure modes but does not fully represent real-world retrieval scenarios where documents may vary in lengths and kinds.

Second, our ideational influence experiments, though methodologically novel, assume that cosine similarity between text embeddings and discourse prototypes is a valid proxy for ideological alignment. This assumption is theoretically sound in light of distributional semantics, but as yet untested by formal validation against human perception studies or other external criteria. It remains an open empirical question the extent to which the discursive proximity we measure corresponds meaningfully to the ways in which human readers perceive ideological framing or bias. Bridging this validity gap will be crucial in securing a wider diffusion of LMDA-based evaluation within the AI community.

Another potential limitation pertains to the subjective nature of the interpretation of ideological dimensions within the LMDA framework. Despite LMDA's quantitative foundation, the assignment of ideological meaning to the extracted dimensions relies on expert judgment, which may not be replicable. Nevertheless, the framework's validity is typically strengthened through independent inter-annotator agreement among domain experts.

Third, the results of the ideological dimension analysis via the LMDA method correlate solely with American presidential debates. Although this material constitutes rich and reliable ecological information, it must be realized that the applicability of these prominent and particular aspects and factors to other forms of politics, other languages, and other domains such as business, science, and media coverage remains unclear. Finally, our experiments were conducted on a limited set of LLMs. Although we selected models spanning different capability tiers and providers, the rapid evolution of foundation models means that newer architec-

tures may exhibit different robustness and susceptibility profiles. Replicating and extending this work across a broader and more current model landscape is a necessary direction for future research.

6.4 Implications for Research and Practice

The research's conclusions have important ramifications for a number of stakeholders. This work highlights the shortcomings of single-dimensional evaluation for AI researchers. Advancing RAG systems requires benchmarks of factual fidelity and discursive fairness to move forward together. The model-dependence of vulnerabilities demonstrated here-for instance, the plasticity of GPT-3.5 versus the resilience of GPT-4o-mini-emphasizes a more nuanced understanding of how model scale, architecture, and training are shaping not just capability but also robustness and bias profiles.

For practitioners and developers deploying RAG systems, this study serves as both a warning and a guide. The choice of the underlying LLM, the design of the retrieval pipeline, and, most critically, the prompt engineering strategy for knowledge integration are not merely technical decisions but ethical and reliability calibrations. Blindly implementing RAG pipelines to "ground" an LLM without considering the quality and perspective of the retrieval sources, or the model's susceptibility to being overridden by them, can inadvertently produce systems that are efficiently biased or factually fragile. This work therefore supports the introduction of guardrails such as source attribution, confidence scoring, and clear policy for contradictions or ideology-laden content.

This finding has far-reaching implications at the societal level: it points out the intrinsic tension in augmenting AI with additional knowledge. The web and other dynamic corpora are not neutral repositories of truth, but arenas of perspective, debate, and influence. Consequently, RAG systems should be regarded as active members of a discursive ecosystem and not as truth-tellers. Besides that, this requires more literacy on the part of the user when dealing with retrieval-augmented AI systems and transparency with regard to sources of information and limitations.

6.5 Future Research Directions

Upon building on the foundation established by this thesis, some potential and essential research directions are identified. The most obvious route is to establish and verify a function that optimizes for factual

robustness and discursive resilience. This includes building discourse-aware retrievers that are able to identify ideological framing within text retrievals, constructing KGs that include source metadata, and building knowledge integration controllers that are able to balance between internal parametric knowledge and external information. Identifying effective methods for real-time discursive scoring and optimal controller policies under varying constraints remains a critical open challenge.

Equally pressing is the need for comprehensive, multi-dimensional benchmarks that move beyond the current siloed evaluation paradigms. Future evaluation suites will need to include factual robustness tracks, discursive influence tracks, as well as hybrid scenarios where both types of perturbations are used. These types of benchmarks will also need to go beyond the current domain of U.S. politics to include a wide variety of cultural, linguistic, and domain-specific sorts of ideological discourse. These will all need to be included as necessary to develop systems that are robust in a wide variety of contexts around the globe.

Fundamental research is also required on multiple fronts. First, the alignment metrics introduced in this thesis require formal validation through human perception studies that examine the correlation between cosine similarity-based discursive alignment scores and human judgments of ideological framing and bias. Second, there is a need for algorithmic innovation to tackle these trade-offs, perhaps via debate-based training approaches, reinforcement learning from diverse human feedback, or adversarial training on contexts with different ideological presumptions to better cultivate critical source examination skills for LLMs. Thirdly, there remains an unexplored avenue to directly incorporate large-scale curated KGs into the RAG framework to provide core functionality for establishing the veracity and deliberative verifications of LLM responses. In all instances, these challenges need to be addressed to guarantee that as LLMs grow more powerful models, their integrity, clarity, and alignment with intricate human values keep pace.

Bibliography

Amamou H, Gagnon S, Davoust A & Avila AR (2025). Towards robust retrieval-augmented generation based on knowledge graph: A comparative analysis. *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 80–85, Vienna, Austria.

Amamou H, Salari E, Chen RB, Delfino MCN, Yamada R, Berber-Sardinha T, Kshirsagar S, Gagnon S, Davoust A & Avila AR (2026). Investigating llms' alignment with ideological discourses in retrieval-augmented generation. *Conference on Computational Natural Language Learning (CoNLL 2026)*. To be submitted.

Bender EM *et al.* (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

Berber Sardinha T (2019). Using multidimensional analysis to detect representations of national identity. *Multiple Perspectives on Language Ideology and Practice*, Routledge.

Berber Sardinha T & Fitzsimmons-Doolan S (2025). *Lexical Multidimensional Analysis*. Cambridge University Press.

Bommasani R *et al.* (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown TB *et al.* (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Buçinca Z *et al.* (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*.

Caliskan A, Bryson JJ & Narayanan A (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chen J *et al.* (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

de Souza JV, Amamou H, Chen RB, Salari E, Gubelmann R, Niklaus C, Serpa T, Lima MMdF, Pinto PT, Kshirsagar S, Davoust A, Handschuh S & Avila AR (2025). Cross-lingual keyword extraction for pesticide terminology in brazilian portuguese and english. *Journal of the Brazilian Computer Society*, 31(1):972–989. DOI:10.5753/jbcs.2025.5815.

Doghman M, Amamou H, Davoust A & Avila AR (2026). Infox-qc: A quebec-focused french corpus for misinformation detection and ai robustness assessment. *International Conference on Language Resources and Evaluation (LREC 2026)*. Accepted, in press.

Edge D *et al.* (2024). From local to global: A graphrag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Feng S *et al.* (2023). From pretraining data to language models to downstream tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Fitzsimmons-Doolan S (2024). Using lexical variables to identify language ideologies in a policy corpus. *Corpora*.

Gao Y *et al.* (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2309.01431*.

Hogan A *et al.* (2021). Knowledge graphs. *ACM Computing Surveys*.

Izacard G & Grave E (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Ji Z *et al.* (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Kadavath S *et al.* (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lewis P *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*.

Lin BY *et al.* (2021). Trig: Transformers for knowledge graph reasoning. *arXiv preprint arXiv:2111.07178*.

Ling C *et al.* (2023). Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Sadfi Y, Amamou H, Davoust A & Avila AR (2026). Comparative analysis of machine learning, llms, and rag for fake news detection. *13th International Conference on Computational Intelligence (CAI 2026)*. Accepted, in press.

Salari E, Delfino MCN, Amamou H, de Souza JV, Kshirsagar S, Davoust A & Avila AR (2026). Uncovering ideological bias in rag with lexical multidimensional analysis: A case study on covid-19. *Journal of Semantics (SEM 2026)*. Under review.

Silva RM, Amamou H, Ferraz LBS, da Silva FKA & Avila AR (2025). Fake news detection in portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1149–1166. DOI:10.5753/jbcs.2025.5525.

Sun J *et al.* (2023). Knowledge graph augmented language models for knowledge-intensive tasks. *arXiv preprint arXiv:2310.01394*.

Thakur N *et al.* (2024). "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. *Findings of the Association for Computational Linguistics: EMNLP*.

Traag VA, Waltman L & van Eck NJ (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Wang S *et al.* (2023). Knowledge editing for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Weidinger L *et al.* (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Yang X *et al.* (2025). Crag: Comprehensive rag benchmark. *Advances in Neural Information Processing Systems*.

Appendix

For the purpose of clarity and reference, this appendix offers additional information for the experimental chapters of this thesis, arranged by chapter.

Supplementary Materials for Chapter 3

Robustness in Knowledge-Augmented RAG Systems

6.6 Complete Prompt Templates for GraphRAG Configurations

6.6.1 GR_{RGB} Prompt

The prompt used for the GR_{RGB} configuration adapts the standard RGB benchmark prompt to operate with Knowledge Graph context. The complete prompt template is illustrated in Figure 6.1.

GR_{RGB} Contextual Prompt

You are an accurate and reliable AI assistant that can answer questions with the help of external documents. Please note that external documents may contain noisy or factually incorrect information. If the information in the document contains the correct answer, you will give an accurate answer. If the information in the document does not contain the answer, you will generate 'I cannot answer the question because of the insufficient information in documents.'. If there are inconsistencies with the facts in some of the documents, please generate the response 'There are factual errors in the provided documents.' and provide the correct answer.

STRUCTURED CONTEXT: {KG context}

OUTPUT: {answer}

Figure 6.1 – GR_{RGB} prompt presented to the LLM, based on the RGB system combined with structured knowledge.

GR_{ext} Contextual Prompt

You are a helpful assistant tasked with answering questions based on the provided external documents. The documents may contain noisy or irrelevant information. Your goal is to extract the necessary information to answer the question accurately, ignoring any irrelevant or noisy content. Instructions: 1. **Focus on Relevant Information**: Carefully analyze the documents and extract only the information that directly answers the question. 2. **Ignore Noise**: If a document contains irrelevant or noisy information, ignore it and focus on the parts that are relevant to the question. 3. **Provide a Clear Answer**: If the answer is found in the documents, provide it clearly and concisely. 4. **Reject if Necessary**: If the documents do not contain sufficient information to answer the question, respond with: *"I cannot answer the question due to insufficient information in the documents.

STRUCTURED CONTEXT: {KG context}

OUTPUT: {answer}

Figure 6.2 – GR_{ext} External-Only GraphRAG Prompt for Robust, Noise-Aware Answering

6.6.2 GR_{ext} Prompt

The external only configuration prompt strictly constrains the model to rely exclusively on the provided structured knowledge, rejecting any use of internal or prior knowledge. The full prompt is shown in Figure 6.2.

6.6.3 GR_{def} Prompt

The default GraphRAG prompt is designed for general question answering using synthesized community summaries derived from the Knowledge Graph. The corresponding prompt template is presented in Figure 6.3.

6.6.4 GR_{comb} Prompt

The hybrid GR_{comb} prompt combines both internal model knowledge and external structured knowledge to produce more comprehensive and validated answers. The complete prompt is provided in Figure 6.4.

GR_{def} Contextual Prompt

You are an AI assistant designed to provide accurate answers based on community summaries. Please answer the question using the structured context provided below, which represents synthesized information from relevant documents.

STRUCTURED CONTEXT: {KG context}

QUESTION: query

OUTPUT: {answer}

Figure 6.3 – GR_{def} GraphRAG's Default Prompt for Accurate, Evidence-Based Generation

GR_{comb} Contextual Prompt

You are a reliable AI assistant that can answer users' questions. Combine the external knowledge present in the retrieved context and your prior knowledge acquired during training to provide the most accurate and detailed answer to the question below. Use both sources to validate information and resolve any contradictions.

STRUCTURED CONTEXT: {KG context}

QUESTION: query

OUTPUT: {answer}

Figure 6.4 – GR_{comb} Hybrid Prompt for Noise-Robust Question Answering Using Documents and Prior Knowledge

Table 6.1 – Hyperparameter settings for Knowledge Graph construction

Parameter	Value	Description
Entity Confidence Threshold	0.7	Minimum confidence for entity inclusion
Community Detection Algorithm	Leiden	Community identification algorithm
Resolution Parameter	1.0	Community detection granularity
Retrieved Documents	5	Documents per query
Chunk Size	512 tokens	Document splitting size
Overlap Size	50 tokens	Chunk overlap

Table 6.2 – Model-specific parameters for answer generation

Parameter	GPT-3.5	GPT-4o-mini	Description
Temperature	0.7	0.7	Sampling temperature
Max Tokens	512	512	Maximum output length
Top-p	0.9	0.9	Nucleus sampling
Frequency Penalty	0.0	0.0	Repetition control

6.7 RGB Benchmark Experimental Parameters

Supplementary Materials for Chapter 4

Ideological Influence in RAG Systems

6.8 Prompt Templates

The two fundamental prompts given to the LLMs are described in Figures 6.5 and 6.6.Without the impact of outside knowledge, these prompts enable evaluating how the LLM's prior knowledge corresponds with the discourses. Additionally, it permits to evaluate the impact of the extensive knowledge on the results. Figure 6.7, Figure 6.8 and 6.9 describe the prompt presented to the LLMs based on the external knowledge.While Figure 6.7 prioritize only the external knowledge, Figure 6.8 and Figure 6.9 allow the assessment of how the external knowledge interacts with the LLM encoded and extensive knowledge.

Basic Prompt (LLM)

You are a reliable AI assistant that can answer users' questions. Use only your prior knowledge acquired during training to provide the most accurate and detailed answer to the question below.

QUESTION: {query}

OUTPUT: {answer}

Figure 6.5 – Basic prompt presented to the LLM, characterizing responses based on encoded knowledge only.

Table 6.3 lists the queries presented to the LLM. Note that the questions are defined by topics within each dimension.

Basic Prompt (LLM+)

You are a reliable AI assistant that can answer users' questions. Combine your prior knowledge acquired during training and your extensive knowledge to provide the most accurate and detailed answer to the question bellow.

QUESTION: {query}

OUTPUT: {answer}

Figure 6.6 – Basic prompt presented to the LLM, characterizing responses based on encoded and extensive knowledge.

Contextual Prompt (RAG)

You are a reliable AI assistant that can answer users' questions. Use only the external knowledge present in the retrieved context to provide the most accurate and detailed answer to the question below. Do not use your prior nor your extensive knowledge to answer the question. Use solely the provided context.

QUESTION: {query}

CONTEXT:
[1] {Passage text}
[2] {Passage text}
...
[5] {Passage text}

OUTPUT: {answer}

Figure 6.7 – Contextual prompt used to retrieve LLM responses based solely on the external knowledge.

Contextual Prompt (RAG+)

You are a reliable AI assistant that can answer users' questions. Combine the external knowledge present in the retrieved context and your prior knowledge acquired during training to provide the most accurate and detailed answer to the question below.

QUESTION: {query}

CONTEXT:
[1] {Passage text}
[2] {Passage text}
...
[5] {Passage text}

OUTPUT: {answer}

Figure 6.8 – Contextual prompt presented to the LLM, characterizing responses based on both encoded and external knowledge.

Contextual Prompt (RAG++)

You are a reliable AI assistant that can answer users' questions. Combine the external knowledge present in the retrieved context, your prior knowledge acquired during training and your extensive knowledge to provide the most accurate and detailed answer to the question below.

QUESTION: {query}

CONTEXT:
[1] {Passage text}
[2] {Passage text}
...
[5] {Passage text}

OUTPUT: {answer}

Figure 6.9 – Contextual prompt presented to the LLM, characterizing responses based on encoded, extensive and external knowledge.

Table 6.3 – Question topics used for ideological alignment analysis, with one question per topic and polarity.

Dim.	Positive Pole		Negative Pole	
	Topic	**Question**	**Topic**	**Question**
1	Economy	*How does inflation affect my everyday expenses, and what can I do to save money?*	Abortion	*How does abortion access affect women's health and rights across the country?*
	Healthcare	*How can I find affordable health insurance if I don't qualify for Medicaid?*	Secular State	*How does the separation of church and state impact public schools and religious practices?*
2	Global Affairs	*How do global affairs affect the U.S. economy and job market?*	Immigration	*How does immigration impact the U.S. economy and job market?*
	Freedom	*What are the limits to personal freedom in the United States?*	Hate Crime	*How can we prevent hate crimes and ensure justice for victims?*
3	Nuclear Threat	*How can the United States protect itself from a nuclear threat?*	Constitutional Amendments	*How do constitutional amendments impact American society?*
	Nuclear Proliferation	*How does nuclear proliferation impact global security and stability?*	Abortion	*How might changes to abortion rights impact women's healthcare access?*
4	Gun Control	*How do stricter gun control laws impact crime rates in the United States?*	Communist Threat	*What role did the Communist threat play in shaping U.S. foreign policy during the Cold War?*
	Criminal Justice Reform	*What are the main issues with the current criminal justice system in the United States?*	Diplomatic Relations	*Why are summit meetings important for international diplomacy?*
5	Clean Energy	*What are the benefits of switching to renewable and clean energy sources for the environment and economy?*	Foreign Policy	*How do U.S. foreign policy decisions affect relationships with key international allies?*
	Job Creation	*How can increasing investments in energy projects help create new jobs and boost the economy?*	Strict Policies	*How does the government intervene during a national emergency strike, and what impact does it have on the economy?*
6	Public Education	*What are the most effective strategies for improving public education in the U.S.?*	Tax Policy Reform	*How will tax policy reforms impact middle-class families and small businesses?*
	Education Funding	*How can local control of education funding improve schools in my community?*	Iraq and National Security	*How has U.S. involvement in Iraq impacted national security and global relations?*