OPEN UNIVERSITY OF KENYA

MASTER'S THESIS

PROBABILISTIC NEURO-SYMBOLIC AGENTIC AI PLATFORM FOR DECISION SUPPORT IN THE KENYAN BUDGET PROCESS

JOSEPHAT ONCHANGWA MOTONU

SCHOOL OF SCIENCE

A Master's Thesis Submitted in Partial Fulfilment of the Requirements for the Award of the Master of Science in Artificial Intelligence of the Open University of Kenya.

OPEN UNIVERSITY OF KENYA

OCTOBER 2026

COPYRIGHT

All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written permission of the author or The Open University of Kenya.

© 2026, Josephat Onchangwa Motonu

DECLARATION

I hereby declare that this thesis is my original work and that it has not been submitted for an award of a degree in any other university.

Candidate:

Name: Josephat Onchangwa Motonu Signature: 

Date: 15-10-2026

This thesis has been submitted with my supervisor's approval.

Supervisor:

Name: Dr. Ronald Ojino, PhD Signature: 

Date: 16/10/2025

DEDICATION

I dedicate this thesis to my beloved wife and family, whose unwavering love, patience, encouragement and support have been instrumental throughout this journey. This milestone is as much theirs as it is mine.

ACKNOWLEDGEMENTS

I would like to thank everyone who has helped me get to this point in my thesis through their guidance, support and encouragement.

Most of all, I owe my eternal gratitude to my supervisor DR. Ronald Ojino for the invaluable and robust academic guidance, consultative feedback, patience, availability, encouragement and allowing me access his valuable academic networks throughout the research process. This played a very significant role in the development of this study and consequently boosting its academic and professional contribution.

I would like to thank the Department of Computing and Informatics, School of Science and Technology, Open University of Kenya for giving me such a stimulating academic environment, resources and support on which this study is based. The faculty members who gave me insightful inputs that enhanced my learning and research.

I am so grateful to my dear wife, and family for their constant love, patience, encouragement and putting up with the times I was not around them. This provided the motivation and strength I so much needed to remain solidly grounded.

In addition, I appreciate my classmates and professional colleagues for their intellectual and technical engagements and willingness to share ideas, perspectives and experiences which added value to this process.

Finally, I would like to express my gratitude to my employer, Parliament of Kenya, for affording me flexibility in pursuing my studies and performing my duties as a Parliamentary employee.

Thanks to all those who have helped directly and indirectly to make this possible and my sincere thanks. This achievement would not have happened had you not supported us.

ABSTRACT

Even though participatory budgeting is constitutionally guaranteed in Kenya, public finance information is scattered across institutional websites, in varying formats, and reported at various stages, making it difficult to trace, analyse and use. This study used Design Science Research (DSR) approach to design, implement and evaluate a probabilistic neuro-symbolic agentic Artificial Intelligence (AI) platform for decision support in the Kenyan budget process. The architecture consisted of a source-centred knowledge graph, a loosely coupled Bayesian probabilistic forecasting model and a bounded natural language agentic interface. The main knowledge graph release included 21,618 statements, which were based on an analytical graph with 2,036 assertions on Controller of Budget expenditure. Assertions were sampled across 130 structures and were 100% schema conformant on structural validation, and transcription/reporting adjustments were made during targeted domain validation. This rolling-origin evaluation, using forecast dates, was better than the scope-median baseline (MAE = 0.1144), and is comparable to the Bayesian three-state forecasting model, which produced an equal-year MAE of 0.1134 under a forecast-date-constrained evaluation of 852 historical execution records and was capable of providing formal distributions of predictive intervals and execution-state probabilities. The model achieved a MAE of 0.1005 on a held-out benchmark of 50 questions across 25 Ministries, Departments, and Agencies (MDAs) for FY2025/26, performing well in comparison to an ungrounded Large Language Model with a MAE of 0.1321 and a retrieval-augmented LLM with a MAE of 0.1395. Bayesian fit diagnostics for the held-out FY2025/26 data demonstrated good convergence of parameters ($\hat{R} \leq 1.0015$ ≤1.0015, bulk ESS ≥1,674, tail ESS ≥3,177); whereas nominal 80% predictive intervals achieved 64% empirical coverage, indicating a need for continued monitoring of parameter calibrations. The separate graph retrieval did not improve the numerical accuracy in the configuration tested, but the combined architecture maintained the Bayesian predictions stored in the graph and provided inspectable source provenance. The results show that integrated neuro-symbolic agentic platform can provide uncertainty-aware and traceable advisory support for the public finance analysis, offering clear boundaries between evidence representation, statistical inference, and natural-language communication.

Keywords: Participatory Budgeting, Design Science Research, Knowledge Graph, Bayesian Probabilistic Forecasting, Neuro-Symbolic AI, Agentic AI, Decision Support Systems.

Table of Contents

COPYRIGHT 2
DECLARATION 3
DEDICATION 4
ACKNOWLEDGEMENTS 5
ABSTRACT 6
Table of Contents 7
List of Tables 9
List of Figures 10
List of Abbreviations and Acronyms 11
CHAPTER ONE: INTRODUCTION 12
1.1 Background of the Study 12
1.2 Statement of the Problem 13
1.3 Purpose of the Study 14
1.4 Research Objectives 15
1.4.1 General Objective 15
1.4.2 Specific Objectives 15
1.5 Research Questions 15
4.5 Significance of the Study 15
4.6 Scope and Delimitations 16
1.8 Assumptions of the Study 17
1.9 Limitations of the Study 17
1.10 Operational Definition of Key Terms 19
1.11 Organisation of the Study 20
CHAPTER TWO: LITERATURE REVIEW 21
2.1 Introduction 21
2.2 Theoretical Framework 21
2.3 Conceptual Framework 22
2.3.1 Overview 22
2.3.2 Conceptual Framework Diagram 23
..... 23
2.3.3 variables and Indicators 24
2.3.4 Relationship among the variables 25
2.4 Empirical Literature Review 25
2.4.1 Neuro-Symbolic AI for Governance data 26
2.4.2 Bayesian Probabilistic Reasoning 27
2.4.3 Agentic AI System 27
2.4.4 Artificial Intelligence for Civic Technology 28
2.5 Critique of Existing Literature 28

2.6 Research Gaps **30**
2.7 Summary of Literature Review **31**
CHAPTER THREE: RESEARCH METHODOLOGY ***32***
3.1 Introduction **32**
3.2 Research Philosophy **32**
3.3 Research Design and Approach **32**
3.4 System Architecture **33**
3.4.1 Component Roles and Information Flow 34
3.4.2 Interface Between the Components 35
3.5 Data Source and Design **35**
3.6 Neuro-Symbolic Design **38**
3.7 Probabilistic Reasoning and Bayesian Inference **42**
3.8 System Integration and Evaluation **47**
CHAPTER FOUR: DATA ANALYSIS, RESULTS AND DISCUSSION ***54***
4.1 Introduction **54**
4.2 Neuro-Symbolic and Evidence-Based Results **54**
4.3 Historical and operational Bayesian forecasting results **60**
4.4 Corrected FY2025/26 benchmark **66**
4.5 Cross-experiment synthesis **73**
CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS ***77***
5.1 Introduction **77**
5.2 Summary **77**
5.2.1 Summary for Objective 1 77
5.2.2 Summary for Objective 2 77
5.2.3 Summary for Objective 3 78
5.3 Conclusions **78**
5.4 Recommendations **80**
5.4.1 Recommendations for Policy 80
5.4.2 Recommendations for further research 80
5.5 Contribution to Knowledge **81**
REFERENCES ***84***
APPENDICES ***88***
APPENDIX A. IMPLEMENTATION IDENTIFIERS AND REPRODUCIBILITY DETAILS **88**

List of Tables

Table 1: Operational Definition of Key Terms

Table 2: MCMC Settings and Convergence Summaries for the Principal Bayesian Configuration

Table 3: Study Roadmap Linking Data, Methods and Evaluation

Table 4: Principal Graph Releases Used in the Analysis

Table 5: Forecast Samples and their Analytical Roles

Table 6: Historical Same-Cutoff Point Forecast Comparison

Table 7: FY 2026/27 Operational Execution Ratio Forecast

Table 8: Point Forecast Metrics for the FY 2025/26 Benchmark

Table 9: Interval and Class Probability Metrics for FY 2025/26 Benchmark

List of Figures

Figure 2.1: Conceptual Framework

Figure 3.1: System Architecture

Figure 4.1: Same-level topology of the development knowledge graph

Figure 4.2: Document-level topology within the institutional clusters

Figure 4.3: Parliamentary document node with file-level performance metadata

Figure 4.4: Air transport program subgraph with year-specific assertions and source-linked amounts

Figure 4.5: Equal-year MAE by model in the earlier historical graph comparison

Figure 4.6: Fiscal year MAE by model across the seven historical test folds

Figure 4.7: Posterior Medians and 80% and 95% prediction interval for the nine FY2026/27 Vote forecast

Figure 4.8: Mean absolute error and MDA-cluster bootstrap intervals for the corrected FY2025/26 benchmark

Figure 4.9: MDA-level paired difference in Mean Absolute Error

Figure 4.10: Forecast error by recurrent and development accounting scope

Figure 4.11: Empirical convergence of the reported central 80% interval

Figure 4.12: Mean predicted class probabilities and observed class frequencies for under, exact and over execution

List of Abbreviations and Acronyms

Abbreviation	Full Form
AI	Artificial Intelligence
DSR	Design Science Research
IFMIS	Integrated Financial Management Information System
LLM	Large Language Model
PDF	Portable Document File
OECD	Organisation for Economic Development and Cooperation
CoB	the Controller of Budget
OAG	Office of the Auditor General
NER	Named Entity Recognition
DV	Dependent Variable
IVs	Independent Variable
MVs	Moderating Variables
NACOSTI	National Commission for Science, Technology and Innovation
NLP	Natural Language Processing
RDF	Resource Description Framework
OWL	Web Ontology Language
ESS	Effective Sample Size
MDP	Markov Decision Process
UI	User Interface
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
NumPy	Numerical Python
SBERT	Sentence-BERT
BIRR	Budget Implementation Review Report
SHACL	Shapes Constraint Language
OCR	Optical Character Recognition
CRPS	Continuous Ranked Probability Score
MCMC	Markov Chain Monte Carlo
MDA	Ministries, Departments and Agencies
WIS	Weighted Interval Score

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

In Kenya, the participation of citizens in key public policy matters such as the budget process is a constitutional requirement. This is anchored in Article 10, 118, 201 and 221 of the Constitution of Kenya (2010) and these articles are further amplified by the Public Finance Management Act and its attendant Regulations. The requirements here are that citizens must be involved throughout the decision-making process from policy conception to approval and implementation.

Despite the existence of the constitutional and legal provisions, evidence shows that public participation especially on the budget process is well established in the ex-ante stage that comprises of conceptualization, formulation, planning, and approval, but is weak on the ex-post stage which encompasses implementation, reporting and evaluation (Parliament of Kenya, 2021; Muyundo, 2024; Gikibi & Kipchumba, 2025).

Whenever citizens make their contributions during the budget process, they do so through heterogeneous methods such as oral submissions in the physical public forums, written memoranda submitted through email or hand-delivered to Parliament and discussions and sentiments expressed on social media platforms. Conversely, official budget data is availed to public through published portable document files (PDFs) and spreadsheets. The datasets are published on the government websites as static reports and dashboards.

Distributed formats and institutional archives demand reconciliation workflows that are transparent, and can be inspected for their source context. This encourages the consideration of integrated decision-support tools instead of whether a portal already provides all analytical capabilities. The problem is practical, whether an AI-assisted approach also is beneficial for the outcomes of the various stakeholders is an empirical question, and how to link evidence between documents without blurring competing claims, changes and omissions.

Advancements in the field of artificial intelligence (AI) in recent times due to increased computational power, scalable data storage, natural language processing, deep learning and agentic AI, among others present a great opportunity to develop robust tools that could help in analyzing, structuring, interpreting and communicating complex budget information. For instance, rather than using the current static reports and dashboards, emerging AI tools such as

neuro-symbolic AI can be harnessed to help make public finance explainable and traceable (Garcez et al., 2022).

Probabilistic reasoning using Bayesian statistics creates systems that reason under uncertainty, which is common in government budgeting because of the delays and unpredictable revision of budgets for competing interests (Murphy, 2022). Additionally, up-and-coming agentic AI offers goal-driven interfaces that can make complex budget information easily accessible to non-expert audiences. All these capabilities would eventually allow for the creation of interactive and uncertainty aware decision support (Zheng et al., 2023).

Previous uses set precedent and are not a lack of technology. Semantic publication of budgets has been recently presented in Chile (Cifuentes-Silva et al., 2020) and representation of government decisions using ontologies has been explored in Greece (Serderidis et al., 2024). Institutional-transfer limitations and opportunities are also revealed by AI-based budget tagging (Guariso et al., 2023). These studies encourage the study of the relationship and comparative analysis of documentary evidence, forecasting and controlled communication with regards to the specific needs in Kenyan budget analysis.

In this context the study explored a source-centred representation that does not assert a single undifferentiated source of truth but which instead holds multiple documented claims. Bayesian forecasts were connected to qualified observations, and bounded agentic orchestration provided communication of evidence and storage of analytical outputs. The research rationale was to see what this integrated approach could prove, where other, simpler approaches could still hold a place, and which limitations would remain with the components combined.

1.2 Statement of the Problem

There is a need to reconcile Kenyan budget decision support data published on different institutional portals and versions of documents. Estimates, approved amounts, implementation reports and audit findings are various phases of the budget process and cannot be viewed as equivalent observations. Differences in identifiers, scope of accounting, reporting periods and revisions limit comparisons, and publication lags limit knowledge at a forecast date. The practical issue is thus not just access to documents: It's how to transform those documents, which are often scattered, into comparable, source-verifiable evidence and properly qualified forecasts. Portals' usability and semantic publication of budget documents are relevant areas of

research yet access and integration are not predictive of reliability (Nikiforova & McBride, 2021; Cifuentes-Silva et al., 2020).

AI can provide tools for budget information extraction, terminology alignment in documents, and answering natural language questions. But the AI-assisted analysis also comes with its own level of assurance issues: values extracted from the analysis should be vetted for accounting meaning, and answers generated by the AI should not be confused for vetted answers or calibrated forecasts. Guariso et al. (2023) show the potential of machine learning for budget tagging and highlight the constraints of the approach when data are not representative of the target institutional setting. This assists with context-specific assessment instead of if off-the-shelf AI works well with Kenyan budget analysis. Trustworthiness is also emphasized in the application of Generative-AI risk-management guidance throughout system design, use, and evaluation (Autio et al., 2024).

The research question is what is the role of an AI-assisted budget workflow in Kenya to maintain source and budget-stage meaning, generate temporally valid execution-ratio forecasts, and convey those forecasts without changing their numerical content – all in the face of limited empirical evidence in the literature. The study, therefore, focused on a source-centred, neuro-symbolic representation associated with separate Bayesian forecasting and bounded agentic orchestration. It assessed and gave scores for representation quality, forecast accuracy and uncertainty, and output preservation on their own, based on historical statistical baselines and a protected language-model benchmark. The problem was not about assuming AI can increase participation or trust or the quality of the decision, but about building the capacity and boundaries of that integration.

1.3 Purpose of the Study

This study was aimed at designing, implementing and evaluating a platform for source-linked and uncertainty-aware analysis of Kenyan Budget information, using AI tools. It used a source-centred neuro-symbolic representation, a separate Bayesian forecasting of execution-ratio and a bounded agentic interface. Evaluation focused on the organisation and verification of documentary claims and on the predictive performance with explicit information cut-offs as well as the preservation of stored forecasts in communication. The intended users were stakeholders in the public-finance system, but the study measured technical abilities in decision support not improvements in the actual decisions made or the participation of the users.

1.4 Research Objectives

1.4.1 General Objective

To develop, deploy and test a neuro-symbolic probabilistic agentic AI decision support system for the budget process in Kenya.

1.4.2 Specific Objectives

1 To build a neuro-symbolic representation framework for modelling Kenyan budget data.
2 To build a Bayesian probabilistic reasoning framework to model uncertainty and forecasting of budget trends in the Kenyan budget process.
3 To test the platform for data for decision making on Kenyan budgets

1.5 Research Questions

2 How might this neuro-symbolic representation be used to enhance the modeling of the Kenyan budget data?
3 What role can the Bayesian probabilistic reasoning play in modelling uncertainty and predicting budget trends in the Kenyan budget process?
4 Effectiveness of the platform for enhancing data-driven Kenyan budget decision support.

Modelling Kenyan budget data for evaluation involved maintaining source-linked assertions and accounting distinctions, not necessarily factual correctness and institutional continuity throughout the graph. Forecasting budget trends was operationalized as forecasting the annual execution ratios and their uncertainty rather than forecasting all revenue, expenditure or macroeconomic performance. The effectiveness of the platforms was evaluated based on technical evidence in terms of retrieval, preservation of the forecast, predictive performance, and eligibility controls. The research questions thus focus on capabilities shown within these limits: improved participation, improved trust, improved comprehension and improved decision quality were not directly established.

4.5 Significance of the Study

The significance of the study is the design and evaluation evidence for the use of AI in context of budget analysis. Technically, it connects documentary claims to their sources, evaluates forecasts of execution, and differentiates between the output of models and explanations generated by them. The benefits for the intended users are easier verification of evidence and clearer communication, but not necessarily more participation, trust or decision quality. The

significance below identifies the wider potential relevance from the contributions stated in Section 5.5.

A bounded architecture is essential because design significance involves the specification of interfaces between data-driven extraction, symbolic representation and separate Bayesian forecasting. The study does not represent a new general theory of neuro-symbolic AI but rather provides context-specific evidence of such interfaces; the availability of provenance and predictive distributions (Garcez & Lamb, 2023) is not directly related to improved explainability for the user.

Practical implications include the ability to enhance budget transparency by attaching claims to the underpinning documents and displaying evaluated forecasts with declared uncertainty and reported restrictions. While the platform tries to solve an identified challenge of analysing fragmented budget information, its impact on public participation and decision quality will need to be evaluated by prospective users (Martin, 2024; Fenuku, 2024; Bajeti Hub, 2024).

Institutional significance lies in the platform's potential to support trustworthy communication by linking outputs to source evidence and reporting predictive uncertainty and system limitations. Because misinformation and disinformation can erode public trust (Ojino et al., 2024), these design features may support more transparent communication by public institutions; however, their effect on trust was not evaluated in this study (OECD, 2025).

Policy significance in supporting the Kenya digital masterplan: this study aligns with the Kenya National Digital Masterplan (2022 – 2032) which lay emphasis on using digital technologies to enhance government services delivery.

4.6 Scope and Delimitations

The study focused geographically and administratively on budget data from the Republic of Kenya. It used publicly available data from the National Treasury, Parliament, the Office of the Controller of Budget and the Office of the Auditor-General. The time-series evidence covered the budget cycle, including planning, formulation, processing, approval, implementation and oversight. The study was therefore delimited to published national-government budget data and did not extend to granular records generated by ministries, departments and agencies or to county-government budgets.

Language – The Constitution of Kenya has already acknowledged cultural diversity and multi-lingualism, although the main language of this platform is English, other indigenous and Swahili languages are important but were not included in this study.

Budget implementation tracking – In this study real-time tracking of budget implementation through the Integrated Financial Management Information System (IFMIS) is not included. Rather, it can only be used for public involvement throughout the budget cycle through officially published budget documents and not to forensic audit spending already made.

Hardware constraints – the proposed platform does not work best in low bandwidth or low resources environments. Thus, the assumption is that intended users have access to basic devices that have Internet and that text-based interaction features can enhance the level of accessibility in regions where connectivity is limited.

Legal implications – The platform will only be a participatory and advisory tool, and will not have legal power to overrule legislative decision-making. It can only be used to give a framework for public feedback, which can assist the analysist, public and ultimately the parliamentarians in an informed decision-making process relating to budget.

1.8 Assumptions of the Study

The study presumed that the official documents collected from the National Treasury, Parliament, the Controller of Budget and the Office of the Auditor-General was a true reflection of the published budget record at the time of the collection. It also took it as a given that the dates of documents relied on in the forecast-availability rules constituted a reasonable basis for determining the information available at each forecast origin. The monetary values, the time periods and the accounting areas were evaluated based on the source tables and labels. Multiple labels or crosswalks that had been reviewed were not considered definitive institutional identifiers but were used as provisional continuity indicators where there was no stable cross-year identifier. The study also assumed that intended users will have access to the platform via an internet connected device and that it will be used as an advisory tool and not a replacement for statute-based judgment.

1.9 Limitations of the Study

The study was constrained by a number of limitations that delineate the interpretation and generalisation of its findings. First, it was based on publicly available national government documents, namely from the National Treasury and Parliament, as well as from the Controller

of Budget and the Office of the Auditor-General. As a result, the evidence base did not include non-public administrative data, real-time IFMIS transactions, granular data in ministries, departments and agencies, and county-government budgets. The varying timing of publications, publication formats, table organization, and revisions and institutional identifiers also resulted in some inconsistencies in publication coverage by fiscal year, and source.

Second, part of the knowledge graph was written in an automatic manner and detailed source-cell review was focused on specific parliamentary content and on smaller validation sets. The checks found conformity in the sampled assertions, but this did not prove complete source correctness for each graph record. Errors in extraction, attribution or entity-linking may be introduced through the use of optical character recognition and complex PDF layouts, as well as ambiguous units and changing administrative labels. Additionally, longitudinal modelling was restricted as there were many regularly occurring entities that were not known to have changed their names or organisational structures through mergers or splits. Therefore, graph-wide forecasting was mostly conducted at the vote/agency level, not at the level of fully harmonised programmes or budget-lines.

Thirdly, there were various denominators and temporal designs for the forecasting experiments. Some operational analysis used original-gross-estimate proxy while the historical graph study and FY2025/26 benchmark used revised-gross-estimates. Thus, their findings address somewhat different and related questions and cannot be used interchangeably. Tabling and repository dates may not identify the first public availability of a particular document version, which limits the certainty of reconstructed historical information sets. The main held-out benchmark included one fiscal year, 25 MDAs and 50 recurrent and development questions. Repeated language-model generations assessed response variability but did not increase the number of independent forecast targets. There were 47 outcomes that represented under-execution, with only one exact-execution outcome and two over-execution outcomes. The class imbalance prevented reliable estimation of the probabilities of the rare states. Furthermore, the nominal 80% Bayesian predictive intervals achieved only 64% coverage in the held-out year, despite lower point-forecast error than the tested language-model alternatives. Run-specific convergence diagnostics for the final FY2025/26 Bayesian fit confirmed chain convergence and mixing across continuous components (maximum rank-normalized $\hat{R} \leq 1.0015$; bulk ESS $\geq 1{,}674$; tail ESS $\geq 3{,}177$), establishing that the observed interval under-coverage (64%

coverage for nominal 80% intervals) reflects empirical predictive calibration rather than MCMC numerical non-convergence.

Lastly, the platform was assessed for the most part, technically, predictively, and using evidence grounded approaches, not by means of a large-scale prospective deployment with parliamentarians and citizens. Hence, the study does not make any claims to long-term impacts on public involvement, institutional confidence or the statutory budget process. Limited accessibility and operational use may also be a result of the English language interface and the lack of real-time execution data, which relies on internet-enabled devices. These limitations qualify the strength, scope and generalisability of the conclusions. The findings support source-linked, advisory decision support within the evaluated documents, samples, forecast origins and model configurations, rather than universal forecasting superiority or demonstrated improvements in public participation and institutional decision-making.

1.10 Operational Definition of Key Terms

Table 1: Operational Definition of Key Terms

Term	Operational Definition
Neuro-symbolic AI	In this study, neuro-symbolic AI refers to the integration of data-driven document extraction and semantic alignment with symbolic ontology, knowledge-graph and validation components. The neural or statistical components identify candidate entities, relations and financial assertions from heterogeneous documents, while the symbolic components preserve accounting meaning, provenance, constraints and queryable relationships.
Bayesian reasoning	This is a probabilistic inference method that updates predictions when new evidence becomes available.
Agentic AI	In this study, agentic AI refers to a bounded orchestration system that selects approved retrieval or forecasting tools, validates required inputs and communicates source-linked outputs to users. It does not autonomously alter budget data, invent missing observations or generate independent numerical forecasts.

1.11 Organisation of the Study

This thesis has five chapters. The study's background, statement of the problem, purpose, research objectives and questions, significance, scope and delimitations, assumptions, limitations and operational definitions for key terms are presented in Chapter One. It sets the stage for the need to come up with an integrated, traceable and uncertainty-aware platform to support the stakeholders in the Kenyan budget process.

Chapter Two reviews the theoretical and empirical literature on neuro-symbolic artificial intelligence, knowledge graphs, Bayesian probabilistic reasoning, agentic AI and civic technology. It presents the conceptual framework, examines the relationships among the study variables, critiques existing approaches and identifies the research gaps addressed by the study.

Chapter Three describes the research philosophy, Design Science Research approach, data sources, system architecture and procedures used to develop the platform. It explains the construction and validation of the source-centred knowledge graph, the Bayesian forecasting models, the agentic orchestration layer, the protected evaluation design and the metrics used to assess extraction quality, predictive performance, uncertainty, retrieval and end-to-end system behaviour.

Chapter Four presents and discusses the findings. It reports the knowledge-graph coverage and source corrections, historical and operational Bayesian forecasting results, and the corrected FY2025/26 held-out evaluation of the platform. The chapter compares Bayesian outputs with baseline and language-model conditions, examines recurrent and development expenditure separately, and discusses point accuracy, interval calibration, class probabilities, provenance and the implications of the results.

Chapter Five summarises the findings for each research objective, draws conclusions in relation to the research problem and purpose, and provides recommendations for policy and further research. It also states the study’s original contribution to knowledge in probabilistic neuro-symbolic representation, uncertainty-aware forecasting, bounded agentic AI and reproducible evaluation for public-sector decision support.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter discussed the literature related to the theoretical and empirical concepts. It focused on the literature associated with decision support systems with the use of artificial intelligence in participatory governance and highlighted the research gaps addressed by this study.

2.2 Theoretical Framework

Growing digitization of public administration has increased availability of government data. national budget, audit reports, budget statements and budget implementations reports are readily available on web portals of government agencies. But then, these budget data and information remain scattered in disparate platforms and are highly technical which makes it difficult to understand for stakeholders such as citizens and even parliamentarians.

Research on artificial intelligence and semantic technologies offers several responses to fragmented public data, including linked-data publication, visual analytics, predictive models and tool-using interfaces. These approaches should be assessed against their intended tasks rather than dismissed collectively as static. This review examines how neuro-symbolic representation, Bayesian inference and bounded agentic orchestration address distinct requirements, and where evidence for their integrated use remains incomplete.

Neuro-Symbolic Artificial Intelligence Theory. This theory integrates neural network technology with symbolic reasoning to support data driven learning and logical inference. This helps to achieve explainable AI in governance where transparency and accountability are needed (Garcez et al., 2022). This theory integrates structured knowledge representation systems and data-driven learning models to enhance interpretability, explainability and logical reasoning as outlined by (Garcez & lamb, 2023; dÁvila Garcez et al., 2021). It is relevant to budget decision support within the wider public finance management in Kenya due to the need to ensure transparent and traceable public policy decisions as required by the constitution and the laws.

Bayesian Inference Theory. Bayesian inference provides a systematic framework for combining prior information with observed evidence to produce posterior distributions and predictive uncertainty (Murphy, 2022). This is relevant to public-finance forecasting because budget execution data may be sparse, revised or delayed. In this study, Bayesian models were

used to estimate execution-ratio distributions and probabilities of under-, exact- and over-execution; their predictive value was then assessed against held-out outcomes and simple baselines.

Theory of Agentic Artificial Intelligence. The concept of agentic AI is inspired by the fields of autonomous agents, planning, and multi-agent systems. In this study, the agentic component was not trained via reinforcement learning but rather through bounded orchestration, where the agent chooses the appropriate retrieval and forecasting tools that were approved, validates the inputs and distributes the outputs with source links (Wooldridge, 2021; Wang et al., 2023). This framing is applicable to the Kenyan budget process because it encourages action in relation to goals, has auditability and restricts autonomy.

The theory behind integration is not that integration of advanced techniques will necessarily enhance decisions, but functional complementarity. Symbolic representation is used to retain meaning and origin of claims; Bayesian modelling presents uncertainty based on the observations that are considered valid; bounded orchestration is used to manage the access and communication of evidence and forecast. Each of these functions are for different failure modes and can be tested independently of each other. The question of whether their sum total outweighs the added complexity is one of fact, not theory.

This framework relies on the assumption that efficient decision support in participatory budget processing is largely influenced by timely availability of high quality budget data, robust knowledge representation of the budget data from both structured and unstructured and disparate sources and formats to support analysis, and the successful application of the developed neuro-symbolic and agentic AI reasoning to create a platform that would offer a unified interactive and intelligent system that can offer interpretability, explainability and predictive insights to support multi-stakeholder audiences in budget process decision-making in the Kenyan context.

2.3 Conceptual Framework

2.3.1 Overview

The conceptual framework underpinning this study integrated neuro-symbolic representation, a separate Bayesian probabilistic reasoning component and bounded agentic AI. The neuro-symbolic component combined data-driven extraction and semantic alignment with symbolic ontology, knowledge-graph and validation functions. Bayesian forecasting consumed eligible

graph-derived observations and returned predictive distributions, while the agentic interface coordinated approved retrieval and forecasting tools and communicated source-linked outputs.

The integrated approach was selected due to the need for both a documentary basis of an amount and an assessment of future execution for a situation in which a budget question might arise. The source-centred layer keeps the information about the institution, fiscal period, measure, revision and source location required for defining a comparable observation. The forecast-date and denominator rules are then applied using the graph-to-model interface prior to statistical estimation. Finally, the bounded interface returns the stored forecast along with the evidence and limitations. Integration is important on these interfaces because it allows a relationship to be inspected between a reported claim, an eligible model input and a communicated forecast, not just three technologies to be placed within the same application.
Modular neuro-symbolic pipeline with a separate probabilistic model, rather than an end-to-end jointly trained neural-symbolic Bayesian system. Where there is equivalent provenance fields could be maintained in a relational database, or where the point accuracy is similar, a simpler model could be kept, and a deterministic interface could be used for fixed queries. The graph is chosen for explicit cross-document relationships, revision links, and the agentic interface for bounded question routing; the probabilistic component for conditional distributions. These selections come at a cost of ontology maintenance, source-review, model-monitoring and orchestration. They have the best justification for tasks that involve all three functions, and less justification for simple look-up (or point forecasting) tasks; net benefit still requires comparative evidence of deployment.

2.3.2 Conceptual Framework Diagram

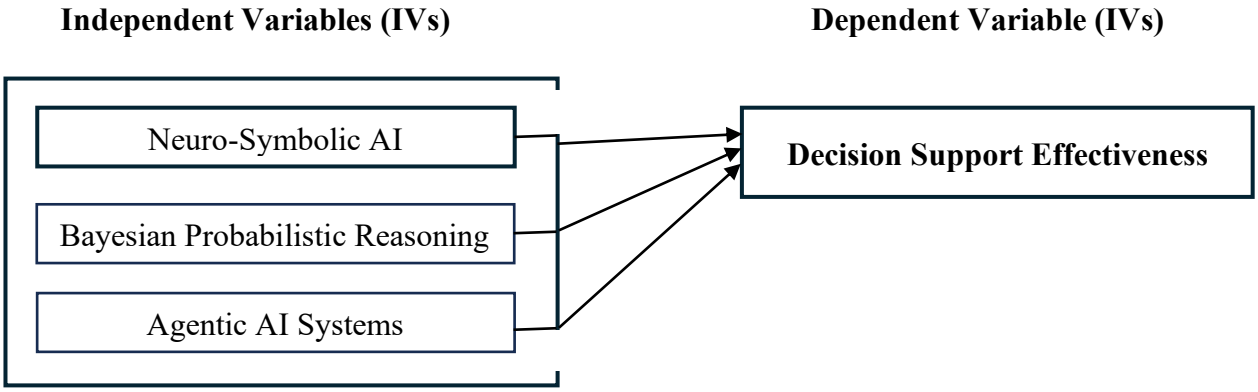


Figure 2.1: Conceptual Framework

The relationship between neuro-symbolic artificial intelligence, Bayesian probabilistic reasoning and agentic AI systems is shown in an independent variable in Figure 2.1. Together, these components affect the effectiveness of decision support in the Kenyan budget process, through the improvement of interpretability, uncertainty modelling and autonomous reasoning capabilities (Garcez & Lamb, 2023; Murphy, 2022; Wooldridge, 2021). This framework was used to design and assess the system by converting the neuro-symbolic reasoning and Bayesian inference and agentic system into measurable variables and indicators.

2.3.3 variables and Indicators

The three main types of variables are the independent variables which convey the main technological components that drive the system and they comprises of neuro-symbolic AI, Bayesian probabilistic reasoning and agentic AI. The dependent variable is the effectiveness of the decision support.

(a) Independent Variables (IVs).

Neuro-Symbolic AI representation combines neural network learning models for extraction and semantic alignment to convert source text into entities and relations, with symbolic knowledge structures such as the knowledge graphs to make the system to discover hidden patterns in budget data while maintaining logical consistency and interpretability (Garcez & Lamb, 2023). It check required fields, canonical identifiers and provenance prior to accepted assertions entering the knowledge graph, which in turn supports cross-document integration, source traceability and relational queries.

Bayesian Probability Reasoning and inference (Murphy, 2022) represents the application of Bayesian inference techniques to model uncertainty and predict budget trends. This technique improves the decision-making process by using prior budget information and updating the beliefs when new data becomes available.

Agentic AI capabilities define the development of autonomous agents that can reason, plan, interact with data sources and perform goal-driven decision-making as they facilitate dynamic analysis and user interaction in the decision support system (Wooldridge, 2021; Zheng et al., 2023). The agentic orchestration coordinates retrieval and approved analytical tools to respond to bounded user questions.

This framework has provided several mechanisms as opposed to a single composite effect where entity links may improve cross-document retrieval, symbolic constraints may reduce structurally invalid assertions, provenance my improve source verification, posterior prediction provide a formal representation of uncertainty and agentic tool selection may reduce the effort needed for multi-step tasks. Thus, each mechanism was evaluated separately because errors in data extraction, entity linking, validation and model assumptions may potentially propagate through the architecture.

(b) Dependent Variable (DV)

This represents the outcome of the framework in view of the budget decision support in Kenya. This constitutes the ability of the system to provide accurate and interpretable budget information and actionable insights to the stakeholders. Also, it encompasses transparency and accountability, accuracy in the budget trends predictions, and overall effectiveness of the proposed prototype in supporting the decision-making process.

2.3.4 Relationship among the variables

The relationships in the conceptual framework are proposed mechanisms, not established causal effects. Source-linked representation is expected to support evidence verification; probabilistic modelling supplies forecast distributions; and bounded orchestration is intended to preserve analytical outputs during interaction. Data quality, institutional identity, source timing, model assumptions and interface behaviour can constrain each mechanism. The study therefore evaluates these properties separately rather than treating integration as a single measured cause of better decisions.

This framework combined state-of-the-art artificial-intelligence techniques to process budget data, create source-linked claims and produce uncertainty aware forecasts. The neuro-symbolic pipeline was responsible for extracting and aligning entities and financial assertions, the knowledge graph for preserving the semantic relationships and provenance, and the Bayesian component for generating predictive distributions, while the bounded agentic interface for coordinating the retrieval and communicating the validated outputs. The architecture was designed to overcome the drawbacks of static reports and dashboards, but its impact on the quality of decisions made by user stakeholders needs to be evaluated by the users.

2.4 Empirical Literature Review

This section presents literature reviews on the core domains of the proposed system: neuro-symbolic AI, Bayesian probabilistic reasoning and agentic artificial intelligence and how they relate to civic technology which anchors the budget decision support. It examines previous studies and their limitations in addressing the problem at hand that study seeks to address.

2.4.1 Neuro-Symbolic AI for Governance data

Neuro-symbolic artificial intelligence addresses the limitations of data-driven machine learning modeling. It integrates neural network learning mechanisms with symbolic knowledge representation, which creates systems that can learn patterns and explaining their reasoning based on logic (Garcez et all., 2022). Knowledge graph presents the semantic connection between entities in the budget discourse such as spending agencies and their various programs in a manner that allows integration of heterogeneous budget data (Hogan et al., 2021) such as budget allocations to programs in ministries and departments, expenditure needs, and competing priorities.

Knowledge graphs have been applied in governance and public administration scenarios to represent institutional structures, legislative decisions and administrative processes. These representations improve transparency by enabling users to trace relationships between organizational sections, policy decision and flow of budgets (Jiang et al., 2022). In public finance where budget process is contained, knowledge graphs facilitate semantic linkages between budget allocation, spending agencies and governments programs in a semantic manner.

The reviewed governance applications emphasise semantic representation, interoperability and evidence access. These are substantive contributions, not merely substitutes for static reports. However, their suitability for execution forecasting cannot be inferred from representation quality alone. The relevant question is whether a particular implementation also models future outcomes, evaluates predictive uncertainty and supports users in interpreting the evidence. These capabilities require separate assessment.

Accordingly, the present study uses neuro-symbolic representation to organise and validate source-linked claims, while a separate Bayesian component models execution outcomes. This division is a task-specific architectural choice rather than a general limitation of neuro-symbolic AI. Source traceability can support an explanation of where a value came from, but

does not by itself establish causal explanation, factual correctness or improved understanding by non-expert users (Hogan et al., 2021).

2.4.2 Bayesian Probabilistic Reasoning

Bayesian inference combines prior information with observed evidence to obtain posterior and predictive distributions (Murphy, 2022). It is relevant to budget forecasting because uncertainty about model parameters and future execution can be stated explicitly. Nevertheless, delayed reporting, revisions and incomplete records require explicit modelling or eligibility rules; the use of Bayes' theorem alone does not resolve them. Predictive reliability depends on the likelihood, priors, target definition and available data, and must be assessed against suitable comparators.

Bayesian inference has been combined with knowledge-graph evidence to support probabilistic reasoning (Zhang et al., 2023). However, these models are often evaluated separately from source-centred semantic structures and interfaces intended for non-technical users. In this study, Bayesian forecasting was implemented as a separate probabilistic reasoning component that consumed eligible observations derived from a versioned knowledge-graph snapshot and returned predictive distributions linked to their supporting evidence.

2.4.3 Agentic AI System

This system represents the move from passive and static computational models to intelligent, dynamic and autonomous systems which can reason and act in a complex environment as argued by (Wang et al., 2023). Agentic AI presents an opportunity where the emerging technologies can be used to create systems that are interactive, based on natural language, and easy to use by stakeholders with varying technical capabilities and budget informational requirements.

The application of agentic AI to public finance in general and budget analysis in particular, remains less developed. Most uses are seen in general purpose rather than domain-specific use case such as budget data analysis. Also, agentic systems are often delinked from knowledge representations which affects their ability to offer traceable reasoning.

The agentic AI component served as a bounded interface between users, the knowledge graph and approved forecasting tools. It selected permitted retrieval or forecast operations, validated

required inputs and communicated source-linked outputs. It did not autonomously alter budget data, invent missing observations or generate independent numerical forecasts.

2.4.4 Artificial Intelligence for Civic Technology

The goal of civic technology endeavours is to enhance democracy by using digital technologies to facilitate citizens for participatory engagements. In the Kenyan budget process use case, such initiatives usually endeavours to provide open data portals, to promote transparency and accountability (Mergel et al., 2021). But then, these initiatives often have limited analytical depth and lack dynamism and interactivity.

Open-data portals and visualisations remain useful for publishing and inspecting budget information. Semantic budget systems can additionally provide structured, machine-readable access, as illustrated by Cifuentes-Silva et al. (2020). The question motivating this study is whether a source-linked analytical workflow can also provide evaluated forecasts and communicate their uncertainty without obscuring documentary meaning. The prototype addresses that technical question; suitability for users with different skills and informational needs remains a user-evaluation question.

2.5 Critique of Existing Literature

The literature can be compared to four different requirements: maintaining accounting meaning and documentary provenance, predicting under uncertainty, communicating results, and controlling system behavior. Achievement of one requirement does not imply achievement of the other(s). A searchable or visually accessible dataset can be used for evidence inspection without predicting future executions; a predictive model can be used to give low error without source context; and an interactive interface can be used to communicate a wrong result fluently. What is important then is which combinations will satisfy the requirements for the intended task, how, and at what extra expense?

Chilean budget legislation and execution are directly relevant precedents represented by a national budget ontology, linked open data and visualisation (Cifuentes-Silva et al., 2020). Their belief contradicts the notion that current civic technologies are just static databases: semantic publication can already be used to enable human and machine access. Serderidis et al. (2024) also show how to reuse ontologies and represent government decisions in an interoperable way with d2kg. These contributions lay the groundwork for semantic integration; however, the representation and access that they report are not themselves sufficient to produce

calibrated forecasts of the execution ratio at a protected forecast origin. The present research follows this trend by connecting source-defined budget claims to separately assessed forecasting and bounded communication but not to suggest that it is beginning semantic budget integration.

It is important to note that knowledge graphs and neuro-symbolic methods are not inherently unable to deal with uncertainty. Deductive and inductive techniques, identity management, contextual representation and quality assessment (Hogan et al., 2021) are part of the broader literature. The key difference with this distinction is between uncertainty as to an extracted claim or an entity match and uncertainty as to a future expenditure outcome. A graph may be able to support conflicting claims and confidence metadata without necessarily supporting the predictive distribution needed for the latter task. On the other hand, a Bayesian execution model based on accepted graph observations does not necessarily carry over OCR or attribution/entity linking mistakes. This is because the current framework distinguishes between these concerns by means of validation and provenance control, and does not model all sources of uncertainty together.

While Bayesian modelling is the only approach that has an explicit notion of conditional predictive uncertainty, this is not unique to Bayesian approaches and does not imply better forecasts. Simple statistical baselines are still valid for short histories and other probabilistic approaches may also provide prediction intervals. The selection in this study is based on the fact that execution states and continuous outcomes can be specified independently and that it makes it possible to examine the model assumptions. However, its usefulness should only be assessed using same-cutoff comparisons, sensitivity analysis, run-specific convergence diagnostics and held out calibration and not by the label Bayesian alone.

Retrieval and agentic tool use solve another problem: coordination of an analytical response and accessing evidence. They cannot verify the amount retrieved or whether or not a numerical forecast is reliable. The risk-management paradigm of Autio et al. (2024) is suitable for considering the trustworthiness during design, use and evaluation; it does not guarantee that a specific interface is safe and effective. As such, the current architecture only allows for authorized operations and communication of outputs stored. The same evidence and forecasts could be presented in a deterministic form or dashboard with less orchestration complexity. The agentic interface is thus a design decision for flexible question handling, but not a requirement for trustworthy forecasting which has not yet been tested for incremental usability.

2.6 Research Gaps

The literature examined provides the bases for semantic publication of budgets, AI-assisted classification and probabilistic reasoning. Cifuentes-Silva et al. (2020) present a linked-data representation of budget legislation and execution, as well as Guariso et al. (2023) who present context-sensitive automated budget tagging. With these contributions, neither the application of AI to budgets nor the building of a budget knowledge graph forms the gap on its own. The gap is with respect to evidence of interfaces between source-grounded representation, forecast-date-constrained statistical inference and controlled natural-language communication in the Kenyan context. It is not a claim that there is no comparable system anywhere – it is an integration/evaluation gap in the literature reviewed.

Representation gap: Objective 1. In the literature examined, there is no clear answer on how to structure the heterogeneous nature of the Kenyan budget documents into a set of assertions that can be queried while keeping the differences between original estimates and revised estimates, as well as recommendations and expenditure. To fill this gap, evidence related to structural conformity and documentary correctness is required, and not just the existence or size of a knowledge graph. Typed, versioned claims, page- or table-level provenance, and structural validation and targeted source review are studied. It is evidential organisation, full source correctness and fully harmonised longitudinal identity is not achieved.

Predictive-evidence gap: Objective 2. Semantic integration does not provide any information on whether observations are eligible for competitive forecasting or whether they provide reliable uncertainty estimates. The question is how do forecasts of execution-ratio values work when there are clear denominator definitions, accounting scope and documentation available, and whether the point accuracy and the interval calibration/class-probability quality are consistent. The study tackles this by rolling-origin evaluation against simple statistical baselines and a separately protected held-out benchmark, instead of relying on the existence of a Bayesian distribution as a sign of calibration.

Controlled-integration gap: Objective 3. The retrieval of evidence, prediction and faithful communication is not the same. The integration question remains, whether a bounded AI interface can link up eligible documentary claims to stored forecasts without altering their values and limitations. Retrieval and no-retrieval language-model conditions and direct Bayesian outputs and forecast-preservation condition are compared. This partly covers technical decision support, but leaves comparative usability, institutional impact and the

incremental value of agentic orchestration compared to a deterministic interface out of scope of what has been demonstrated.

2.7 Summary of Literature Review

The literature demonstrates that semantic technologies can be used to integrate public data and enable access, and that probabilistic models and tool-using interfaces can provide different analytical and interaction needs. It does not excuse to take all the present systems as given, or assume all knowledge graphs can be assumed to have the same capabilities, or that all the Bayesian outputs are calibrated. The result is a research rationale which is a controlled task-specific integration with a separate evaluation of the source quality, of the forecasts and of the interface behaviour.

Bayesian probabilistic reasoning is a way of dealing with uncertainty which very often exists in budget data. Previous works have integrated knowledge graphs and Bayesian models, however, if only the output of the models is provided, the results might still not be understandable for non-technical users. In this work, Bayesian forecasting was connected to the facts obtained from the knowledge graph, and the agentic interface issued the predictions, uncertainty and provenance.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the methodological approach adopted to design, implement and evaluate a probabilistic neuro-symbolic agentic artificial intelligence system for decision support in the Kenyan budget process. This system is a socio-technical solution since it that applied the emerging technologies to provide solution to real-world problems in participatory budget process (Floridi et al., 2022) to enhance interpretability of complex budget data, and handle uncertainties that are common in budget data. It also provides for the evaluation of the system (Drechsler & Hevner, 2022) on the effectiveness of AI-driven solution in budgeting.

3.2 Research Philosophy

The study was guided by a pragmatic research philosophy because it sought to solve an existing problem by developing and evaluating a system for budget analysis and evidence-based decision support. Quantitative methods assessed extraction and graph quality, computational convergence, predictive accuracy, interval calibration and comparative performance, while qualitative source review supported the interpretation and correction of extracted assertions and provenance. The philosophy therefore supported the combination of methods needed to evaluate the platform's technical performance, traceability and evidential reliability.

3.3 Research Design and Approach

To design, build and test the proposed intelligent system, the study applied Design Science Research, a research method used to address a socio-technical problem (Drechsler & Hevner, 2022). Extraction and graph quality, computational convergence, point-forecast accuracy, interval calibration, class-probability quality and comparative performance were assessed using quantitative methods. Extracted assertions, provenance and table-cell mappings were interpreted and corrected using the qualitative source review. All these methods were used jointly to assess the technical performance of the platform, its traceability and evidential reliability.

Assessment was done in a multidimensional way, rather than aggregating into a single composite score. The quality of knowledge-graph was measured using source review and provenance and structural conformance; the quality of Bayesian performance was measured using convergence diagnostics, point error, interval quality and class-probability scores, while the quality of platform behaviour was measured using retrieval grounding, reproduction of

frozen forecasts and explicit abstention. These measures were reported separately as they are different properties of the system and should not be added together in any arbitrary way.

3.4 System Architecture

The system was implemented in Python because its ecosystem supported transparent and reproducible integration across the platform's components. A layered and modular architecture separated documentary ingestion, knowledge representation, probabilistic modelling, agentic orchestration and decision-support outputs, thereby improving maintainability, traceability and control over the information passed between components (Garcez et al., 2022; Wooldridge, 2021).

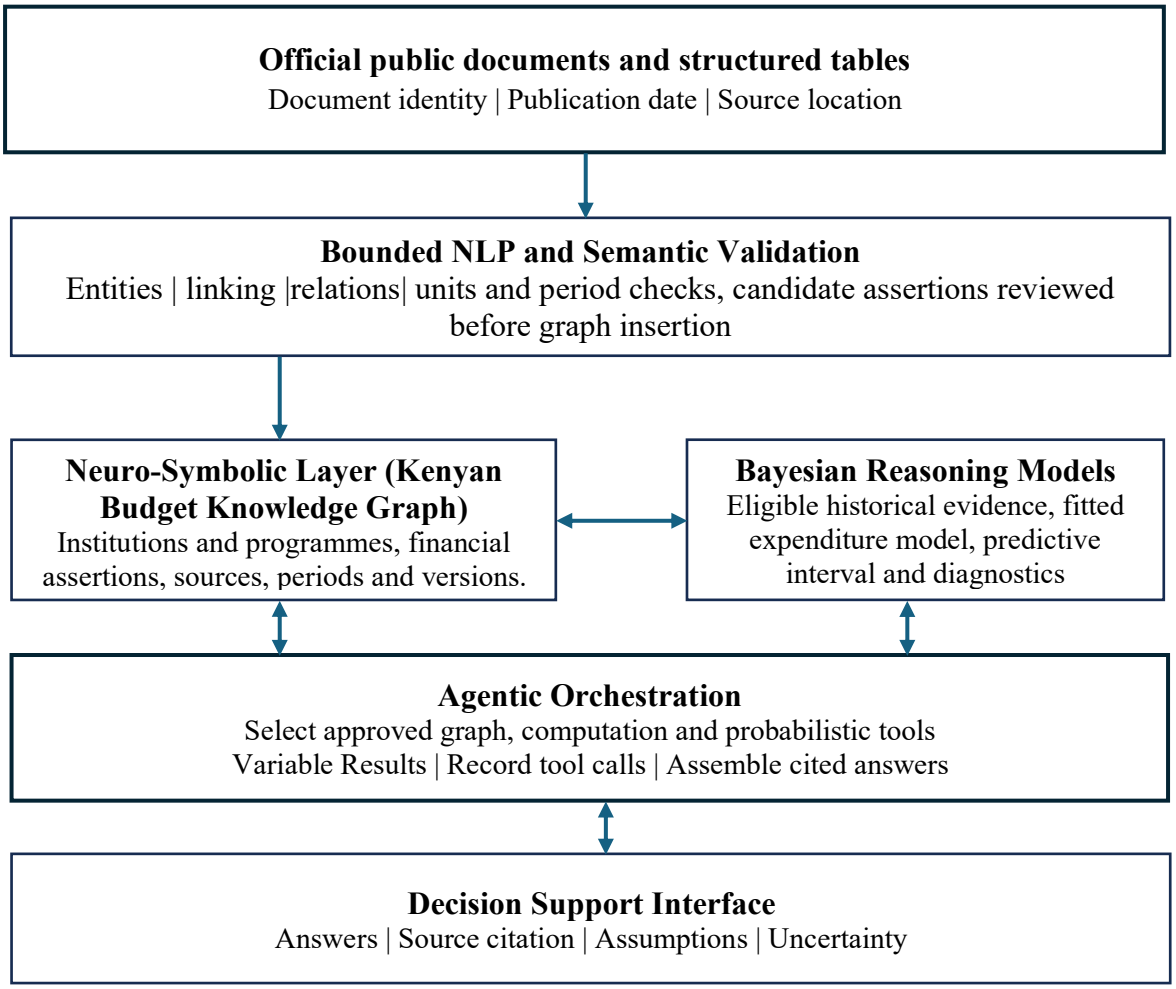


Figure 3.1: System Architecture

This architecture consists of the input layer which comprises of data ingestion, intelligent layer which comprises of the neuro-symbolic AI where the knowledge graph feeds to Bayesian reasoning by probabilistic inference and agentic AI´s autonomous agents complete the

orchestrates the analysis and reasoning. The output layer is for decision support which comprises of interpretable budget insights, budget predictive analytics and transparent decisions. Each layer performs a transformation that can be represented by:

$$X_{l+1} = f_l(X_l, W_l)$$

Where;

X_{l+1} = transformation to the next layer,

X_l = data representation at layer l,

W_l = parameters across various highlighted layers, and

f_l = the transformation function.

Therefore, the end-to-end flow of the system reasoning can be summed as:

$$Rf = f_{UI}(f_{agent}(f_{Bayes}(f_{ns}(f_{data}(X)))))$$

Where;

Rf = the overall reasoning flow of the system,

f_{UI} = the interactive user interface function,

f_{agent} = the agentic AI function component,

f_{Bayes}= the Bayesian inference function,

f_{ns}= the neuro-symbolic function component,

f_{data}(X) = is the data representation function

This is a pipeline that provides for traceability of reasoning according to (Floridi, et al., 2022; Paulheim, 2023).

3.4.1 Component Roles and Information Flow

Structured and unstructured documents were published to an ingestion layer where they were stored with a record of the document identity, date and location of origin. Using neural natural-language processing and table extraction, candidate entities and financial assertions were created. Semantic alignment identified candidate mentions and matched them with the canonical entities and SHACL and task-specific rules verified required fields, period consistency, measure type, and provenance prior to graph insertion.

The neural part was responsible for the extraction from documents, for the resolution of multiple references to the same budget entities in the documents and for the interpretation of user questions. The symbolic part contributed the formal structure of consistent entities in the

form of ontology and the canonical names, and the relationships and validation rules in the knowledge graph. The Bayesian model was then run on a valid dataset of observations from the versioned graph snapshot to generate predictive distributions and diagnostics. The agentic layer orchestrated access to the knowledge graph; it was used for selecting the proper statistical tools based on the task and presented the result with the evidence and uncertainty information for interpretation.

3.4.2 Interface Between the Components

A document identifier, location of a page or table, entity identifier, fiscal period, measure type, numeric value, accuracy, scale, and the revision status/extraction confidence were passed from the ingestion-to-knowledge-graph interface. The graph-to-model interface also provided a versioned observation table, the selection query, and the as-of date. Checks on eligibility were carried out, and observations published after the forecast origin eliminated and values for each year separated from cumulative quarterly values.

The model provided diagnostic information and uncertainty ranges in addition to forecasts. These outputs were segregated from facts observed and correlated to evidence presented to arrive at them. The agent was used to make read-only graph queries, provide documentation of arithmetic functions and approve model endpoints. It didn't produce nonexistent observations or free-form language to generate probabilities. Forecasts were precomputed or provided from a fitted versioned model and a changed evidence snapshot caused controlled refitting instead of model fitting with every user interaction.

3.5 Data Source and Design

In this study, four official families for budget-document were merged. Much of the estimate and programme information is provided from National Treasury publications. Where applicable, Parliamentary records gave legal context and implementation/expenses evidence, which was complemented by Auditor-General reports. The held-out evaluation results were provided in a later Controller of Budget report. Each analytical record maintained the institution, version of the document, fiscal period, measure and source location to keep the values from various stages of the budget process distinguishable and interpretable.

Initially, source inventory was handled in a versioned manifest, where a single document record was added for each document and a stable document ID was maintained. The manifest includes the following information: source institution and role, fiscal year, partition, archive path,

extracted path, hashes, URL and date metadata, and revision status and preliminary eligibility. These fields contain source discovery and provenance information; the specific location of the manifest is included in Appendix A.

There are currently 370 document records in the source inventory: 320 development records; 36 evaluation records and 14 records outside the scope of the analysis. It covers the years from FY2016/17 to FY2026/27 and includes 262 National Treasury records, 80 CoB records, 26 Auditor-General records and two Parliament records.

The primary source for budget estimates and the structure of government expenditure was from National Treasury publications. They were used on their tables to distinguish allocations at various administrative stages, including votes and programmes, as well as between budget stages (e.g. original estimates, revised estimates). Since financial information can be found in various types of tables in the same document, every amount was interpreted based on the table context to which it belonged, in terms of fiscal year and budget stage, before being added to the dataset.

Parliamentary records include appropriation legislation and scrutiny documents. They distinguish enacted amounts from estimates, proposals and committee recommendations. Records are linked to Treasury and CoB evidence when vote, programme, period and amount can be reconciled. Budget and Appropriations Committee recommendations retain recommendation status, while enacted appropriations remain a separate measure. Original estimates, supplementary proposals, enacted supplementary amounts and revised estimates are represented as separate assertions when the sources distinguish them.

Each institution contributes the claims supported by its documents: Treasury records describe estimates and budget structure, Parliament records legislative status and scrutiny, and CoB records implementation and reported expenditure. Cross-source comparisons connect those assertions while preserving the values associated with each budget stage.

The CoB is the primary source for reported budget implementation and expenditure outcomes. Its National Government Budget Implementation Review Reports appear at quarterly, half-year, nine-month and annual intervals. Each reporting interval remains a separate period type, so cumulative in-year observations and annual outturns enter analysis under their own period labels. The report and table establish whether an observation is annual or cumulative.

CoB tables distinguish original gross estimates, revised gross estimates, supplementary revisions, actual expenditure and execution percentages. Each value retains its source label and table context downstream. For an execution-ratio forecast, the denominator is the budget measure specified for that analysis and available at the forecast origin. The FY2025/26 benchmark uses unrounded actual expenditure divided by the same-scope revised gross estimate. Other development experiments use an original-gross-estimate proxy, so their ratios answer a different forecasting question and are reported separately.

Auditor-General reports provide audit and assurance context for Government entities and financial years. They are involved in this study through source tracing and contextual evidence. Audit claims continue to be associated with the report/passage they support, expenditure targets from CoB and budget-authority measures from Treasury or Parliament.

The development source base spans from FY2016/17 through to FY2024/25 and is not evenly spread between institutions and across document types. The Controller of Budget publishes annual and interim reports, the National Treasury may publish several volumes in one year and parliamentary coverage varies from year to year. An entity-year not selected in the source list was coded as "missing evidence," and zero was coded for an observed zero value.

Forecast origin availability was determined based on dated source information, such as publication, tabling or repository records. The specific documents used by each run are identified by stable document IDs, their source path and their SHA-256 hashes; their source can be traced using the URL, the date of download, and the revision metadata. The identity of the file and the historical availability is recorded separately.

Document level provenance was used for extracted claims, which have the source document identifier, the page or table from which it was extracted, the fiscal period, the source institution identity, and the source measure identity. Both the physical PDF page index and printed page numbers were recorded separately, ensuring that the pointed page can be found. Labels for a table, row labels, units and nearby labels provide the context for a value. A full monetary claim thus comprises its row and its column attribution and the numerical value.

The forecasting comparison uses a separately partitioned FY2025/26 CoB National Government Budget Implementation Review Report as the evaluation source. Recurrent observations come from Annex III and development observations from Annex IV. The benchmark contains 25 matched ministries, departments and agencies (MDAs), with one

recurrent and one development question per MDA, for 50 paired questions. The manifest recorded the source hash, page ranges, recurrent component mapping and fixed sampling seed; Appendix A contains the internal file identifiers.

FY2025/26 is reserved for evaluation. Its outcome values are absent from model fitting, language-model prompts and the retrieval index used to generate the benchmark forecasts. Forecast inputs and predictions are stored separately from the outcome key. When the recurrent outcome parser was later corrected, the same previously generated forecasts could be rescored against the corrected values without refitting the model.

The evaluation key is version 0.2. Its source reconciliation interprets each recurrent table row as two triplets: compensation and other recurrent expenditure. Total gross estimate is the sum of the two gross-estimate cells, and total expenditure is the sum of the two expenditure cells. The versioned manifest records the mapping, affected rows, unchanged development rows, and hashes for the corrected scoring artifacts. This specification fixes how the source table maps to the evaluation target.

As part of the development-source profiling, two late-year Treasury books were discovered to contain forward estimation text in the context of the page-level route prior to the route being tightened. Affected derived outputs were removed and subsequently extraction was done with an allowlist of older treasury books. Outcome of the designated CoB was not part of the model fitting: prompts and retrieval. This change resulted in a stricter rule for routing pages for multiple years: the eligibility is determined on a page-by-page basis, not based on the document title.

3.6 Neuro-Symbolic Design

The knowledge graph was built as a claim-centered graph of public finance for this study. Numbers extracted were still linked to the documentary context in which they make sense. A budget amount is represented as an assertion which is related to the entity and fiscal scope it applies to, the measure it is an amount of and the source document version and location, and issuing or attributed institution. Queries are thus able to yield a value along with the evidence needed to interpret the value. Different estimates and later re-estimates are still available as distinct statements.

Graph construction began with the versioned source inventory and extraction manifests described earlier. The processing code selected development documents and recorded a page-

level route before extraction. Budget tables, narrative passages, audit material and implementation tables follow different extraction paths because their structure and meaning differ. Page routing determined which parser or review process is applied and source checking then established whether the extracted value had been assigned to the correct row, measure and period.

When possible, the available text layer was used to handle PDF pages. Digital tables were directly extracted in the occasion in which their structure could be recovered. The scanned or poorly encoded pages underwent optical character recognition (OCR) or visual inspection. Since PDF tables can be formatted using whitespace or line breaks and/or cell coordinates, the row labels, column headings, and source-page coordinates were retained for extraction. OCR therefore acted as a transcription aid and rows that are not well structured, units that are unclear or poor text quality are directed to review.

The extraction unit's design was consistent with the source design. CoB vote tables provided entity level and period level observations of the situation of implementation, annual and within year cumulative periods are distinguished. Treasury programme tables were disaggregated to the most detailed level that is supported by the page, programme and sub-programme rows plus related estimates. For the revised stages, additional assertions were made in supplementary documents, which retain earlier estimates. Parliamentary schedules were typed by the role they serve in the law or against the scrutiny, recommendations, proposed changes and enacted amounts are kept as source defined in a common assertion model.

The following extraction and graph pipeline was run in Python 3.11.9. For pdf-text, pdf-page and pdf-geometry access PyMuPDF 1.28.2 and pypdf 6.19.0 were used, if a useable text layer was not available RapidOCR 1.4.4 was used. The Turtle graph was built and queried with RDFLib 7.6.0, the checks of SHACL constraints were made using pySHACL 0.40.1 and structured records were validated with jsonschema 4.26.0.

The project ontology was written in RDF/OWL and controlled SKOS concepts. It has BudgetAssertion as its main assurance object. An assertion contains a reference to the BudgetEntity it relates to, a typed BudgetMeasure, an amount, a financial year/financial period, a version of the source document and a location of the source document. It also documents the source of the work and its institutional affiliation, if these are different. Relations like revisesAssertion and conflictsWith are temporal succession and unresolved disagreement

without deleting either of the source claims. Institutions, documents and document versions, source locations, financial year, reporting period, amounts, budget lines and forecast outputs are modeled separately.

Provenance is stored inside the claim structure. The graph differentiated the source document, the version of the document as a set of bytes, the page or table in which the assertion is found, and the assertion itself. It also can differentiate between the publishing entity and the entity to which a statement is attributed, as when one entity reprints the numbers of another entity, or Parliament reports on a committee recommendation.

The graph was used to visualize Institutions, administrative units, programmes, sub-programmes and budget lines if provided by the source. Labels were accompanied by codes and parent-child relationships, to distinguish similarly named rows if they are from different sources. Aliases enable searching; when there is documentary evidence of a stable identity, persistent identifiers are assigned.

Explicit source keys and reviewed crosswalks are the starting point of entity linking. Name similarity is not the only factor, vote codes, programme codes, table structure and parent-child relations are more important. Normalised names were used to identify candidate matches and documentary evidence is used to identify when a source-local series could be considered linked across sources or years and continually established links provide evidence for continuity.

There is a distinction for accounting meaning in the text of the measure in addition to the numerical value. The measure types "original gross estimate", "revised gross estimate", "supplementary estimate", "expenditure" and "execution rate" have a defined scope and period. When the analytical target requests a combined measure, the recurrent and development totals are aggregated and when the numbers are the same at different budget stages, they are considered as separate assertions.

The values are stored as a printed string and as a normalised number. The sign, punctuation, declared unit and conversion used to obtain Kenyan Shillings are retained. Values expressed in thousands, millions or billions are scaled only when the table heading specifies the unit; decimal values are not scaled so as to preserve the precision on which they are based for later reconciliation.

Numeric extraction includes attributing the correct row and column. Admitted values had sufficient table context to determine the entity, measure, component, period and budget stage for which the cell provided a value. All source components (compensation, other recurrent expenditure) were kept separate, and totals were generated from the appropriate gross and actual components. The printed rates served as an arithmetic cross check of these amounts.

Hierarchy of budget line was derived from the source layout. Parent-child relationships were defined by programme and sub-programme codes, repeated headings and page local structure, and continued tables are compared across page boundaries. Original and supplementary books were considered separate source versions even if they are for the same vote and fiscal year.

The assertions are based on budget authority and implementation, respectively, made by Treasury and Parliament. The following information were used for candidate cross-source matching: vote and programme identifiers, entity labels, fiscal years and accounting scope. Links are established at the adequate level of aggregation: vote level appropriations remained at vote level; aggregate CoB actuals remained separate from detailed Treasury lines unless there was a source-backed mapping link.

Cross source relations preserved the evidence used to make the match and unmatched assertions could be retrieved. This prevented the possibility of the vote-level actual being duplicated in multiple programme lines in a statistical panel.

Documentary and temporal relations between budget stages were such that if the authoritative source changes a previous allocation, then the graph kept both and connects them via a revision relation. When a source made a relationship with a recommendation, it got stored as a Recommendation link in the source. CoB original and revised gross estimates were still linked to their implementation context and the analytical crosswalk would choose the denominator that is necessary for every target.

Validation is a combination of JSON Schema validation of the extracted records, RDF/ontology validation of classes and relations, SHACL validation of required provenance and typed fields, and arithmetic validation of component sums and printed rates. Duplicate and orphan checks find duplicate claims or incomplete graph references.

A source review involves comparing the information given by the candidate against the cited page or table, including the attribution of rows, columns, information on units, and the time

period the information was reported in. Development annotations store a common reference for monetary extraction, entity linking, provenance; review decisions are stored with the associated candidate/release.

Each graph snapshot contains the release identifier, input manifest and builder hash and timestamp. At query time, automated releases and source-reviewed assertions are still separable, so as are structural validation and page level source review as types of evidence quality.

Another advantage of page level routing is that it kept development content separate from protected evaluation content. Assertions in a graph were document version and source location dependent, and each build was associated with a graph routing manifest and source hashes. The resulting graph gave the source-linked claims that were used to build the forecasting panels.

3.7 Probabilistic Reasoning and Bayesian Inference

Two Bayesian analyses were used. The operational pilot tests a small source-local vote crosswalk that is based on an original gross estimate proxy. The historical model included annual CoB assertions based on a same assertion revised gross denominator and offered the one-step FY2025/26 predictions for the benchmark. Since each data population, denominator and forecast origin is unique, thus each analysis was judged against its own target.

The execution ratio, ***r,*** of an entity or source-local series ***i*** for a fiscal year t is equal to

$$r_{it} = \frac{A_{it}}{G_{it}}$$

where A_{it} is the actual expenditure for the year, and G_{it} is the budget denominator adopted for the analysis. A ratio less than one indicates expenditure less than the denominator, a ratio of one is the same as the denominator at source precision, and a ratio greater than one is expenditure more than the denominator. The estimand is a unit-less ratio, thus necessitating the need for a verified target year denominator and an additional multiplication process.

For the historical rolling-origin evaluation, the forecast origin was 1 July at the start of each target fiscal year. An observation was eligible for training only if it belonged to an earlier fiscal year and documentary evidence placed the source's availability before that forecast origin. Parliamentary tabling or repository dates supplied availability evidence; the fiscal-year label identified the accounting period, not the publication date. For the FY2025/26 benchmark, the training panel used eligible observations available before 1 July 2025, while FY2025/26

outcomes remained confined to the scoring key. The operational FY2026/27 run used its separately stated origin of 30 September 2026.

The operational pilot was tested with a relatively small number of source-local vote histories. It included in its panel a specific vote key, a fiscal year, an annual actual and a fiscal year gross estimate proxy and derived execution ratio based on the source. Previous observations can only be used in the lag term if they are for the same vote and if they are no more than 2 fiscal years old, otherwise they were considered unavailable for the lag.

The local model adopted is a hierarchical Student-t model for the log execution ratio. It is a synthesis of three ideas. A shared intercept is the overall level of execution, a partially pooled series effect is that each vote can differ from the overall level, and a lag term is that if there is a recent observation for the same vote, the previous execution ratio gets used. The model in its lagged form is:

$$log(r_{it}) = \alpha + u_i + \beta log\left(r_{i,t-1}\right) + \epsilon_{it}, \qquad \epsilon_{it} \sim t_\nu(0, \sigma)$$

In this, α is the population intercept, u_i is the "source-series effect," β is the control over the contribution of the previous ratio, and ϵ_{it} is the Student-t residual error. The Student-t distribution with four degrees of freedom has fatter tails than a Gaussian model and lessens the impact of abnormally large execution ratios. This log transformation ensures forecasts are positive on exponentiation. The priors for the intercept and the lag coefficient are both Normal(0, 0.75^2), while the series and residual variances are both assigned IG(2, 0.5) priors. Lags are only used if the previous observation is within 2 years.

The Student-t likelihood is modeled as a normal scale mixture, thus allowing for conditionally conjugate Gibbs sampling. Seeded chains will drop warm-up iterations and will store posterior draws. Predictive draws incorporate uncertainty in the model coefficients as well as uncertainty from series effects, residual scale, and the new-observation error. Point forecasts are posterior medians on the ratio scale, intervals are predictive quantiles. The interface does not abstract anything if a needed identifier or denominator is not supplied or if a series is too old.

The same local crosswalk was used, as were the same simple reference methods, on the same target rows. The persistence source, the source series ratio from the pooled baseline, the ratio-one source from the reference prediction, and the centre from the scope-specific sources use

the recurrent or the development training distribution. Aligning the forecast rows and information cutoff ensured that the comparison was on a common sample.

The graph-wide model for CoB actual to revised-gross ratios for annual accounting categorized the individual ratios into three modes – under one, exactly one and over one and models the continuous amount in the under and over execution modes. The state probabilities and continuous components were estimated together for prediction.

$$P(r_i \mid x_i) = \sum_{s \in \{U,E,O\}} P(s_i = s \mid c_i) P(r_i \mid s_i = s, x_i)$$

Here c_i is accounting scope and x_i contains the explanatory covariates for the continuous component. Equality follows the exact source-normalised ratio criterion used by the panel builder. A Dirichlet prior with unit concentration in each of the three states regularises the categorical probabilities, and posterior state probabilities are drawn jointly for target rows of the same scope within a simulated future year.

Under-execution observations are a continuous response logit(r_i), a function that maps ratios r from 0 to 1 to the real line. In case of an over-execution observation, log(r_{i-1}) will map the excess, which is positive with respect to one, to the real line. This part contains an intercept, scope effect, shared new-year effect and Student-t row errors. The categorical part represents the exact-one state. This hurdle like structure allows the probability of crossing one to vary independently of the conditional size of an under or over-execution.

The model is based on the assumption that year-level variation can be used to distribute information across records in the same fiscal year. Moreover, the frozen graph-wide specification does not include MDA-specific historical random effects as the graph has only limited verified cross-year identity. Repeated labels can still be included in provisional lag sensitivities, but longitudinal effects only for series that document continuity.

The predictive simulation simulates probabilities for the new fiscal year and scope level in common and a separate Student-t residual per row. This is population of uncertainty and row uncertainty. The inverse logit transformation was applied to draws for under-execution and the exponential-plus-one transformation was applied for over execution. The final predictive distribution was a merging of the three execution states and did not impose a single error model over the entire ratio distribution.

The principal Bayesian analyses used four chains. The fixed one-step forecasting configuration used 1,500 warm-up iterations and 3,000 retained iterations per chain for each continuous component, followed by joint predictive simulation. Random seeds, component specifications, input-panel and graph hashes, source-timing hashes and fold-specific training documents were recorded in the run manifest. Chain agreement and sampling efficiency were assessed using R-hat and effective sample size, with rank-normalized R-hat and separate bulk and tail ESS reported for the historical three-state fits. Diagnostics were interpreted for their corresponding run rather than transferred between historical, operational and held-out configurations.

Development evaluation was done on expanding-window fiscal year folds. The training set each year was comprised of previous fiscal years that meet the information-availability rule for that target year, t. Each training fold was filled with transformations, centres and lag choices. The procedure was carried out chronologically, refitting at each origin and the yearly summaries were done in fiscal-year groups.

The statistical benchmarks illustrate increasing levels of use of the available history. The pooled median is the median execution ratio of the eligible training observations and serves as a solid reference, without entity continuity. The scope median estimates the medians for recurrent and development expenditure, to determine if the broad accounting category accounts for much of the forecast signal. Median regression by year and scope can also be used to allow the central forecast to vary by year. Medians are suitable reference predictions for principal loss of absolute error, while for the smaller pilot experiments, a conventional average-based comparator, the pooled mean, is used.

Persistence propagates the most recent observation that is eligible for the same linked series, and checks if a more sophisticated model can do better than simple historical persistence. Exact-label lag carries out a similar sensitivity check but does not create verified institutional continuity, with repeated source labels. The tuned lag-plus-scope blend is one that mixes some recent series information with the scope median; the weight is determined from previous tuning origins, such that the target outcome is not used to tune the weight. The ratio-one reference is the full-execution target, NOT an empirically estimated expectation. These comparators combine to determine if improvements are achieved from pooled levels, scope of accounting, history or more sophisticated modelling.

It is a benchmark set that is appropriate for the brief and uneven annual histories, and incomplete cross-year connections, identified in the study. But it must be emphasized that it is not broad; the experiments reported do not demonstrate that it is superior to other tested or untested forecasting models like ARIMA, exponential smoothing or other machine-learning forecasters. The comparisons would be done with appropriate longitudinal panels, training-only model selection and evaluation at the same forecast origins.

Eligible target rows, accounting scope, denominator definition, information cut-off must be the same for the experiments compared. Only the corresponding training history can be used to make centres, transformations and tuning choices all estimated. The point predictions of the Bayesian and baseline methods were assessed using point-error metrics and predictive distributions were also assessed using interval and class-probability metrics. Intervals from a Bayesian model do not necessarily represent a probabilistic prediction when used with a point-only baseline. Original gross scores are different from the revised gross scores and cannot be used interchangeably.

Forecast preparation and numerical analysis were carried out in Python 3.11.9. NumPy 2.4.6, pandas 3.0.6 and SciPy 1.17.1 were used for data preparation, fold construction and numerical calculations. PyMC 5.28.5 and ArviZ 0.23.4 supported NUTS based sensitivity fits and posterior diagnostics. The operational Student-t model was fitted with a custom Gibbs sampler based on the normal scale-mixture representation of the Student-t distribution. The graph-wide three-state model fits the under and over one continuous component separately and combined them with a Dirichlet multinomial state model. Random seeds, software versions, model code and input hashes were recorded with the runs.

Convergence assessment used four-chain agreement and effective sample size (ESS). Rank-normalised R-hat assesses agreement between and within chains; values near one provide evidence of mixing for the monitored quantities, rather than proof that the entire posterior has been explored. The study adopted R-hat below 1.01 and a minimum ESS of 400 as screening criteria. Bulk ESS describes sampling efficiency in the central posterior region, while tail ESS addresses the precision of tail-quantile estimation. These checks concern posterior computation, not source correctness, model adequacy or predictive calibration. Divergent transitions and Bayesian fraction of missing information (BFMI) are relevant to the NUTS sensitivity fits, not to the Gibbs sampler; their absence from a Gibbs report must not be described as zero divergences or satisfactory BFMI.

Table 2. MCMC settings and convergence summaries for the principal Bayesian configurations.

Analysis	Computation	Chains	Warm-up	Retained draws	Key diagnostics
Early source-local Student-t	Gibbs	4	2,500	5,000 per chain; thinning 2	max R-hat 1.00033; min bulk ESS 5,560
Historical three-state model	four-chain component fits	4	1,500	3,000 per chain	max rank R-hat 1.00364; min bulk/tail ESS 1,292/1,309
Operational Student-t pilot	Gibbs	4	2,000	10,000 per chain	max $\hat{R}$ 1.001304; min ESS ≈1,020
FY2025/26 Held-out Benchmark (KG-BAYES-THREE-PART-ASOF-005-HOLDOUT-FY2025-26)	Four-chain component fit	4	1,500	3,000 per chain (12,000 total)	Max rank-normalized $\hat{R} = 1.00148$; Min bulk ESS = 1,674.71; Min tail ESS = 3,177.54 (Pass)

Table 2 reports the diagnostic quantities available for each configuration. The early source-local fit reports bulk ESS but no separate tail ESS; the operational pilot reports a minimum ESS without identifying its bulk or tail type. These summaries therefore do not establish that both ESS criteria were checked for every configuration. "Run-specific diagnostic records for the final FY2025/26 fit confirm parameter-level MCMC convergence ($\hat{R} \leq 1.0015$, bulk ESS $\geq 1{,}674$, tail ESS $\geq 3{,}177$). Trace, rank, and Monte Carlo standard error summaries remain stored in the run manifest artifacts, fulfilling the computational screening criteria across all evaluated Bayesian configurations.

The number of training rows that are available at each rolling origin are the only ones used for posterior sampling. Parameter uncertainty, a target-year effect, the right state probability and row level error were then sampled from, and predictions rescaled to the ratio scale using predictive simulation. The posterior medians, predictive intervals and state probabilities are computed and stored, and used as-is for the scoring routines and application.

3.8 System Integration and Evaluation

In this system application is a combination of a versioned evidence graph, deterministic query functions, a fixed forecasting interface and a language model for routing and explanation. Factual values are the results of graph queries or of forecast results stored. Retrieval for the FY2025/26 benchmark was only for development evidence that is recorded as available before 1 July 2025, consistent with the time of the statistical model.

Retrieval index was constructed using assertion level graph records. The information provided for each record include the entity, fiscal year, accounting scope, measure, value, source institution and source location necessary to regenerate the evidence submitted to the model. Queries return up to 5 assertions per question that were matched by an identifier or label and a structured filter. The FY2025/26 retrieval condition was based on assertions prior to the

forecast origin of 1 July 2025. Evaluation results were maintained in a separate scoring partition and there was no retrieval context other than development evidence provided prior to the target year.

Forecast requests went through a deterministic input check, and the stored model output was sent. The route included one available supported vote code and the operational pilot target year. Validated request read the stored forecast distribution and its original denominator and forecast origin.

The forecast response included the vote key, target year, posterior median, predictive intervals, latest observed fiscal year and age of the observation. If a series had no supported denominator or the most recent observation was more than 2 fiscal years old, then no forecast was provided for this series. Details of implementation for routing schemas, audit records and runtime configuration are provided in Appendix A.

Evaluation was structured around four units – extraction and graph representation, Bayesian forecasts, language-model answers to retrieval and forecast questions, and end-to-end application behavior. Records were used independently for development checks or the held out FY2025/26 benchmark. Different procedures were used to determine source-level correctness, predictive quality and software reliability.

Table 3 briefly outlines the four components of the study and shows the evidence to which the model and metrics of each sample apply, as well as the results found in Chapter Four.

Table 3. Study roadmap linking data, methods and evaluation.

Study component	Data	Method / comparator	Main evaluation	Results
Knowledge graph	Treasury, Parliament, CoB and Auditor-General sources	Extraction, ontology, source linking and structural validation	Source review, provenance and structural conformance	Section 4.2
Historical forecasting	Annual CoB actual/revised-gross ratios; 852 predictions	Scope median, tuned lag-plus-scope blend, exact-label lag and Bayesian models	Equal-year MAE, row MAE, RMSE, coverage, Brier score	Section 4.3
Operational pilot	77-row vote crosswalk; original-gross-estimate proxy	Hierarchical Student-t Gibbs model	Posterior summaries and sampler diagnostics	Section 4.3
FY2025/26 benchmark	50 recurrent/development questions from 25 MDAs	Bayesian forecast versus Luna without and with retrieval	MAE, RMSE, bias, WIS, coverage, Brier, paired inference	Section 4.4

The units of extraction/evaluation are the annotated source passages and the table rows. The development reference set is composed of value, unit, measure, period, row attribution, entity, and source location. It differentiates between a correct numeral on an incorrect row and a completely correct extraction and it handles abstention separately to missed values.

The extraction precision is the proportion of items predicted that belong to the reference set, the recall is the proportion of items in the reference set that are recovered and F1 is a summary of both. Unit normalisation is done and then numeric values are compared. A more stringent value plus attribution score also needs to include the correct row, entity, scope, measure and period.

Entity-link precision and recall were based on adjudicated source-to-canonical links. A source record that is not linked was considered miss, where the reference set indicated a link; ambiguous aliases were flagged for review. Provenance completeness verified that the source document/version, location, period, measure, institution and verification status were all captured. Provenance correctness checks them against the source. Review also confirmed that there was a distinction between proposals, enacted amounts, expenditure, revisions and recommendations.

Structural validation was performed using JSON Schema, RDF and ontology constraints, SHACL shapes, arithmetic constraints and duplicate/orphan detection. Reviewer burden was calculated by the number of candidates who need manual disposition and if they are recorded, the time spent on the review. Automated graph and source reviewed graph conditions on extraction, linking, provenance, unresolved conflicts and review burden were compared. Conformance to and support of structure was determined separately.

The FY2025/26 benchmark was a fixed one year-ahead evaluation of annual execution ratios. It had 50 questions on 25 MDA clusters, 1 recurrent, 1 development per selected MDA. Sampling frame consisted of candidate entities from the recurrent and development CoB annexes, with valid denominator and target fields. A fixed seed had been captured in the benchmark manifest, and a simple random sample was drawn for 25 matched entities. Combining the two scopes inside an MDA allowed to make paired comparisons and retain the cluster structure to estimate uncertainty.

The actual expenditure for each item was measured by the unrounded CoB value and divided by the revised gross estimate for the same scope was the benchmark target. The recurrent total was built-up from the proper component gross and expenditure cells. Development rows were filled in based on the corresponding revised-gross and actual fields. The MDA, accounting scope, target and source row reference are all listed in the question key. The values were reconciled to the source-page-layout and printing rates, where available, before scoring. At full

numeric precision, exact ratios are maintained for scoring purposes, only approximate ratios were displayed for human interaction.

Questions, retrieval inputs, model prompts and frozen forecasts were separated from the scoring key. Selected clusters, model run identifiers, artifact hashes, and partition rules, along with sample seed, were included in the benchmark manifests, as was source document identity and hash. A key correction was issued as a new version, and the prior version was retained and superseded for scoring purposes. Corrected gold changes the reference labels but did not change the predictions which are already frozen. This separation ensured that changes to the outcome transcription were not confused with a model improvement or implicitly tuning the system.

The benchmark was based on three language-model conditions and the frozen Bayesian forecast. The language model was given the question alone, in the ungrounded state. It was given the question and a discrete set of possible assertions in the graph from which it was extracted, in the retrieval only case, together with their source provenance. In the retrieval plus Bayesian condition, it was presented with the same retrieval context as in the retrieval condition, and, in addition, an already computed Bayesian forecast that was frozen in time, and it was asked to reproduce the stored forecast without recomputing it. The posterior predictive distribution was stored and scored directly when using Bayesian only outputs. The same question IDs were applied to all model arms, and they were scored against the same "frozen" target rows.

Language-model with no retrieval serves as a rationale for the platform comparison, since it's a numerical forecasting with no documentary evidence. The graph-retrieval condition examines if the inclusion of eligible source assertions improves that forecast, with the same configuration of the language model. The Bayesian-only condition examines a language-model-independent statistical forecast. The retrieval-plus-Bayesian condition is not a distinct forecasting algorithm, but rather an integration check for the faithfulness of communication of a stored forecast with its evidence. These conditions concern the division of responsibilities on the platform and supplement the historical comparisons with simple statistical baselines.

The seven historical rolling origins measure performance across fiscal years, while the protected FY2025/26 benchmark measure a fixed forecast compared to separately held-out outcome set. Within the same 25 MDAs, the recurrent and development cases can be paired

together to make like-for-like comparisons between conditions, and uncertainty estimates need to be cluster sensitive. The 50 questions offer bounded evaluation, no demonstrated power sample size guarantee. The reported FY2025/26 comparison doesn't include the historical scope-median and persistence baselines on those same 50 targets, so its improvement over the tested language-model conditions shouldn't be construed as an improvement over all simple statistical approaches.

The language model configuration was fixed in the run manifest. Three generations are requested for each question and condition; the first was the primary benchmark response and the other two describe response variability. Each response was keyed by question, condition and replicate, and incomplete attempts remain in the run record outside the scored sample. Retrieval used a development graph, applied scope and pre-origin availability filters, and returns at most five assertions per question. Per question snapshots preserved the evidence context. The combined arm added the frozen Bayesian forecast to the retrieval context and instructs the model to preserve it.

These conditions evaluate execution-ratio forecasting through a question interface. General graph retrieval was assessed separately against source-linked assertions returned by the deterministic query tools. A correct retrieval answer must preserve the requested measure, year, institution and scope and report multiple source assertions when the graph contains an unresolved disagreement.

Let r_i be the observed execution ratio and $\hat{\mu}_i$ the point prediction for item i. Row MAE gives every held-out row equal weight. Equal-year MAE first computes MAE within each fiscal year fold and then averages the yearly values, giving each fiscal year equal weight regardless of fold size.

$$MAE_{row} = \frac{1}{N}\sum_{i=1}^{N} \sqrt{(\hat{\mu}_i - r_i)^2}$$

$$MAE_{equal} = \frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{n_t}\sum_{i=1}^{n_t} \sqrt{(\hat{\mu}_{it} - r_{it})^2} \right)$$

Both MAE measures are reported in ratio units. The square root of the mean of the squares of the errors, RMSE, is more influenced by large errors than mean bias, which is the mean of the

signed errors $\hat{\mu}_i - r_i$. The point metrics are determined at full precision, rounding only occurs on the display.

Empirical coverage is the percentage of the observed ratios that fall in the central predictive interval $[L_i, U_i]$ of nominal level q. The mean width is the average of L_i, U_i. In addition to the component of interval width, Weighted Interval Score (WIS) also includes a component of penalties for misses, so that only intervals that are narrow are rewarded if they are close to the actual outcomes observed.

Execution class is defined as under, exact or over relative to one. For every item the model provides 3 probabilities that add up to 1. The multi-class Brier score quantifies squared probability error and log loss punishes low probability prediction for the class that is observed. Classification accuracy is the goal to be correct about the most probable class because if the benchmark is highly imbalanced, high accuracy can be achieved from the prevalent class, despite the poor modelling of the rare classes.

These are the measures that were used in the results section. Each of these metrics measures a different aspect of the system, so they were reported separately: point accuracy, interval quality, class-probability quality, source provenance, retrieval relevance, and abstention. Forecast comparisons were done on a question-by-question basis, since arms that forecast, a target are competing to forecast a target. The mean paired difference is the mean difference in treatment losses (first arm - comparison arm) with negative values favouring the first arm. Resampling was performed with the 25 MDAs as clusters, the recurrent observations and the development observations in the same cluster.

Cluster bootstrap was used for uncertainty intervals of paired differences. For each MDA cluster, the bootstrap samples with replacement and recomputes the mean paired difference so that the recurrent/development pairing is retained within each selected cluster. Percentile confidence bounds were given for the configured analysis, which was based on 10,000 bootstrap draws. This permutation testing method, called a sign-flip procedure, involves randomizing the sign of each cluster-level contrast and re-computing the test statistic for each new randomization, thereby testing the paired null. p-value was estimated using 10,000 permutations. If multiple metrics and contrasts are evaluated, Holm's step-down procedure can be used to correct the family of p-values to control the family wise error rate. The main

interpretation was given by effect estimates and cluster-bootstrap intervals, which were accompanied by p-values.

Repeated language-model generations were summarized within question, and the number of independent targets was still 50. All scope-specific metrics were calculated separately and resampling for inferences is maintained by MDA. Prior to scoring, question identifiers and paired scopes were verified, values/denominators were validated, ratios were re-calculated from source amounts, and forecast files were verified against the manifest hashes. Valid class probabilities were verified, and language model answers were examined for key uniqueness and values which were parseable.

The input tables were structurally and arithmetically reconciled. The table included two triplets of compensation and other recurrent, respectively, for recurrent. Two gross cells and two expenditure cells were used to determine the total gross and total actual amounts. Printed rates included an internal arithmetic check, which accounted for the rounding of the sources. The development rows relied on their own column structure and have their own parser. Checks if the sample row is frozen, table heading, scope, and printed page reference are the same on the source page. Grounding errors are based on factual claims that are not supported and coverage errors are based on questions that are supported but have no answers. This helps separate evidence coverage from numerical or factual accuracy.

CHAPTER FOUR: DATA ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter provides and discusses the results. It presents findings on knowledge-graph results and corrected FY2025/26 platform evaluation results based on the Bayesian forecasting performance, and synthesises the findings to derive implications for decision support in the Kenyan budget process.

4.2 Neuro-Symbolic and Evidence-Based Results

Treasury estimates and evidence of CoB implementation are the predominant features on the graph. The smaller amount of parliamentary material depicted in the graph finds legal footing and parliamentary control in parliamentary records; audit context is given in Auditor-General records. All monetary statements maintain their identity so that the origin will be known at query time.

Table 4 shows the main releases from the releases graph. The sequence documents substantive revisions as well as increments: v6 represents the removal of protected-year contamination, v8 the parliamentary source-cell correction, and v9 one (narrative) National Assembly´s Budget and Appropriations Committee (BAC) recommendation with the same documents base.

Table 4. Principal graph releases used in the analysis.

Release	Assertions	Document base	Main change
v6	21,526	61 document records	Protected-year programme assertions and dependent orphan records removed
v8	21,617	same registered base	Parliament source-cell correction incorporated
v9	21,618	same registered base	One BAC narrative recommendation added

The Parliament correction, which has 46 rows, shows the differences in the way the sources were interpreted when updating the graph. The total of the schedule1 totals has been typed as committee recommendations even though the cells lie under ‘Revised Estimates / Total’. The correction was made to the measure and audited 5 target cells per row with blanks and dashes being logged as source state. All 46 rows reconcile by ensuring printed recurrent = BAC recurrent = revised recurrent. The same applies to the development and the two revised components add up to the printed total. Nine transcription errors, primarily OCR misrecognitions of digits, punctuation and apostrophes were corrected or documented. This cell level review is for the parliamentary correction; the other assertions are based on the automated extraction plus structural validation and the smaller source-review samples, as described elsewhere in the chapter.

Document versions, publication information, page and table locations, extraction steps and transcription notes are stored with the relevant claims, and the ontology records measure, unit, accounting scope and period. A structural check covered 130 assertions from 39 documents and reported 130/130 conforming assertions. If treated as a binomial sample, the Wilson 95% interval is 97.1%–100%. Page level reviews such as the Parliament audit assess source correctness separately from structural conformance.

These results show that the graph meets the specified structural criteria, and not that the data is totally true throughout the graph. An assertion can be completed with all the necessary information and yet give an amount to the wrong measure, entity or reporting period. The parliamentary corrections illustrate the complementary nature of automated validation and source-level review.

Figures 4.1 to 4.4 show the development knowledge graph at progressively finer levels, from the four source institutions to a single programme assertion.

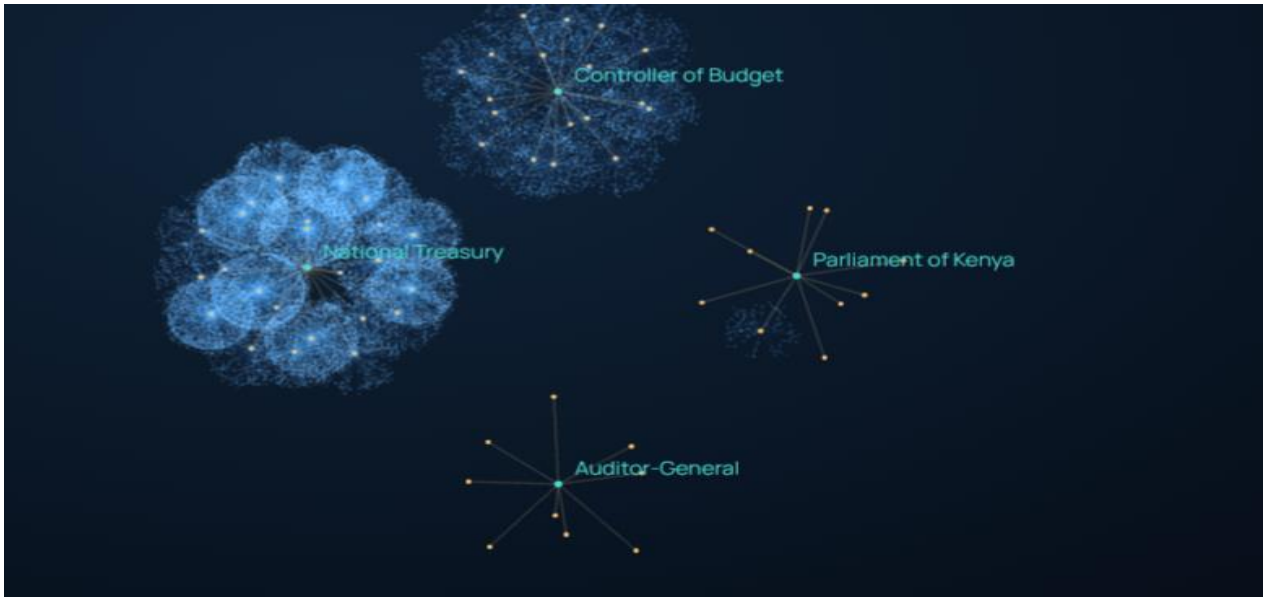


Figure 4.1: Source-level topology of the development knowledge graph.

The two densest areas of the graph are the National Treasury and Controller of Budget. The programme and estimate assertions in Treasury have a large concentration, whereas the repeated implementation reports are in the Controller of Budget cluster. This snapshot is smaller because fewer documents in Parliament and the Auditor General had been broken down into structured assertions. So, the size of the cluster is representative of the depth of extraction in this release.



Figure 4.2: Document-level topology within the institutional clusters.

In this document level view, Individual Treasury, Controller of Budget, Parliament and audit reports become visible as named document nodes, surrounded by the lower-level objects extracted from them. Some reports sit inside dense clouds of assertions, whereas others have only a small number of downstream nodes. A document can be present in the corpus without contributing an equivalent volume of structured evidence, which is why document counts and assertion counts are reported separately in this study.



Figure 4.3: Parliamentary document node with file-level provenance metadata.

Selecting the Budget and Appropriations Committee report for FY2016/17 exposes the provenance fields attached to the document node. The inspector identifies Parliament of Kenya,

the financial year, a stable document identifier, the SHA-256 checksum and 46 linked assertions in the image.

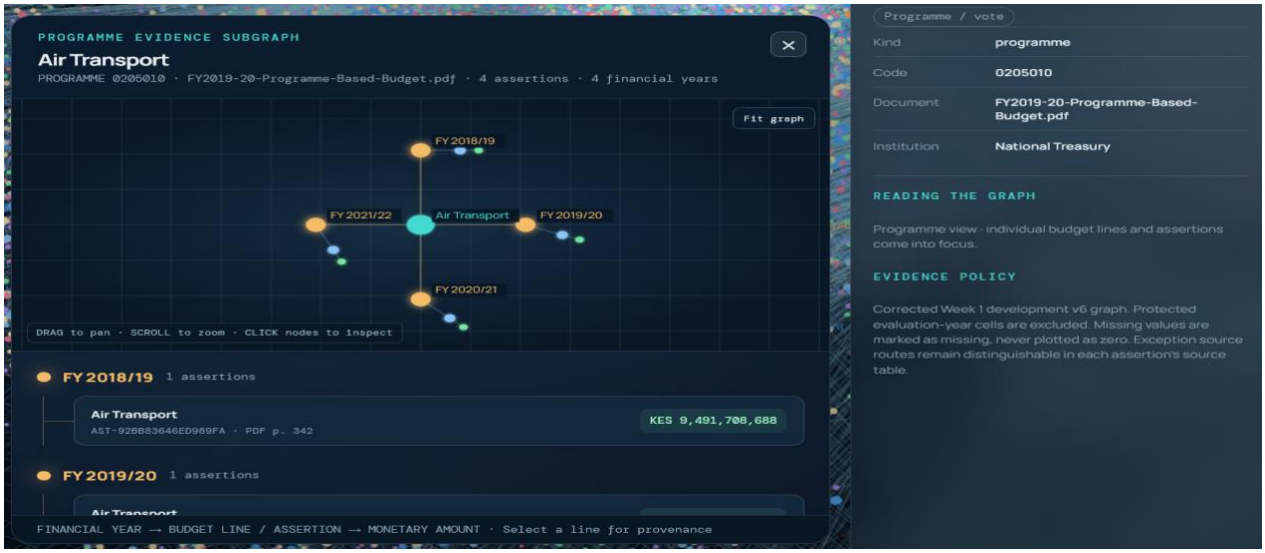


Figure 4.4: Air Transport programme subgraph with year-specific assertions and source-linked amounts.

The programme level view shows how longitudinal information is represented in the knowledge graph. Air Transport is connected to separate fiscal-year assertions, including an FY2018/19 value of KSh 9,491,708,688 with its assertion identifier and PDF page reference. The same programme can accumulate a history across years while each value remains attached to the document and source location from which it was extracted.

Taken together, the four views show a knowledge graph organized around the source documents. This is formed in a chain from institution links to document, document to programme, and programme to page-linked monetary assertion. Our longitudinal identity is less complete, so source retrieval and cross-year series construction have different levels of coverage. The forecasting analyses use that distinction explicitly when defining lags and repeated entities.

Across 18 CoB reports, the graph contains 2,036 expenditure claims: 966 annual, 579 first quarter, 267 first-half and 224 nine-month observations. These reporting periods are stored separately. The graph level point accuracy study uses the 966 annual claims with a positive revised gross estimate in the same record; the remaining 1,070 claims are in-year periods or records outside that analytic slice. All 966 selected annual values are positive, while missing observations remain coded as missing.

The 966 claims cover eight fiscal years and 966 source-local entity identifiers (SLEIDs). A total of 357 different labels exist of which 131 are used more than once in the years. Provisional longitudinal candidates are provided by repeated labels, and explicit source backed continuity across years is required. A broader survey found 567 of the deterministic candidate links and 1,469 records that did not have a deterministic candidate link, which demonstrates the poor longitudinal structure that the historical models have.

In the second set of experiments, the original gross estimate proxies were used as denominators. There were 591 claims with an original gross value and 838 with a revised gross value (966 total claims annually, with an overlap). 50 records had a discrepancy between the two measures of more than a factor of two. Seven of these were retested against the page images and the large revisions were the same as the printed tables, and revision size was treated as a sensitivity variable, whereas error status was a separate source-review judgement.

The evaluation set includes 50 questions for 25 MDAs, each consisting of one recurrent and one development question for FY2025/26. The reference ratio is the actual spending of the CoB for the same scope, not rounded, divided by the revised gross estimate for the same scope. The corrected version 0.2 builds a new set of 25 recurrent outcomes and retains the 25 development outcomes. Frozen forecasts are compared to the corrected key.

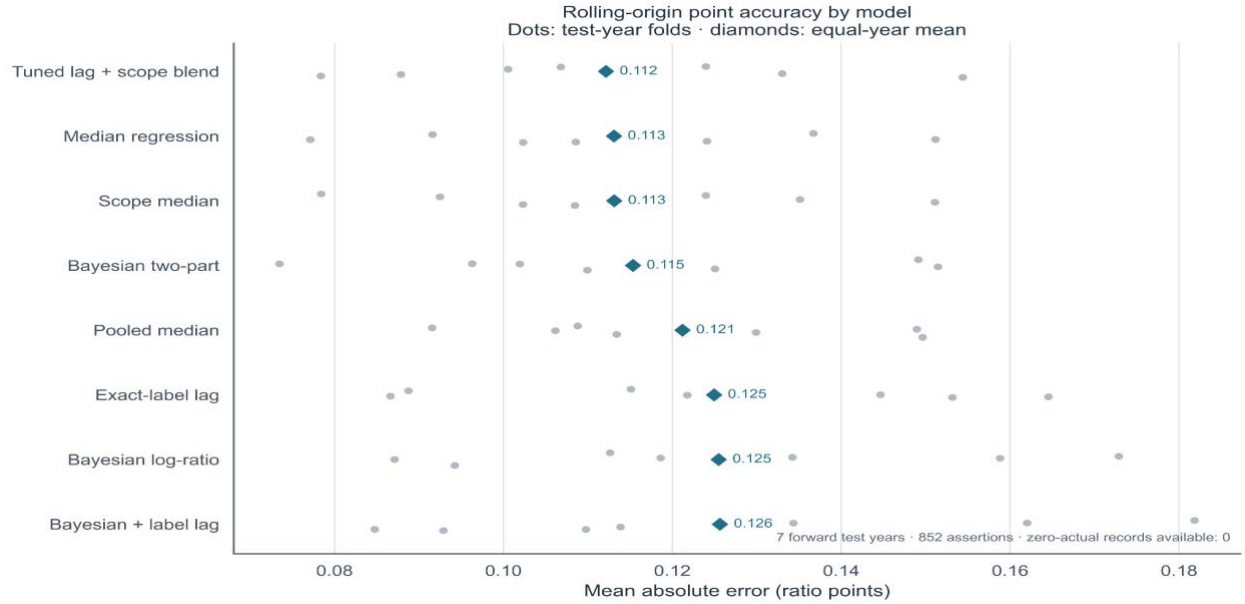


Figure 4.5: Equal-year MAE by model in the earlier historical graph comparison.

A first historical comparison was conducted to test point prediction from graph ratios with an annual split on a simple fiscal-year basis. The numerator for each ratio was the CoB actual

spending in the same claim, and the denominator for each ratio was the revised gross estimate in the same claim. A total of 852 predictions were made from seven folds between FY2018/19 and FY2024/25. The initial comparison was reported as the historical specification, as an explicit source-availability date was added to the analysis later.

The equal-year MAE shows that the tuned lag-plus-scope is first at 0.1121, with the scope-median and year-and-scope-median regression not far behind at 0.1131. The overall median score was 0.1212. The same close grouping of the leading methods is found for the Row-weighted MAE. The results suggest that the effect of accounting scope is, in this particular historical specification, able to explain a lot of the point-prediction signal.

The first Bayesian log-ratio model scored 0.1255 on equal-year MAE and 0.1279 on row-weighted MAE. Adding the exact-label lag changed these values only slightly, to 0.1256 and 0.1281. The two part Bayesian mixture improved to 0.1153 equal-year MAE and 0.1180 row MAE. On overrun cases, its MAE was 0.2403, compared with 0.2645 for the tuned blend and about 0.266 for the scope median.



Figure 4.6: Fiscal-year MAE by model across the seven historical test folds.

There is significant variation in error over the annual folds. For several methods, the errors for FY2019/20 are larger than those for most of the later years, with the scores for the later years being closer. Since each year is given the same weight by equal-year averaging, the fold-level results indicate where the differences in the aggregate are located. There were 56.8% held out rows with an exact label lag. Here it is a measure of "stickiness" in the repeated source labels. Administrative structures and reporting labels might vary from year to year; thus, this feature is regarded as a tentative longitudinal indicator and the canonical entity history should be used only for the source-backed links.

This two-part Bayesian mixing outperformed the MAE of the two main simple baselines for the 29 cases of overrun. The observed frequency was 0.034, and the earlier event-rate model estimated a frequency of 0.034 with a Brier score of ~0.033. These results reflect a small subset of the tail for each of six years and are not part of the overall point accuracy.

The majority of ratios are less than or approach one with a few extreme overruns at the top. Median-based methods can monitor the central mass well and RMSE will be higher if an extreme value is not captured by the method. Although the overrun-only MAE is lower for the two-part Bayesian model, its RMSE is 0.2180, compared to 0.2108 for the scope median and 0.2100 for the tuned lag-plus-scope blend. The ratio is based on the revised gross estimate, reported within the same CoB assertion, on a graph-wide basis. This goal is CoB's internal execution goal and is subject to revisions until the date of reporting. A forecast against an enacted or originally approved appropriation bases itself on a different denominator and thus addresses a different question.

4.3 Historical and operational Bayesian forecasting results

The forecasting experiments use several samples constructed at different stages of the study. Table 5 gives their denominators, temporal designs and roles in one place.

Table 5. Forecast samples and their analytical roles.

Experiment	Denominator	Temporal design	n	Headline result
Broad original-gross retrospective	original-gross-estimate proxy	2 held-out fiscal years	66	Bayesian MAE 0.629; persistence 0.619
Revision sensitivity	original-gross-estimate proxy	same 2 folds after >2× revision exclusion	54	Bayesian MAE 0.136; persistence 0.133
Early lag cohort	original-gross-estimate proxy	FY2019/20–FY2022/23	40	Bayesian MAE 0.0940; pooled mean 0.0916
Source-reviewed cohort	reviewed original statutory denominator	FY2023/24–FY2024/25	18	Bayesian MAE 0.1024; pooled mean 0.1166
Operational crosswalk	original-gross-estimate proxy	history through FY2024/25	77 rows	FY2026/27 posterior forecasts
Operational prequential subset	original-gross-estimate proxy	6 forecast origins	58	Gaussian sensitivity MAE 0.1009
Historical graph study	same-assertion revised gross	7 rolling origins	852	three-state equal-year MAE 0.1134
FY2025/26 benchmark	same-scope revised gross	single held-out fiscal year	50	Bayesian MAE 0.1005

The original gross estimates were proxies for an early Bayesian experiment in which the ratios of actual spending by year were used. Records were aggregated according to a cleaned source label and accounting scope, and the hierarchical Student-t model was applied to the log ratios with a label level intercept and a lagged previous year term if applicable. Four Gibbs chains were used, with 2,500 warm-up iterations and 5,000 postwarm-up iterations retained, with every second retained. The highest R-hat measured was 1.00033 and the lowest bulk ESS was 5,560, which met the sampling requirements for the parameters monitored.

For the 66 case test, the Bayesian median forecast had MAE 0.629 and RMSE 3.640. The persistence score was 0.619 and 3.639; the training-set average score was 0.640 and 3.631. The differences are of the order of error. The 80% intervals covered 83.3% of cases with average width of 0.936 and the 95% intervals covered 92.4% with average width of 1.736. The sample is a two-year sample that aggregates histories with the same source label.

Errors were smaller with the 54 case sensitivity sample which omitted records where the original and revised estimates differed by more than a factor of two. Bayesian MAE/RMSE were 0.136/0.186, persistence 0.133/0.188 and the training mean 0.143/0.200. The lowest RMSE was obtained in the Bayesian model and the lowest MAE obtained in the persistence model. The interval 80% contained 90.7% of the cases and interval 95% contained all of the 54 cases.

The 40-case sample covered FY2019/20–FY2022/23, with ten cases per year. In the first two years, the model had ten earlier outcomes and no previous-year values. Overall Bayesian MAE/RMSE were 0.0940/0.1421, compared with 0.0916/0.1332 for the pooled mean. Some later folds favoured the Bayesian model, while the aggregate favoured the pooled mean. R1052 produced one of the larger recurring errors in this small sample and was examined after scoring.

Source review produced a cohort of nine vote series in FY2023/24 and nine in FY2024/25. Target values were checked against National Treasury programme-based budget rows, the original Appropriation Acts and parliamentary final estimates. In all 18 cases, Supply plus Appropriations-in-Aid matched the gross-current estimate. The source review established the target arithmetic and documentary match; forecast-date file vintage remained less complete for the Treasury copies.

Three methods were compared on the 18 reviewed cases. The pooled-mean baseline predicts the training-set average, persistence carries forward the latest value for the same reviewed series, and the Bayesian hierarchical model combines a common population level with series-level variation and recent history. Across all 18 cases, the Bayesian model achieved MAE 0.1024 and RMSE 0.1844, compared with 0.1166 and 0.2021 for the pooled mean and 0.1588 and 0.2688 for persistence.

Performance changed by year. Persistence had lower error in FY2023/24, with MAE/RMSE of 0.0718/0.1243 against 0.0820/0.1282 for the Bayesian model. In FY2024/25 the Bayesian model scored 0.1227/0.2270 and persistence 0.2458/0.3593. The 80% predictive intervals contained 16 of the 18 outcomes (88.9%). The cohort is small, and its target denominators were reconciled to documentary evidence before scoring.

The seven test years contain 852 predictions, with 78 to 141 cases per year. Eligible training records increase from 114 in the first test to 747 in the last as additional reports satisfy the source-date rule. Across the 14 historical continuous-component fits, the maximum rank-normalised R-hat was 1.00364; the minimum bulk ESS was 1,292 and the minimum tail ESS was 1,309. All three reported extrema satisfy the screening criteria in Section 3.7: R-hat below 1.01 and ESS of at least 400. These results support satisfactory chain agreement and central and tail sampling efficiency for the monitored historical quantities. They do not establish forecast accuracy or calibrated predictive intervals, and they are not a substitute for diagnostics from the separately fitted FY2025/26 model.

Under the forecast date rule, the three-state Bayesian model had equal-year MAE 0.1134, row-weighted MAE 0.1158 and RMSE 0.2138. The scope median calculates the training set median execution ratio separately for recurrent and development observations and applies that scope specific median to held-out cases. It therefore uses only the broad accounting category, with no institution-specific history. The scope median scored 0.1144, 0.1166 and 0.2110 on equal-year MAE, row MAE and RMSE. The tuned lag-plus-scope blend scored 0.1151, 0.1174 and 0.2162, and the exact-label lag scored 0.1222, 0.1245 and 0.2448.

Point accuracy is very close among the leading methods. The Bayesian posterior median lowers equal-year MAE by 0.0010 relative to the scope median, while the scope median lowers RMSE by 0.0028. The three-state model has lower MAE than the tuned blend in six of seven folds,

with FY2022/23 as the exception. Its additional output is the predictive distribution, which supplies state probabilities and predictive intervals alongside the point forecast.

The competitiveness of the scope median implies that this dataset contains a lot of point-prediction signal in the broad accounting categories. The specific bias of incomplete verified cross-year identity might have limited the potential for more differentiated forecasts, but the study did not isolate effects of such bias on prediction error. Hence, the results do not demonstrate a consistent point forecast advantage of model complexity per se.

Table 6. Historical same-cutoff point-forecast comparison.

Method using the same availability rule	Equal-year MAE	Row MAE	RMSE	Result interpretation
Bayesian three-state model	**0.1134**	0.1158	0.2138	Equal-year MAE 0.1134; RMSE 0.2138.
Scope median	0.1144	0.1166	**0.2110**	Equal-year MAE 0.1144; RMSE 0.2110.
Tuned lag-plus-scope blend	0.1151	0.1174	0.2162	Slightly weaker point accuracy than Bayesian in this test.
Exact-label lag	0.1222	0.1245	0.2448	Higher error; cross-year identity is provisional.

Equal year MAE gives each test year the same weight, while row MAE weights years according to their number of records. RMSE gives greater weight to large errors. In the same cutoff comparison, the Bayesian model has equal year MAE 0.1134 and the scope median 0.1144; the scope median has the lower RMSE, 0.2110 versus 0.2138.

This is due to the fact that the forecast date analysis has a more limited information set than the historical comparison, and therefore the two rankings reflect different definitions of a backtest. The more stringent rule causes the top methods to be more tightly grouped. Both analyses rely on past development outcomes that occurred before the held-out FY2025/26 benchmark.

Across the seven test years, equal-year coverage was 55.6% for 50% intervals, 84.4% for 80% intervals and 96.95% for 95% intervals. The average 80% interval width was about 0.490 ratio units. By year, 80% coverage ranged from 74.4% to 94.2%. Overall coverage was close to its target but varied from year to year. The Brier scores for the three outcome probabilities (below one, exactly one and above one) ranged from 0.0080 to 0.1698 across years.

The results package includes a finite-draw continuous ranked probability score (CRPS), but the transformed Student-t overrun component has an infinite first moment and therefore an

infinite population CRPS. For this model, distributional assessment uses interval coverage, interval width and class-probability scores; point assessment uses MAE and RMSE. Gaussian sensitivity models provide a finite-moment comparison under a different tail assumption.

Interval coverage varies across the historical experiments. The 66-case test achieved 83.3% coverage for nominal 80% intervals, the 54-case sensitivity achieved 90.7%, the earlier 852-record two-part model achieved 83.9%, and the later forecast-date model achieved 84.4% when years were weighted equally. Each value belongs to its own sample and model specification.

Among cases where actual spending exceeded the revised estimate, Bayesian MAE was 0.2403, compared with 0.2645 for the lag-plus-scope blend, 0.2545 for the year-and-scope median and 0.2661 for the scope median. Overall RMSE is lower for the scope median and lag plus scope blend, so the overrun comparison is reported as tail specific performance. The operational analysis that follows uses a different denominator and target.

For FY2026/27, the operational package applies a hierarchical Student-t model to 77 development records through FY2024/25, using 30 September 2026 as the forecast origin. The target is annual CoB reported actual spending divided by the original-gross-estimate proxy. Nine vote-level series receive forecasts. R1052 is excluded because its latest annual observation is FY2022/23, four years before the target and beyond the two-year maximum lag age.

Eight of the nine posterior medians are below 1.00, ranging from 0.855 to 0.994; the Teachers Service Commission median is 1.008. Table 7 reports the nine forecasts and the excluded R1052 series in a consistent order. All nine 95% intervals include 1.00.

Table 7. FY2026/27 operational execution-ratio forecasts.

Vote	Series	Median	80% interval	95% interval
R1095	State Department for Public Works	0.855	0.656–1.111	0.534–1.368
R1174	State Department for Trade	0.882	0.669–1.159	0.550–1.418
R1184	State Department for Labour	0.917	0.709–1.189	0.578–1.455
R2011	Kenya National Commission on Human Rights	0.961	0.747–1.235	0.610–1.516
R2051	Judicial Service Commission	0.992	0.766–1.286	0.626–1.574
R2071	Public Service Commission	0.994	0.771–1.276	0.634–1.559
R2091	Teachers Service Commission	1.008	0.783–1.294	0.639–1.571
R2111	Auditor General	0.977	0.761–1.258	0.624–1.531
R2121	Controller of Budget	0.890	0.689–1.144	0.564–1.401
R1052	Ministry of Foreign Affairs	No forecast	—	—

The settings of the operational model were: Gibbs chains of length four, burn-in of 2,000 iterations and 10,000 retained iterations per chain. The highest R-hat in the final run is 1.001304 and the minimum ESS in the run is around 1,020, both of which are within screening limits of the study. The precision for tail cannot be determined from the summary value as bulk and tail are not distinguished. Longer chains used in the retained run to provide more effective sampling of the intercept, which was not sampled sufficiently in a previous shorter run. Replication of the retained run resulted in the same CSV and JSON forecasts, which is evidence of replicability of the set computation but not evidence of convergence or predictive validity.

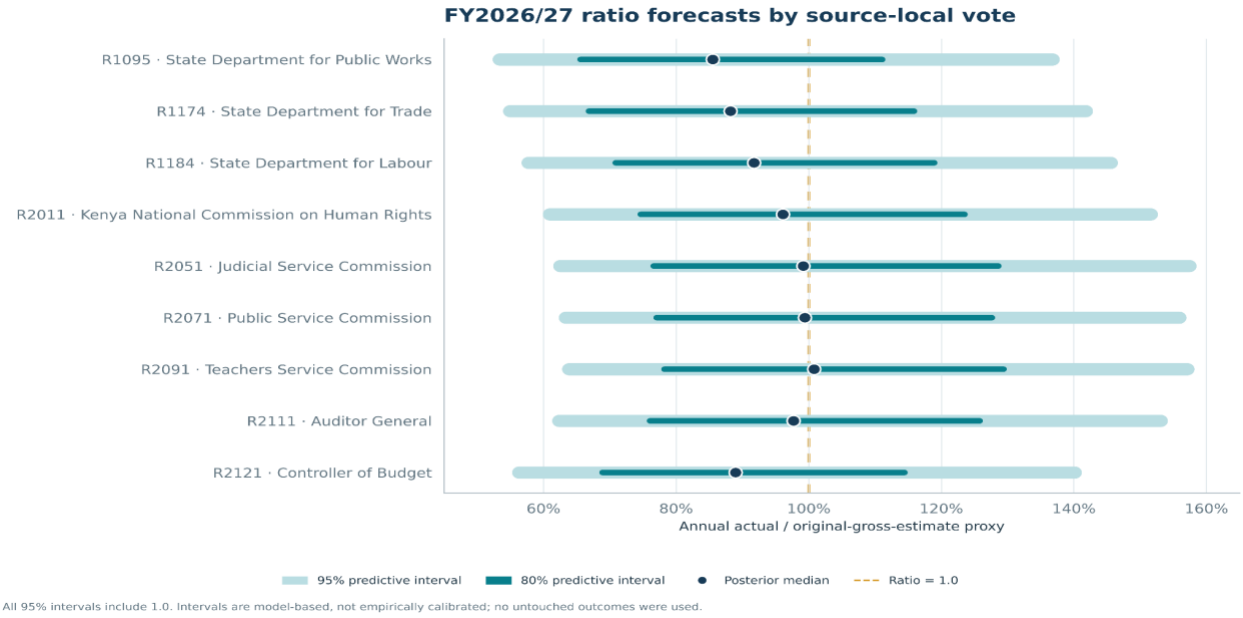


Figure 4.7: Posterior medians and 80% and 95% predictive intervals for the nine FY2026/27 vote forecasts.

These forecasts are plotted in Figure 4.7. 8 medians are below 1.00 and 1 is slightly above, with the lowest being 0.855 and the highest 1.008. The 80% and 95% intervals overlap 1.00 for all of the forecasts series and the level of uncertainty is high around the central estimates.

Nine series end in FY2024/25, while R1052 ends in FY2022/23. Currently, year 2025/26 is the evaluation year and is not part of the operational training history. Nine recent series are forecasted, and the older series R1052 is outside the maximum of 2 years of lag. Each series has a brief annual history, so that there is not much information about the series available to the hierarchy.

Because the model shares information across agencies, short or extreme histories are partially pooled toward the common distribution. This reduces the influence of a single extreme annual

ratio and also means that abrupt agency-specific changes may be smoothed. The FY2026/27 outcomes will determine how well that trade-off performs prospectively.

Alternative specifications were tested on the same 58 sequential forecasts. A hierarchical Gaussian sensitivity model, whose exponentiated predictive distribution has finite moments, produced equal-year MAE 0.1009, row MAE 0.1008, RMSE 0.1602 and 80% coverage of 87.9%. A pooled Gaussian model without series intercepts scored 0.1010 on equal-year MAE and 0.1582 on RMSE. These central estimates are close despite different pooling assumptions.

Prior sensitivity on the same 58 forecasts gave MAE 0.0960 under the reference prior and 0.0969 under a tighter prior. Under the reference prior, 80% coverage was 93.1%; under the tighter prior it was 87.9%, which happens to match the 87.9% coverage of the hierarchical Gaussian sensitivity despite coming from a different specification. For the FY2026/27 full-training fits, Gaussian and Student-t medians were close for most votes, with larger differences for R2051 and R2121, while the interval tails changed more than the central forecasts.

4.4 Corrected FY2025/26 benchmark

The FY2025/26 benchmark is a one-year-ahead hold-out of 50 questions from 25 MDAs, with one recurrent and one development case for each MDA. The target is the actual expenditure (CoB) for each accounting scope, which is not rounded. Version 0.2 fixes the recurrent reference outcomes (forecasts made prior to the fix remain unchanged).

Corrected reference data shows 47 cases of under execution, 1 of exact execution and 2 cases of overruns. An accuracy like this is sensitive to the accuracy of the dominant class that is under-executed (in this case the majority class), and as is Brier score. Benchmark results are independent of model fitting, language-model prompts and graph retrieval. The Bayesian training panel includes 888 records spanning 7 fiscal years and 7 source documents: 515 recurrent observations and 373 development observations, and the latest recorded availability date as of 9 October 2024.

Verified sampler diagnostics for the FY2025/26 benchmark fit. Run-specific sampler diagnostics for the held-out FY2025/26 Bayesian model (KG-BAYES-THREE-PART-ASOF-005-HOLDOUT-FY2025-26) were verified against the pre-specified screening thresholds ($\hat{R} < 1.01, \text{ESS} \geq 400$). For the under-one continuous component, maximum rank-normalized $\hat{R}$ was 1.001468 with a minimum bulk ESS of 1,674.71 and tail ESS of 3,177.54. For the over-one component, maximum rank-normalized $\hat{R}$ was 1.001481 with a minimum bulk ESS of

3,380.48 and tail ESS of 4,833.95. These diagnostics confirm chain convergence and parameter mixing for the fitted model prior to generating the 20,000 joint predictive draws, confirming that the observed interval under-coverage (64% coverage for nominal 80% intervals) is an empirical calibration property rather than a numerical sampling failure.

The comparison is between the forecast scored directly from the stored predictive distribution in Luna, and Luna without and with up to five graph assertions before the origin event. A fourth interface condition requires Luna to repeat the stored Bayesian output, and it is an implementation check in that Luna's numerical forecast is "locked in" by the Bayesian output. The three sets of generations were gathered for each question from the language-models, the first generation being the benchmark observation, with the repeats representing variability in the responses.

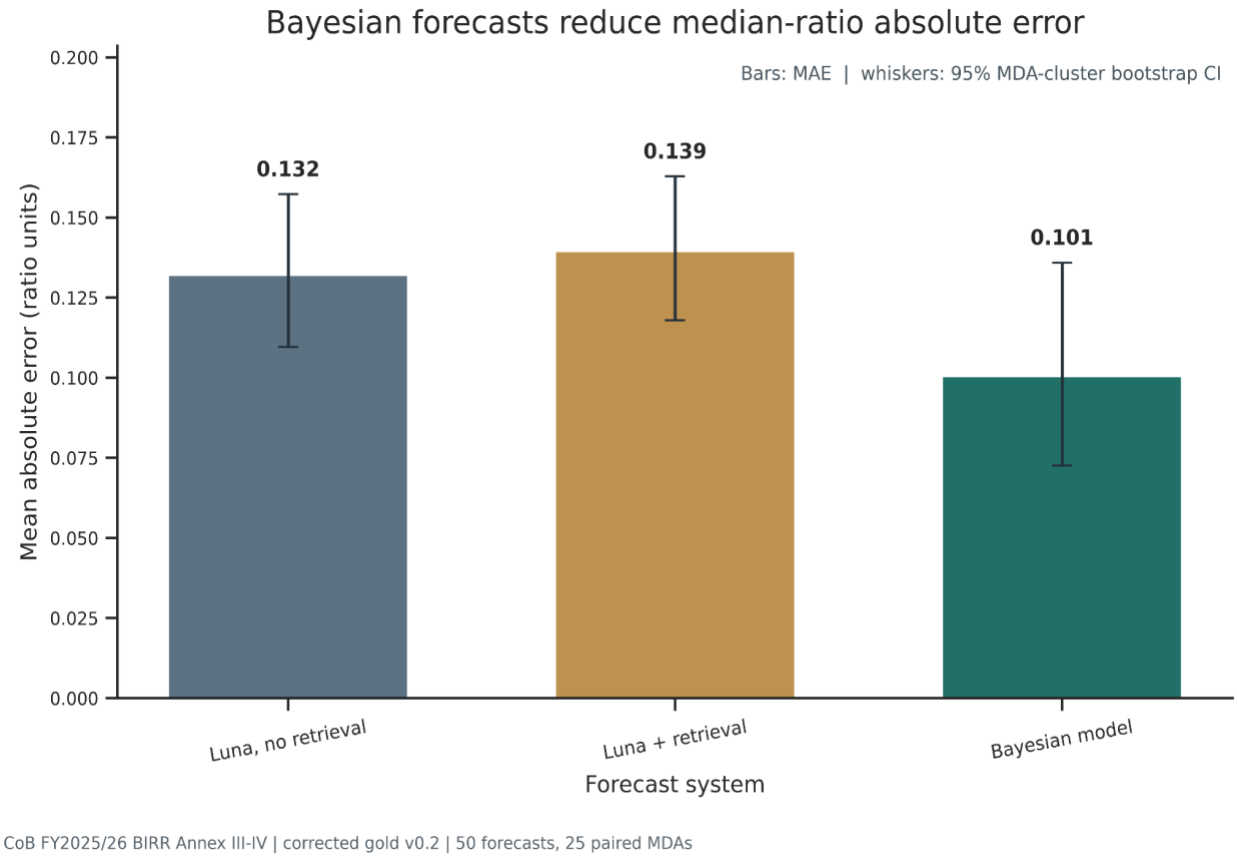


Figure 4.8: Mean absolute error and MDA-cluster bootstrap intervals for the corrected FY2025/26 benchmark.

The Bayesian model had MAE 0.1005, compared with 0.1321 for Luna without retrieval and 0.1395 for Luna with retrieval, corresponding to 23.9% and 28.0% lower MAE. Because recurrent and development cases from the same MDA are paired, uncertainty was estimated by resampling the 25 MDA clusters with both scopes retained together.

Table 8. Point-forecast metrics for the FY2025/26 benchmark.

Forecast arm	MAE	RMSE	Bias
Bayesian-only	**0.1005**	**0.1474**	**−0.0103**
Luna, no retrieval	0.1321	0.1666	−0.0584
Luna with graph retrieval	0.1395	0.1754	−0.0730
Luna retrieval plus Bayesian result	0.1005	0.1474	−0.0103

Table 9. Interval and class-probability metrics for the FY2025/26 benchmark.

Forecast arm	80% interval coverage	Mean 80% width	WIS-80	Multiclass Brier	Accuracy
Bayesian-only	32/50 (64%)	0.4505	**0.0679**	0.1148	94%
Luna, no retrieval	47/50 (94%)	0.4714	0.0796	0.1870	94%
Luna with graph retrieval	48/50 (96%)	0.5212	0.0833	0.1852	94%
Luna retrieval plus Bayesian result	32/50 (64%)	0.4505	0.0679	0.1148	94%

Lower values indicate better MAE, RMSE, absolute bias, WIS and Brier score. Coverage is interpreted relative to its nominal level and together with interval width. WIS combines interval width with penalties for misses, while Brier score measures error in the three class probabilities. The retrieval-plus-Bayesian row matches the Bayesian row because the interface reproduces the stored Bayesian forecast.

The identical results for the Bayesian-only and combined conditions demonstrate preservation of the stored forecast, rather than an independent forecasting gain from integration. Likewise, the absence of an improvement from retrieval applies to the tested model, prompts and limit of five retrieved assertions per question. It does not establish that retrieval is generally ineffective: source grounding and numerical forecasting are separate capabilities.

The paired MAE differences match the aggregate ordering. Bayesian minus Luna without retrieval was -0.0316, with a 95% cluster-bootstrap interval of -0.0463 to -0.0153 and a Holm-adjusted p-value of 0.0081. Against Luna with retrieval, the difference was -0.0390, with an interval of -0.0588 to -0.0168 and adjusted p=0.0162. On this FY2025/26 benchmark, the Bayesian point forecasts have lower average absolute error than both language-model conditions.

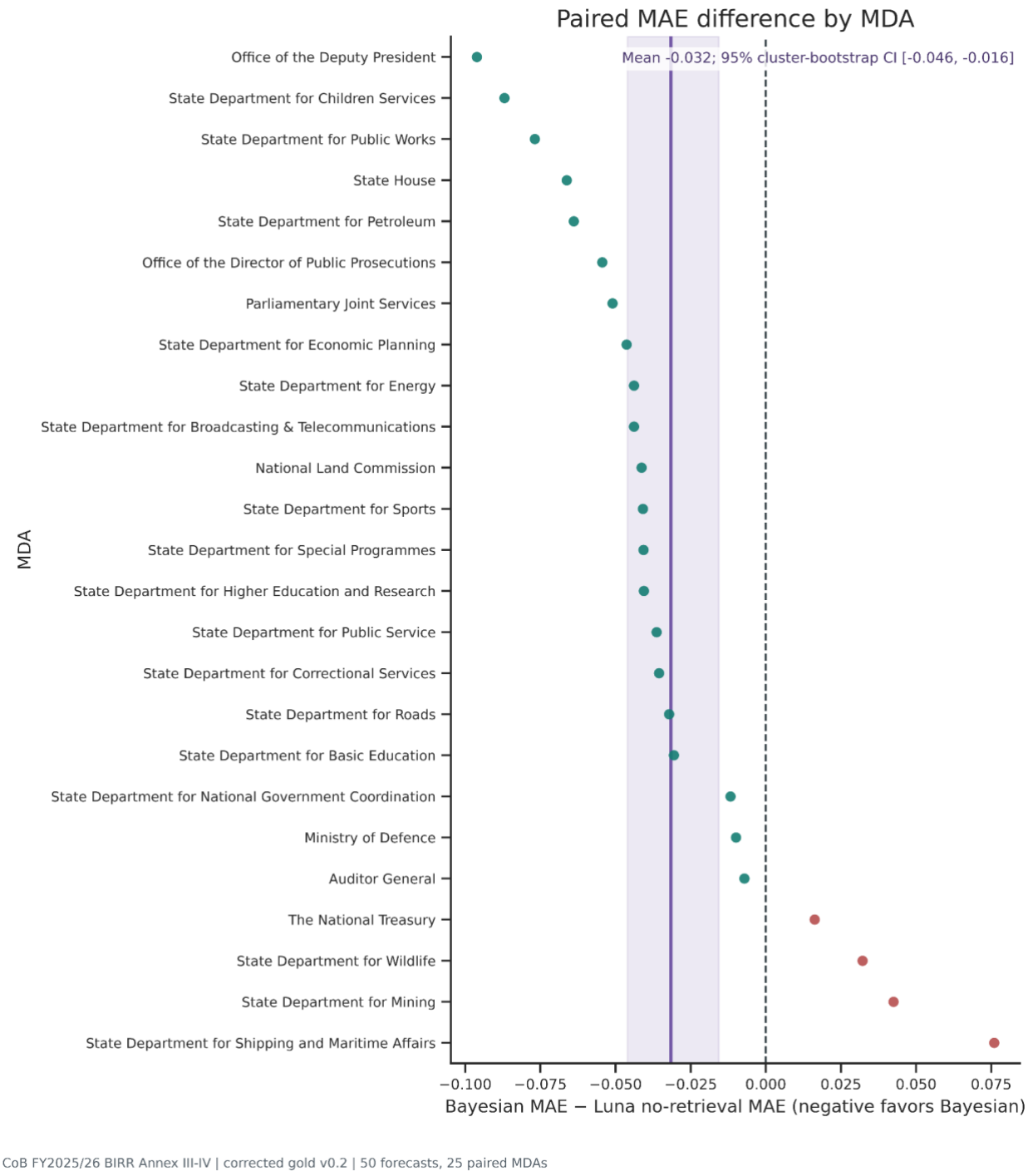


Figure 4.9:. MDA-level paired differences in mean absolute error.

The MDA-level contrasts in Figure 4.9 show how the aggregate difference is distributed across agencies. Most clusters favour the Bayesian forecast, while four favour Luna. Keeping recurrent and development observations together within each MDA preserves the paired structure used by the bootstrap and permutation tests.

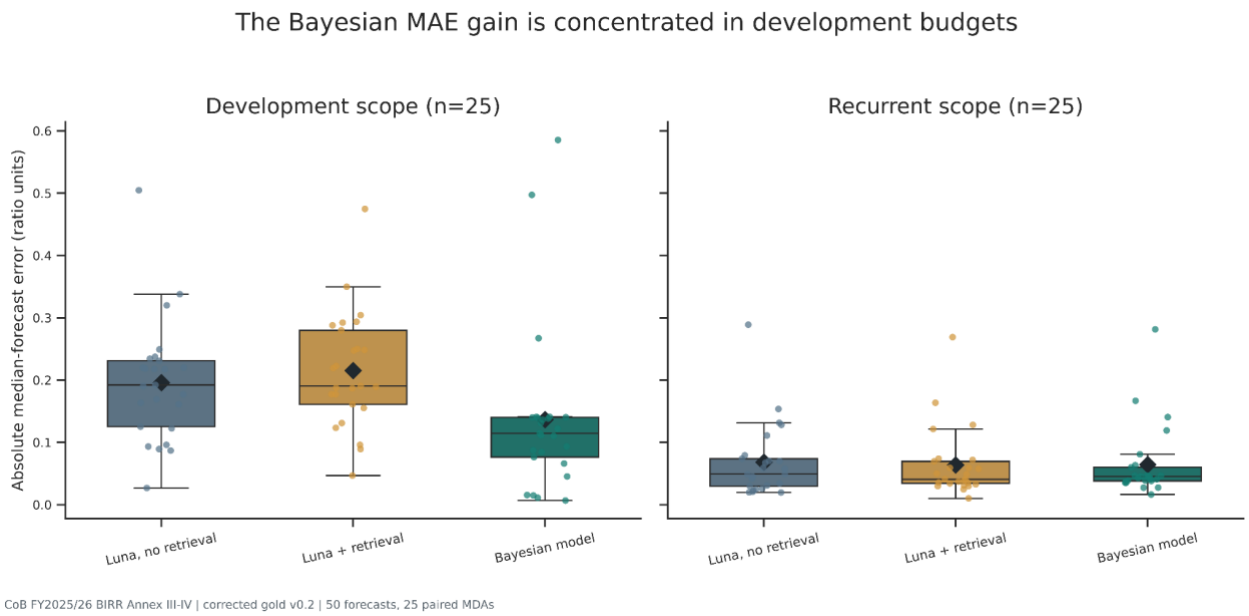


Figure 4.10: Forecast error by recurrent and development accounting scope.

Development-budget questions account for most of the aggregate MAE difference. Bayesian MAE is 0.1367 on development cases, compared with 0.1961 for Luna without retrieval and 0.2153 for Luna with retrieval. Recurrent MAE is 0.0643, 0.0681 and 0.0637 respectively. The overall difference therefore reflects the development scope more than the recurrent scope.

This holdout is more spread out with respect to development execution ratios, and the recurrent cases tend to be more clustered. The three-state model has fiscal-year and scope-level structure; MDA-specific longitudinal effects necessitate a more complete cross-year identity panel. If the same scope pattern occurs, further benchmark years will identify whether this is the case.

Across both scopes combined, mean bias was -0.0103 for the Bayesian model, -0.0584 for Luna without retrieval and -0.0730 for Luna with retrieval. The median absolute error of the Bayesian model was 0.0647. RMSE was 0.1474 for the Bayesian model, 0.1666 for Luna without retrieval and 0.1754 for Luna with graph retrieval. The difference between MAE and RMSE indicates the influence of a few high-magnitude errors.

There are some large increases in actual ratios as a result of the recurrent correction, notably for the Ministry of Defence and the State Department for Basic Education. The predictions were fixed prior to the correction, so the corrected scores are against the unchanged forecasts.

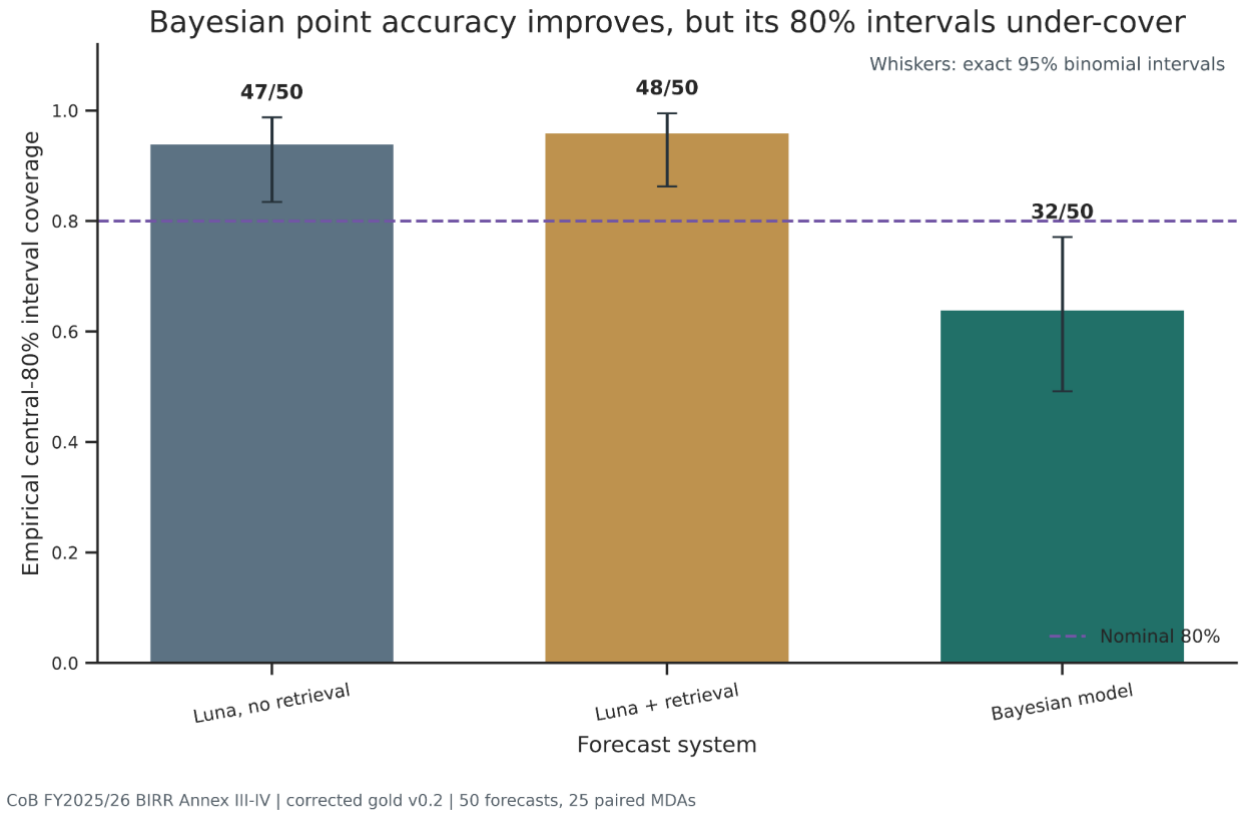


Figure 4.11: Empirical coverage of the reported central 80% intervals.

The Bayesian 80% intervals captured 32 of 50 outcomes (64%, exact 95% binomial interval: 49.2% – 77.1%, not taking into account MDA clustering). Luna without retrieval covered 47/50 (94%) with mean width 0.4714 and retrieval Luna covered 48/50 (96%) with width 0.5212; the Bayesian mean width was 0.4505. WIS-80 nevertheless favoured Bayesian at 0.0679, compared with 0.0796 and 0.0833. The Bayesian intervals are too narrow for their nominal 80% level although they are sharper.

A narrower interval may achieve a lower, and therefore better, WIS when its width and penalties for missed outcomes jointly produce a favourable score. However, the lower Bayesian WIS does not establish that the intervals attained their nominal coverage. In this holdout, the Bayesian model achieved lower point-forecast error than the tested language-model alternatives, while its nominal 80% intervals covered only 64% of outcomes. Point accuracy and interval calibration therefore support different conclusions.

The evaluation artifact records a finite-draw Bayesian CRPS of 0.0834. As in the historical three-state model, the transformed Student-t overrun component has an infinite first moment, which makes population CRPS infinite. The benchmark therefore uses WIS for the reported interval and reports point and class metrics separately.

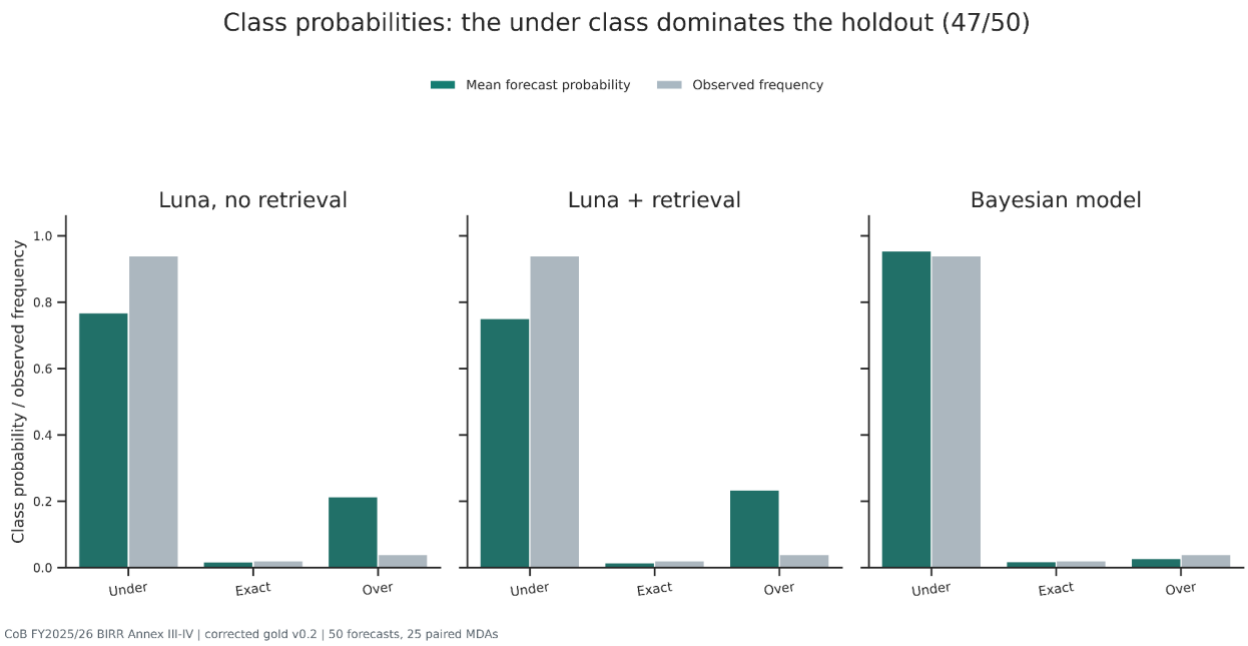


Figure 4.12: Mean predicted class probabilities and observed class frequencies for under, exact and over execution.

The three-class Brier score is 0.1148 for the Bayesian model, 0.1870 for Luna without retrieval and 0.1852 for Luna with retrieval; all three have 94% class accuracy. A prevalence reference calculated from the holdout has Brier 0.1144 and the same 94% accuracy. Because 47 of 50 outcomes are under-execution, the aggregate class metrics are dominated by that class, and the one exact and two over-execution outcomes provide little information about rare-state calibration. Because the prevalence reference was calculated from the observed holdout outcomes, it is a descriptive comparison rather than a forecast that could have been issued at the forecast origin.

The mean Bayesian probability assigned to under-execution is 0.9544, compared with an observed frequency of 0.94. This close aggregate match describes the dominant class in this benchmark. Assessment of the exact and over-execution probabilities requires additional years with more observations in those states.

The scores are the reobserved results when the recurrent parser was corrected. Gross and actual values are in cells 0 and 1 for the first component and cells 3 and 4 for the second. The earlier parser had a gross amount in cell 0, an execution rate in cell 1, and an actual amount in cell 4, thus intermingling a gross amount with an execution rate cell. The corrected mapping shows that the gross and actual spending cells are both cells 0 and 3, respectively. The mapping was done with a source-page review and checks against the reported percentages.

All 25 recurrent outcomes were rebuilt and the 25 development outcomes were unchanged. Re-scored the same 50 questions and frozen forecasts with version 0.2. The updated outcome hash was added to the benchmark manifest, and the original key was retained for audit history.

In this holdout, the MAE of the Bayesian model was about 24–28% lower than that of the two language-model conditions, mainly due to fewer errors on development-expenditure cases and its mean bias was -0.0103. Its nominal 80% interval had however only 64% empirical coverage. Thus, the point-forecast advantage does not provide reliable interval calibration. To assess whether the shortfall in coverage is a consistent problem from year to year and under different mixes of fiscal year accounting-scope problems, further calibration assessment and some additional held-out years are required.

4.5 Cross-experiment synthesis

The results indicate that a differentiated comparison is more appropriate than an overall ranking of the architectures. The seven-year historical study resulted in an equal-year MAE of 0.1134 for the Bayesian model and 0.1144 for the scope median, with the scope median being lower in terms of RMSE. This takes away some of the justification for complexity as far as point accuracy. In the corrected FY2025/26 benchmark, the paired cluster-bootstrap intervals for Bayesian MAE were below zero, whereas those of the tested conditions with language models were not, and the former was lower than the latter, partly due to cases on development expenditure. The held-out advantage is for different sets of comparators: Held-out advantage vs language models is not determining advantage over scope median on the same 50 targets.

The primary constraints are related to temporal span, class balance, source timing and review of identity and extraction of longitudes. The bench mark is 25 MDA clusters for one fiscal year. 47 out of 50 outcomes are under-execution with just 3 under the remaining two classes. Availability dates are based on documentary evidence like tabling or repository records and may differ from the first public release of a specific PDF version. Detailed cell level source review is focused on the parliamentary correction and many recurring labels are still not cross-year verified. These features bound the area covered by the present estimates, and the principal areas of further validation.

The variation between experiments is partly due to different denominators. The operational forecast includes a comparison of actual spending to an original gross estimate proxy; the graph comparison and FY2025/26 benchmark compares spending to the revised gross estimate

reported by CoB. The denominators are the various stages of the budget process and the various target definitions, thus each score has to be understood in the light of its respective target definition.

The timing of the forecasts varies among studies. The first graph test employs an historical split without a full reconstruction of the source-availability. For the later graph test only training is possible on records that have been documented to be available before 1 July in the forecast year. The FY2025/26 benchmark is based on a fixed forecast date and a separate outcome set that is held. Its effective period is primarily based on parliamentary tabling or from the repository and spans one fiscal year.

Cross-year identity is another source of variation. Repeated labels provide provisional lag features, while the operational model uses a reviewed vote crosswalk. Both approaches contain shorter or less fully verified longitudinal histories than a canonical entity panel. More complete source-backed links through renames, splits and mergers would allow richer entity-specific models.

The results also differ by accounting scope. Bayesian forecasts improve more on development than recurrent questions in the corrected benchmark. Overall scores average across these differences. More detailed models by agency or programme will require reliable links between the same budget items over time.

Sampling diagnostics and predictive performance answer different questions. R-hat and ESS describe numerical sampling of the fitted posterior; held-out error and coverage describe forecast behaviour. Nominal 80% intervals achieved 84.4% equal-year coverage in the seven-year graph study and 64% in the FY2025/26 benchmark. The benchmark intervals are narrower and have lower WIS than the language-model intervals, but their coverage is below the stated level.

Student-t tail behaviour affects distributional scoring. Exponentiating Student-t log errors gives the ratio an infinite first moment, which makes population CRPS infinite. Therefore, what is used here is the median error, interval width and empirical coverage for these models. Class metrics are also read alongside prevalence because 47 of 50 benchmark outcomes fall in the under-execution class.

The graph represents institutions together with source-supported programme and budget-line structure. When a table supplies an amount, measure, year, accounting scope and page, the graph can store that budget line with its documentary evidence. Treasury and Parliament contribute budget-authority claims, CoB contributes implementation and outturn claims, and recommendations or revised estimates retain their own measure types.

Current graph-wide forecasts operate at vote or agency level because systematic mappings from CoB actual expenditure to detailed Treasury and Parliament programme lines are still incomplete. The 966 modelled claims are annual CoB records by vote or agency. Programme- or budget-line forecasting therefore depends on source-backed mappings that connect each budget-authority line to its corresponding actual expenditure at the same level of aggregation.

Most of the graph assertions are obtained from automated extraction and detailed source-cell review is only performed in selected parliamentary material. A larger sample examined could enable quantification of the impact of source verification on retrieval accuracy and inputs to downstream forecasts.

Controlled evidence-to-output continuity is the best argument for the integrated approach. The graph retains the meaning and source of a claim, the statistical piece of the graph creates a clearly defined forecast, and the interface maintains that forecast when communicating it. The apparent equal performance of the combined and the Bayesian-only scores corroborates the consistency of the numbers rather than the additional forecasting value of combination. Similarly, although retrieval does not decrease forecast error, this does not detract from its evidential value; it just means it cannot be said that access to the graph alone improves prediction.

Weaknesses in components are also carried over into integration, but not solved. The combined interface delivers coverage of the nominal 80% interval that is consistent with the coverage in the Bayesian model (64%). Run-specific diagnostics for the final FY2025/26 fit show MCMC convergence of the parameters ($\hat{R} \leq 1.0015$, bulk ESS ≥1,674, tail ESS ≥3,177). Model inputs can be impacted even before forecasting by source extraction and incomplete identity problems. Traceability, therefore, should expose these limitations and not put false trust in the numbers that it provides. The evaluation did not separate out the causal effect of the graph representation from an equivalent relational representation, nor did it measure costs of maintenance or quality of user decisions for the bounded agent, nor determine the effect of the

deterministic interface. The following tests, which are missing, are those of whether integration is worth the extra effort than simpler alternatives.

The framework is advanced as a technically feasible system for conducting traceable advisory analysis, rather than a system of superior forecasting or decision-making. The operational use in FY2026/27 provides nine recency checked ratio forecasts, with outcomes to be determined. Future tests will have to be conducted to determine if the integrated workflow provides enough benefit for the evidence toward verification and interpretation versus its operational complexity while maintaining simple forecasting benchmarks and human oversight.

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter summarizes the findings of the study, focusing on the goal of a designing, implementing and evaluating a probabilistic neuro-symbolic agentic AI platform for decision support in the Kenyan budget process. It provides a summary of the results obtained for each specific objective, concludes by addressing the problem of fragmented budget information, which is technically difficult and fraught with uncertainty, and suggests some recommendations for policy and further study. Thus, the chapter links the findings from evidence in the knowledge graph, the Bayesian forecasting experiments, and the held-out platform evaluation to the pragmatic needs for transparent, traceable and accessible budget decision support.

5.2 Summary

5.2.1 Summary for Objective 1

The first goal was to create a neuro-symbolic representation framework to model the budget information of Kenya. The study built a source-centred knowledge graph that retained the institutions, documents, document versions, fiscal periods, monetary values, measures, accounting scope, and document-level and table-level provenance. The principal release included 21,618 assertions from the registered document base, and the analytical graph had 2,036 assertions from the Controller of Budget for expenditure in annual reporting periods and in-year reporting periods. Of 39 documents sampled, 130 assertions were sampled with 130 conforming records found for structural validation and source correctness was done separately. 46 rows were detailed reviewed in the parliament and the totals of their components were reconciled, whilst nine transcription errors were corrected or documented. Through the resulting framework, it was found that estimates, revisions, recommendations, expenditure in the heterogeneous treasury, parliament, controller of budget and auditor-general evidence can be combined without the evidence collapsing into one value. Longitudinal identity was incomplete, however, as only 131 of 357 labels were matched across years, and many records were not deterministically matched across years.

5.2.2 Summary for Objective 2

The second objective was to create a framework of probabilistic reasoning using a Bayesian method to model uncertainty and forecast budget trends. The three-state Bayesian model under

the forecast-date rule had an equal-year MAE of 0.1134, a row MAE of 0.1158 and an RMSE of 0.2138 on the 852 historical forecasts. Its point accuracy was numerically close to that of the scope-median baseline, which achieved an equal-year MAE of 0.1144 and a lower RMSE of 0.2110. The Bayesian model did provide predictive intervals and probabilities for under-, exact- and over-execution, however. Across the historical test years, nominal 80% intervals achieved 84.4% equal-year coverage, with variation between fiscal years. For the operational run in FY2026/27, 9 of the series were recent, 8 posterior medians were below 1.00, all the 95% intervals contained 1.00 and one stale series was excluded according to the recency rule. The sampling diagnostics were within the specified thresholds, indicating computational convergence, and the large intervals indicated that the uncertainty was material.

5.2.3 Summary for Objective 3

A third objective was to assess the platform for data-based decision support in the Kenyan budget. The evaluation evaluated four properties that are interdependent: source linked graph retrieval, reproduction of the outputs of the frozen Bayesian models through the agentic interface, predictive performance on the corrected FY2025/26 hold-out, and controlled abstention when evidence was not available. Using the Bayesian forecast on the 50 questions benchmark with the 25 MDAs yielded an MAE of 0.1005, while the MAE of Luna without retrieval was 0.1321 and 0.1395 with graph retrieval. Most of the aggregate MAE advantage arose from lower errors on development-expenditure cases, while recurrent-expenditure MAE was numerically close across conditions. The Bayesian forecast also had the lowest RMSE and smallest absolute bias, but its nominal 80% predictive intervals covered only 64% of outcomes, indicating a need for further calibration assessment. The difficulty of assessing rare states was caused by the high number of outcomes under-execution, resulting in all conditions achieving 94% class accuracy. The agentic interface reproduced the stored Bayesian outputs. Graph retrieval supplied source evidence but did not improve numerical forecasting relative to the no-retrieval condition in this benchmark. The results thus endorse the platform as a governed orchestration layer for evidence retrieval and validated forecast/explanation, not as a stand-alone language model that can forecast.

5.3 Conclusions

Objective 1. The study concludes that a neuro-symbolic, source-centred knowledge graph can structure disjointed budget evidence from Kenya in a coherent and traceable manner without losing the distinction between the different budget values such as estimates, actual amounts,

revisions, recommendations and expenditure. This directly responds to the research problem of budget information widely distributed in institution portal and is hard to compare. The graph's provenance and validation mechanisms make it possible to trace represented claims to their documentary sources and inspect their accounting meaning. However, the study did not directly establish improved user understanding or transparency relative to stand-alone reports and dashboards. In addition, identity is incomplete and the review of sources is uneven across years, so that the claims are strongest at document level and at vote and agency level, but less so at the fully harmonised programme or budget-line level.

Objective 2. Use of the Bayesian probabilistic framework provides a principled, rigorous way of modeling the trend of budget execution and quantifying uncertainty in the budget not just in terms of point predictions. In the past it provided the same level of point accuracy as simple scope-based baselines and, in addition, delivered other secondary information, such as predictive intervals and discrete probabilities of execution-states. The system ensured governance controls, stopping predictions being withheld if input data did not meet the importance or recency rules. No MCMC non-convergence was found as a possible source of model error for the FY2025/26 held-out benchmark, as sampler diagnostics for the run indicated convergence ($\hat{R} \leq 1.0015 \leq 1.0015$, bulk ESS $\geq$1,674, tail ESS $\geq$3,177)). The benefits of the posterior distribution are not obvious, however: point prediction errors on held out development budgets were lower when using the posterior medians, and posterior predictive intervals of nominal 80% coverage achieved 64% coverage. The results show that there is a need to monitor convergence of the MCMC algorithm routinely with the aim of interval calibration, and that in the context of Bayesian decision support, it is important to account explicitly for and impose temporal constraints when verified MCMC convergence and point-forecast accuracy do not imply interval calibration.

Objective 3. The integrated framework was technically feasible, and it was able to display the ability to connect source evidence, statistical forecasts and bounded natural-language communication. The two language-model conditions tested did worse than the stored Bayesian forecasts in terms of average absolute error, mostly on development-expenditure cases, on the corrected FY2025/26 benchmark. The combined interface did not improve or add value to those forecasts, either in terms of the numerical accuracy or interval under-coverage. It was discovered that, in the testing configuration, the provision of a source context did not enhance graph retrieval for prediction. It is not because of a demonstrated additive forecasting gain that

the framework is vindicated, but because it is inspectable, it breaks down responsibilities, and it is continuous between evidence and output. It is not tested for relative advantage to a similar database, statistical model and deterministic interface; interval recalibration is only empirically tested ($\hat{R} \leq 1.0015$, bulk ESS ≥1,674, tail ESS ≥3,177), and just one held-out fit is tested for run-specific MCMC convergence. The results support the necessity for traceable advisory analysis and in the future the usability, user understanding as well as the quality and cost-effectiveness of the decision should be assessed.

5.4 Recommendations

5.4.1 Recommendations for Policy

National Treasury, Parliament, the Controller of Budget and the Office of the Auditor-General should share and use a common and machine-readable budget data format that will have constant identifiers for institutions, votes, programmes, fiscal period, scope of accounts, type of measure and version of document. Every published value should be accompanied by its source location, publication date, revision status and unit to ensure that claims will not overrule each other or be confused with revised claims. There should be a joint data-governance mechanism to ensure crosswalks on institutional renames, splits and merges and to publish versioned correction notices. Public institutions should also implement an assurance process for AI-assisted extraction which includes automated schema and ontology, arithmetic and duplicate checks, and risk-based human follow-up to tables that are ambiguous and claims that are of high impact. Platforms to use operationally should include the denominator, forecast origin, posterior median, interval, evidence links, and explicit abstention reason. The agentic layer should only be used for approved queries for read-only data, validated model endpoints, audit logs, role-based access, model/graph versioning, periodic bias and calibration check, and human confirmation of consequential decisions. A phased pilot should involve parliamentary analysts and public finance analysts first, regular public reporting of performance by recurrent and development scope should be made, and a straightforward public interface should be built to explain not obscure uncertainty.

5.4.2 Recommendations for further research

Future studies should replicate the held-out evaluation over a multiple year evaluation of fiscal years and multiple MDAs and county-government budgets to determine whether the point-forecast advantage and development-scope pattern continues through FY2025/26. If a larger sample is available, deliberate effort should be made to obtain exact- and over-execution cases in order to credibly calibrate the probability of rare states. Research should create source-

backed longitudinal identifiers, not repeated labels, via institutional renames and mergers and programme restructurings. Further research should focus on the comparison of other specifications for the tails of the distributions, the use of hierarchical dispersion, conformal and post-hoc interval calibration, and denominator-specific models for original, enacted, and revised budgets. A larger independently reviewed knowledge graph sample should be used to numerically estimate the errors in the extraction and entity-linking and provenance, and their impact on the forecasts. Controlled user studies with Parliamentarians, Analysts and Citizens should include measures for task accuracy, comprehension, trust, usability, and decision time, as well as comparisons to other reports and dashboards. Other topics for future research are multilingual and low bandwidth interfaces as well as real-time or more frequent updates where authoritative data is available, robustness to document changes and quick manipulation, and governance needs for privacy, security, accountability and auditability. Finally, possible deployments must be evaluated for their ability to enhance the quality of budget scrutiny and public participation without diminishing or altering any automation bias, or replacing legislated decisions with the results of models.

5.5 Contribution to Knowledge

The study delivers design knowledge and empirical evidences in a context-specific manner for an AI-assisted budget-analysis workflow. The three gaps from the Section 2.6 are correlated to the subsequent contribution of representation, forecasting and integration: The source-centered assertion and verification workflow of Section 2.6 provides for preservation of documentary meaning; The temporally constrained comparative evaluation of Section 2.6 provides for addressing uncertainty with respect to predictive performance; The tested preservation of stored forecasts of Section 2.6 provides for controlled communication. The methodological contribution is a transversal component, connecting source provenance, information cut-offs and protected scoring. In combination, these works define and evaluate the operation of source-grounded representation, conditional probabilistic forecasting and bounded communication in the context of explicit accounting and temporal constraints. They provide only partial answers, and they don't close the loopholes: graph-wide source correctness and reliable held-out calibration and comparative benefits to users remain unresolved. The study does not include a new general-purpose algorithm for AI, and the combination of the existing algorithms is not considered novel enough to warrant a new algorithm.

Representation contribution: Objective 1. The study responds to the representation gap through operationally implementing a source-centred assertion model for national-budget data from multiple sources in Kenya, where the version, fiscal period, scope of accounting and measure type is maintained and revisions and conflicting assertions are distinguished. This estimated release comprised 21,618 assertions, 130 assertions were sampled, all of which passed structural tests, 46 rows underwent targeted parliamentary review and were reclassified, and nine transcription errors were corrected or documented. These findings are in support of an implemented and evaluated representation and verification workflow for a domain rather than an invention of provenance or proof for the factual correctness in the graph. Structural conformity and documentary correctness are still two separate criteria of assurance. The institutional-continuity requirement is not fully satisfied: repeated labels are only provisional and is not full harmony of the identity for the years.

Forecasting contribution: Objective 2. The study provides a solution to the predictive-evidence gap by showing that there are cases, cases in which the budget-forecasting environment is the one described, where they will not overlap: competitive point accuracy and reliable uncertainty estimates. It provides rules, based on observation from the graph, and an explicit denominator and forecast date. Bayesian equal-year MAE is 0.1134 for 852 historical forecasts and scope median has lower RMSE of 0.1144. The benchmark for BM was 0.1005 for Bayesian MAE, and 0.1321 no retrieval, 0.1395 with retrieval, for the tested language model, mostly due to development-expenditure cases, on the corrected FY2025/26 benchmark. The final held-out fit did not indicate any problems with the continuous components ($\hat{R} \leq 1.0015$; bulk ESS ≥1,674; tail ESS ≥3,177). But nominal 80% intervals were found to have only 64% empirical coverage. The contribution focuses on the comparative-performance issue and corroborates the stability of the MCMC sampling for the reported samples, but leaves the need for interval recalibration as an ongoing methodological requirement. It shows that one needs to be careful when using Bayesian prediction: If one fails to perform an empirical validation of convergence of the parameters, of the competitive nature of the point forecasts, and of the validity of the uncertainty quantification, one will not obtain consistently good results.

A qualification of the forecasting contribution is also provided: execution-state probabilities. The benchmark score was 0.1148 and the scores without retrieval, and with retrieval (language model tested) were 0.1870 and 0.1852, respectively. Class accuracy was 94% for each of the conditions, 47 out of 50 results were under-execution. But, the holdout-prevalence reference

was not a forecast from the forecast origin, but rather a holdout prevalence with actual holdout outcomes, and thus had a Brier score of 0.1144. These findings offer a partial answer to the class-probability part of the predictive-evidence gap: they can perform well on this sample when it comes to comparative probability scoring, but they will not be reliable when it comes to one exact-execution and two over-execution results.

Integration contribution: Objective 3. The study proposes a numerical method to maintain the continuity of forecasts within a bounded orchestration pattern, where the source retrieval, statistical prediction and natural-language communication are decoupled to cope with the controlled-integration gap. The condition with only graph retrieval resulted in no improvement for the numerical forecasting in the tested condition, and the condition with combined condition reproduced the stored Bayesian outputs. This leads to the following principle for the assessed statement of the design: access to evidence through language must not be as a substitute for a separately assessed predictive model and a distinction between a numerical forecast and a generated explanation must be kept. The contribution therefore focuses on the technical relationship between evidence and stored forecasts, but does not set up on its own the faithfulness of all generated explanations to communicate all limitations. There is no independent forecasting gain in the output preservation and technical feasibility, and no additional benefit seen to the non-expert user of understanding or to any benefit over the use of a deterministic interface.

Methodological contribution in the achievement of the three objectives. The research outlines a traceable evaluation process that covers versioned source manifests, page-level provenance, eligibility for the forecast date, protected outcomes, frozen predictions, corrected scoring keys and clustered paired comparisons. If the recurrent outcome key is corrected, and rescoring is carried out on the unchanged predictions, then it differentiates between source-transcription correction and model improvement. The methodological importance lies in this combination of controls, and the separate reporting of the quality of representation, point error, calibration and computation diagnostics. Completed computational audit trails, with verified run-specific diagnostics for the final held-out fit, consider mixing of model parameters and calibration of predictive intervals as distinct properties. The workflow offers guidance for similar documentary applications; any transfers outside of the Kenyan records and configurations evaluated will have to be replicated, not assumed.

REFERENCES

Aboucaya, W., Guehis, S., & Angarita, R. (2023, May). Building online public consultation knowledge graphs. In *Text2KG 2023: International Workshop on Knowledge Graph Generation from Text, Co-located with the ESWC 2023* (p. 12).

Adhikari, G. R. (2020, August). *Strategies for Selecting a Research Topic*. *Mining Engineers' Journal, 22*(1), 27–31.

Aisopos, F., Jozashoori, S., Niazmand, E., Purohit, D., Rivas, A., Sakor, A., ... & Vidal, M. E. (2023). Knowledge graphs for enhancing transparency in health data ecosystems. *Semantic Web*, *14*(5), 943–976.

Alshahrani, S., Fong, S., & Mohammed, S. (2022). Knowledge graph-based data integration for e-government services. *Government Information Quarterly, 39*(4), 101721.

Atzori, L., Iera, A., & Morabito, G. (2023). Semantic technologies and AI for smart governance: A review. *Future Generation Computer Systems, 140*, 1–14.

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. National Institute of Standards and Technology.

Bajeti Hub. (2024). *Open budgets, open dialogue: County budget transparency survey 2024*. Bajeti Hub.

Barngetuny, J. (2024). *Rethinking public budget in Kenya*. *Africa Journal of Public Sector Development & Governance, 7*(1), 22–50.

Center for Strategic and International Studies (CSIS). (2026). *Lost in Definition: How Confusion over Agentic AI Risks Undermining U.S. Governance Frameworks*. Center for Strategic and International Studies.

Chen, X., Liu, Y., & Zhang, H. (2024). Probabilistic knowledge graphs: Models and applications in decision support systems. *IEEE Transactions on Knowledge and Data Engineering, 36*(2), 567–580.

Chen, X., Liu, Y., & Zhang, W. (2022). Knowledge graph construction and reasoning in intelligent systems: A survey. *IEEE Access*, 10, 123456–123478.

Chen, X., Wang, H., & Li, J. (2022). Probabilistic knowledge graphs: Models and applications. *IEEE Transactions on Knowledge and Data Engineering*, 34(7), 3352–3366.

Cifuentes-Silva, F., Fernández-Álvarez, D., & Labra-Gayo, J. E. (2020). National budget as linked open data: New tools for supporting the sustainability of public finances. Sustainability, 12(11), 4551. https://doi.org/10.3390/su12114551

Competition and Markets Authority. (2026). *Agentic AI and consumers*. GOV.UK.
Creswell, J. W., & Creswell, J. D. (2021). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

d'Avila Garcez, A., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2021). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Foundations of Logic and Artificial Intelligence*, 8(1), 1–28.

Detthamrong, U., Laochankham, S., Emperador-Garnace, X. R., Jitsaeng, K., Chaichuay, V., Chansanam, W., & Li, C. (2025). Thematic shifts in e-governance research: From foundational frameworks to emerging technologies. *Social Sciences & Humanities Open*, *12*, 101888.

Drechsler, A., & Hevner, A. (2022). Knowledge paths in design science research. *Foundations and Trends in Information Systems, 6*(3), 171–243. https://doi.org/10.1561/2900000028

Dwivedi, Y. K., et al. (2023). Artificial intelligence in government: Applications and challenges. *Government Information Quarterly, 40*(1).

Fenuku, Q. (2024). Budget transparency, accountability and governance in Sub-Saharan Africa. *European Modern Studies Journal, 8*(3), 44–56.

Floridi, L., Cowls, J., & Beltrametti, M. (2022). AI governance and ethical implications. *AI & Society*, 37(1), 1–17.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2022). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, 28(3), 1–19.

Franklin, S., & Graesser, A. (2022). Is it an agent, or just a program? A taxonomy for autonomous agents. *Intelligent Agents and Multi-Agent Systems*, 45(2), 21–35.

Garcez, A. d'Avila, & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56, 125–162.

Garcez, A. d'Avila, et al. (2022). Neuro-symbolic AI: The state of the art. *Artificial Intelligence Review*, 55(7), 1–51.

Garcez, A. d'Avila, Lamb, L. C., & Gabbay, D. (2022). Neural-symbolic learning systems: Foundations and applications. *Artificial Intelligence*, 310, 103756.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2021). Bayesian workflow. *Statistics and Computing, 31*(1), 1–28.

Gikibi, C. N., & Kipchumba, H. E. (2025). The impact of public awareness and participation on the Kenyan National Assembly legislative process. *Journal of Public Policy and Governance*, 15(1), 77–92.

Guariso, D., Guerrero, O. A., & Castañeda, G. (2023). Automatic SDG budget tagging: Building public financial management capacity through natural language processing. Data & Policy, 5, e31. https://doi.org/10.1017/dap.2023.28

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys, 54*(4), 1–37.

Jiang, B., Yang, S., & Wu, J. (2022). Knowledge graph construction for public sector data integration and governance. *Information Systems, 103*, 101878.

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.). Online manuscript released January 6, 2026.

Kenya. (2010). *The Constitution of Kenya, 2010*. National Council for Law Reporting.

Knublauch, H., & Kontokostas, D. (Eds.). (2017). Shapes Constraint Language (SHACL). W3C Recommendation. https://www.w3.org/TR/shacl/

Koller, D., & Friedman, N. (2022). *Probabilistic graphical models: Principles and techniques* (Updated ed.). MIT Press.

Martin, M. M. (2024). The effects of public participation on budget transparency in Kenya. *European Journal of Economic and Financial Research, 8*(2), 79–94.

Mergel, I., Edelmann, N., & Haug, N. (2021). Defining digital transformation: Results from expert interviews. *Government Information Quarterly, 38*(1), 101585.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

Nikiforova, A., & McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics, 58, 101539. https://doi.org/10.1016/j.tele.2020.101539

OECD. (2025). *Governing with Artificial Intelligence: The state of play and way forward in core government functions*. OECD Publishing, Paris.

Ojino, R., Mutunú, G., & Githaiga, G. (2024). Civic tech in Kenya's anti-Finance Bill protests. *Kenya ICT Action Network*.

Open Government Partnership. (2023). *Open Data for Development (KE0034)*. Open Government Partnership.

Parliament of Kenya. (2021). *3rd Annual National Parliamentary Symposium: Strengthening public participation in the budget process*. Parliament of Kenya.

Paulheim, H. (2023). Explainable artificial intelligence with knowledge graphs. *Semantic Web Journal, 14*(2), 1–18.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M.

(2024). ToolLLM: Facilitating large language models to master 16,000+ real-world APIs. *ICLR 2024 Spotlight*.

Schmitz, C., Rystrøm, J., & Batzner, J. (2025). Oversight structures for agentic AI in public-sector organizations. arXiv. https://arxiv.org/abs/2506.04836

Serderidis, K., Konstantinidis, I., Meditskos, G., Peristeras, V., & Bassiliades, N. (2024). *d2kg: An integrated ontology for knowledge graph-based representation of government decisions and acts*. *Semantic Web, 15*(5), 1677–1699.

Sheng, J., Sun, Z., Yang, B., & Zhang, A. L. (2025). Generative AI and asset management. *Review of Financial Studies*. Forthcoming.

Sun, T. Q., & Medaglia, R. (2024). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 41*(1), 101912.

Toroghi, A., & Sanner, S. (2024). Bayesian inference with complex knowledge graph evidence. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 20550-20558. https://doi.org/10.1609/aaai.v38i18.30040

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2023). A survey on large language model based autonomous agents. *arXiv*, arXiv:2308.11432.

Wooldridge, M. (2021). *Foundations of multiagent systems* (2nd ed.). Wiley.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *ICLR Proceedings*.

Zeginis, D., & Tarabanis, K. (2024). An event-centric knowledge graph approach for public administration as an enabler for data analytics. *Computers*, *13*(1), 17.

Zhang, Q., Chen, M., & Li, X. (2023). Bayesian-enhanced knowledge graphs for intelligent decision support. *Expert Systems with Applications, 213*, 118983.

Zheng, L., Chen, G., & Li, Y. (2023). Autonomous AI agents: Architectures and applications. *ACM Computing Surveys*, 55(6).

Zheng, Z., Song, Y., & Leung, C. (2023). Agentic artificial intelligence: Autonomous systems for complex decision-making. *ACM Computing Surveys*, 56(4), 1–36.

Zuiderwijk, A., Shinde, R., & Janssen, M. (2022). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? *International Journal of Information Management, 62*, 102393.

APPENDICES

APPENDIX A. IMPLEMENTATION IDENTIFIERS AND REPRODUCIBILITY DETAILS

The main chapter uses short descriptive names for files, graph releases and model runs. Appendix A retains the internal identifiers needed for reproduction. The application routes plain-language requests through typed tool schemas, validates year and entity fields before execution, and records hashes of the evidence and forecast artifacts used for benchmark questions. Retrieval and forecasting read fixed local artifacts, while the language-model layer explains the resulting structured outputs.

Table A1. Reproducibility aliases for identifiers omitted from the main prose.

Short name	Internal identifier / path	Purpose
Current source inventory	data/manifests/source_manifest.csv	Versioned documentary inventory
FY2025/26 evaluation report	DOC-B713250A73D03C27	Held-out CoB benchmark source
Frozen retrieval graph	GR-AUTO-DEVELOPMENT-KG-014	Development graph used by benchmark retrieval
Historical three-state model	KG-BAYES-THREE-PART-ASOF-004	Forecast-date-constrained historical Bayesian model
FY2025/26 Bayesian benchmark run	KG-BAYES-THREE-PART-ASOF-005-Houldout-FY2025-26	Verified holdout Bayesian model run record (Bayesian_asof_run_v0.1.json)

Runtime configuration, routing schemas, audit-record fields, retry logs and credential handling are preserved in the project artifacts and run manifests. The benchmark's statistical separation comes from keeping the FY2025/26 outcome key outside the development evidence and frozen forecasts; the remaining implementation records document how the run can be reproduced.