Automating Policy Analysis: A Human-Centered Prompt Engineering Approach for Policy Memo Drafting

Nicolas Rau
Matriculation No. 22-612-238
University of St. Gallen
St. Gallen, Switzerland
nicolas.rau@student.unisg.ch

Fabian Alberti
Matriculation No. 22-607-378
University of St. Gallen
St. Gallen, Switzerland
fabian.alberti@student.unisg.ch

Laurin Pan
Matriculation No. 23-610-140
University of St. Gallen
St. Gallen, Switzerland
laurin.pan@student.unisg.ch

Abstract

Effective policy communication demands the synthesis of complex information into concise, actionable documents defined by rigid formatting and a neutral executive tone. While Large Language Models (LLMs) demonstrate significant generative capabilities, they frequently struggle to meet these professional constraints in a zero-shot setting, often suffering from structural hallucinations and the inclusion of irrelevant background information. In this paper, we present a modular Policy Memo Generator designed to assist economists and public administrators by decoupling the research phase from the drafting phase. Our solution utilizes a multi-agent pipeline that employs hierarchical "Evidence Masks" to extract relevant context before a sequential drafting agent composes the document chapter by chapter. We evaluated this approach against a standard zero-shot baseline using real-world memos from the Brookings Institution. Our results demonstrate that this modular pipeline significantly improves content accuracy and structural integrity compared to standard prompting methods, effectively adhering to the strict stylistic constraints of professional policy analysis.

Keywords

Prompt Engineering, Policy Memos, Economics, LLM, Multi-Agent

1 Introduction

Effective policy communication relies on the ability to synthesize complex information into concise, structured, and actionable documents. The policy memo serves as a primary instrument in public administration, demanding rigid adherence to format and a neutral, executive tone. While Large Language Models (LLMs) have demonstrated significant capabilities in general text generation, they often struggle to meet the specific constraints of professional policy memo writing when prompted in an zero-shot manner. Common failure modes include structural hallucinations, the inability to distinguish between background evidence and policy recommendations, and "context bleeding," where irrelevant information contaminates specific sections of the document.

Recent developments in the public sector underscore both the growing prevalence and the significant pitfalls of LLMs as writing assistants in policy administration. As evidenced by the recent case involving Deloitte and the Australian government (Karp, 2025), generative models are already being actively deployed to draft complex policy reports and memos. However, this implementation also serves as a cautionary tale: while AI can generate volume, it demonstrates critical limitations when not guided by a specialized, constraint-based system. The most severe risk in automated policy analysis is the generation of plausible but fabricated information, rendering documents not only useless but actively dangerous for decision-makers relying on accuracy.

In this paper, we present the development and evaluation of a modular Policy Memo Generator designed to assist economists and public administrators in the drafting phase. Our solution is positioned not as a replacement for human research or ideation, but as a specialized writing assistant. We explicitly decouple the research phase from the writing phase: the user provides the "Gold Standard" inputs, including reference documents, specific policy options, and the addressee's stance, and the system functions as a constrained drafting engine. Our technical approach utilizes a multi-agent pipeline that processes the specified references through a two-fold workflow, utilizing hierarchical "Evidence Masks" to separate global context from chapter-specific details. By strictly defining the inputs and structure, we aim to solve the coherence issues inherent in long-form generation.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI for policymaking and LLM writing assistants. Section 3 defines the user goals and the problem scope. Section 4 details our prompt engineering solution and pipeline architecture. Sections 5 and 6 present our experimental setup and evaluation results, highlighting both automated metrics and expert qualitative feedback. Finally, Section 7 discusses the strengths, limitations, and ethical considerations of our approach and Section 8 concludes.

2 Review of Related Work

Our work is situated at the intersection of two research areas: AI in policymaking and automated writing assistants.

2.1 AI in Policymaking and Public Administration

The integration of Artificial Intelligence into public sector operations has graduated from experimental pilots to a core component of modern governance strategies, fundamentally altering how policy is researched, drafted, and evaluated. Aoki (2024) provides a comprehensive survey of this transformation, highlighting the technology's capacity to streamline legislative processes through automated document classification and policy drafting. The OECD (2025) further observes that governments are increasingly deploying AI to alleviate civil servant workload, exemplified by Brazil's MARIA system which produces first drafts of judicial reports, and

HSG PE '25, St. Gallen, Switzerland
2025.

the UK's initiatives for strategic communication planning. However, the OECD simultaneously warns that the reliance on such systems necessitates rigorous human oversight to mitigate risks regarding accountability, transparency, and bias.

In the complex domain of international negotiations, Ziegler et al. (2025) observed that while delegates now utilize chatbots to draft interventions and conduct background research, there is a duality where AI could both level the playing field for developing nations and exacerbate existing inequities. Gao (2023) similarly posits that LLMs can accelerate environmental policymaking by synthesizing scientific reports, though they caution against the high risks of hallucination in contexts such as legal compliance monitoring.

Beyond general administrative assistance, specialized tools have emerged to address distinct, high-stakes policy tasks. Wu et al. (2025b) proposed Sci2Pol, a system that fine-tunes models on a five-stage taxonomy (from understanding to verification) to generate policy briefs directly from scientific papers. Focusing on the integration of public input, Wang et al. (2025) developed PolicyPulse to synthesize large-scale public perspectives from online forums into structured policy themes, while Kuo et al. (2025) introduced PolicyCraft to support collaborative and participatory policy design through case-grounded deliberation.

2.2 LLM Writing Assistants

The domain of automated writing support has undergone a paradigm shift, moving beyond simple syntax correction to encompass sophisticated, agentic systems capable of managing the entire cognitive lifecycle of document creation. Lee et al. (2024) map the diverse design space of these intelligent assistants, emphasizing that effective systems must balance user control with automation depending on specific task requirements. To address the coherence challenges inherent in long-form generation, SuperWriter (Wu et al., 2025a) employs a reflection-driven framework that decomposes generation into planning, writing, and refining loops. Similarly, WriteHERE (Xiong et al., 2025) introduces a recursive planning mechanism that dynamically integrates retrieval, reasoning, and composition to maintain adaptability throughout the writing process.

Addressing the critical pre-writing and research phase, STORM (Shao et al., 2024) automates information gathering by simulating multi-perspective conversations between synthetic experts to generate comprehensive outlines. To ensure factual accuracy in professional reports, DeepWriter (Mao et al., 2025) differentiates itself by utilizing an offline knowledge base and dynamic memory modules, effectively minimizing the hallucinations often found in web-search-based approaches. In the narrative domain, DOME (Wang et al., 2024a) utilizes dynamic hierarchical outlining combined with temporal knowledge graphs to ensure plot consistency.

In specialized domains requiring rigid adherence to standards, multi-agent frameworks have proven particularly effective. AutoPatent (Wang et al., 2024b) utilizes distinct planner, writer, and examiner agents to automate the highly technical task of patent drafting. In a similar vein, QRAFT (Sahnan et al., 2025) leverages agentic collaboration to refine fact-checking articles through iterative editorial review, simulating the workflow between a journalist and an editor. These approaches collectively demonstrate the efficacy of structured, multi-stage pipelines for generating high-quality professional documents.

Our solution adapts the concepts of these intelligent writing assistants to the specific domain of policy memos. By designing a flexible, human-in-the-loop system, we address the need for nuance and strategic framing in sensitive topics while moving beyond the constraints of standard long-form generation. The following section will define the specific user goals, the 'Gold Standard' of policy writing, and the challenges inherent in automating this task.

3 Problem Definition

3.1 User Group & Goals

A policy memo is designed to provide a concise and analytically rigorous assessment of a specific situation or issue and to derive actionable recommendations for a decision-maker (Massachusetts Institute of Technology, 2004). Boys and Keating (2009) define its "core [as] to evaluate succinctly policy options on a specific issue for a specific policymaking audience," while noting that policy memos are increasingly employed beyond the public sector, particularly in private organizations. Accordingly, the policy memo is not confined to political decision-making but is widely used across a range of institutional contexts, as summarized in Table 1. Given this broad range of applications, policy memos may take various forms, including briefs, petitions, classical internal memoranda, and white papers (Pennock, 2011). Correspondingly, the professional background and institutional position of the author may vary substantially. Nevertheless, with respect to policy writing, the author's role can generally be characterized as that of an analyst and advisor supporting a decision-maker (Pennock, 2011). Typically, the author occupies a subordinate or advisory position within an organization and prepares a structured assessment of policy options for a supervisor, client, or the public, whether as an internal staff member or an external (paid or unpaid) consultant.

3.2 Gold Standard Definition

To inform our assessment of what constitutes an excellent policy memo, we relied on both primary and secondary research. While the latter involved not only reviewing multiple guides on policy memo writing but also analyzing a range of published policy memos

Table 1: Institutional Contexts and Functions of Policy Memos

Context	Primary Function
Intl. Organizations (e.g., IMF)	To guide member states and shape development policy.
Think Tanks	To inform and influence policymakers.
NGOs	To structure and contribute to public debate.
Academic Settings	To provide analytically grounded policy advice.
Private-Sector	To inform strategic responses to regulatory change.

issued by leading institutions, the latter consisted of evaluating our findings through an interview with a current MPA student at the Harvard Kennedy School.

Given that policy memos are typically addressed to a predetermined audience, often decision-makers with limited time and varying levels of domain expertise, they must prioritize brevity, clarity, and relevance, while deliberately avoiding jargon and unnecessary detail (Leadership for Educational Equity, nd). Consequently, unlike traditional academic papers, policy memos do not require extensive theoretical frameworks. Instead, they focus on feasible and actionable recommendations that enable decision-makers to move swiftly toward implementation (University of Southern California, 2025). The Writing Support Team at Boston University further emphasizes that policy memos differ structurally from other academic formats, often resembling an "inverted pyramid," in which the most essential information is presented first and progressively followed by less critical details (Boston University School of Public Health, nd). While the precise structure of a policy memo may vary depending on context, institution, and author, it is commonly organized into sections such as executive summary, background and problem definition, analysis, recommendations, and conclusion (University of Chicago Harris School of Public Policy, nd).

Given this variability, modularity emerges as a key feature for our application, allowing users to adapt the memo's structure and content to their specific needs and decision contexts.

The MPA student underscored the importance of extreme conciseness and the use of clear, neutral language. Moreover, he noted that, depending on the policy context, visualizations, such as charts illustrating risk-return or effort-impact trade-offs across policy options, are increasingly employed to support rapid and informed decision-making.

3.3 User Challenges

3.3.1 General Challenges in Policy Memo Drafting. The creation of a policy memo is a high-stakes exercise in information management that extends well beyond simple composition. While the research phase is inherently time-intensive, the drafting phase presents its own distinct set of cognitive and structural challenges. The primary difficulty lies in information density and constraints. As noted by the writing support guidelines at Boston University, policy memos often resemble an "inverted pyramid," requiring the most critical conclusions to be presented first. Writers must synthesize vast amounts of complex information into a document of strictly limited length (typically 3–4 pages) without sacrificing analytical depth. This requires a time-consuming process of iterative refinement to strip away non-essential detail while retaining necessary context.

Furthermore, authors face the challenge of audience calibration and tone. Unlike general expository writing, a policy memo is often addressed to a specific decision-maker with limited time and varying levels of domain expertise. The writer must tailor the argument to the addressee's specific stance and prior knowledge while maintaining a neutral, executive tone. Balancing this requirement for objective neutrality with the necessity of making persuasive, actionable recommendations creates a "neutrality paradox" that often leads to writer's block and multiple revision cycles.

3.3.2 Challenges with LLM Automation. Recent developments in the public sector underscore both the growing prevalence and the significant pitfalls of LLM as writing assistants in policy administration. As evidenced by the recent case involving Deloitte and the Australian government (Karp, 2025), generative models are already being actively deployed to draft complex policy reports and memos. However, this implementation also serves as a cautionary tale: while AI can generate volume, it demonstrates critical limitations when not guided by a specialized, constraint-based system. The gap revealed by such cases stems from four specific misalignments we observed between standard LLM behavior and policy memo requirements.

The most severe risk in automated policy analysis is the generation of plausible but fabricated information. In the absence of strict grounding mechanisms, LLMs are prone to inventing statistics, misattributing quotes, or creating non-existent citations to support an argument. For decision-makers relying on accuracy, such hallucinations render a document not only useless but actively dangerous.

In addition, policy memos demand rigid adherence to institutional formats. However, standard models prompted in a zero-shot manner frequently hallucinate structural elements, inventing unrequested sections that disrupt the formal flow.

Furthermore, in long-form generation, models often struggle to distinguish between background evidence and specific policy analysis. This phenomenon, which we term "context bleeding," results in irrelevant historical data contaminating the recommendation sections, obscuring the actionable advice.

Lastly, standard LLMs are often trained on narrative texts that build tension towards a conclusion. This conflicts with the "inverted pyramid" structure required in policy work, where the conclusion must be stated upfront. Without specific architectural intervention, models tend to bury the lead, reducing the document's utility for rapid decision-making.

To directly address these structural and cognitive challenges the following section will detail our specialized multi-agent architecture and the specific engineering choices designed to enforce strict adherence to the defined requirements.

4 Prompt Engineering Solution

4.1 Task Definition

Given a set of user inputs $\mathcal{X} = \{\mathcal{I}_{meta}, \mathcal{S}, \mathcal{R}\}$, where $\mathcal{I}_{meta}$ represents the memo's metadata (title, addressee, addressee's stance, occasion, purpose, length, and preliminary policy options), $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ represents the required structure consisting of n annotated chapters, and $\mathcal{R}$ is the set of user-provided reference documents, we aim to generate a comprehensive policy memo $\mathcal{M}$. The output $\mathcal{M}$ is composed of a sequence of chapters $c_1, c_2, \ldots, c_n$, where each chapter c_i corresponds to the structural annotation s_i and is synthesized according to the parameters in $\mathcal{I}_{meta}$.

We assume that the input datapoint $\mathcal{X}$ is explicitly provided by the user. Consistent with the distinction between pre-writing and writing stages in long-form generation, we do not address automatic open-domain information retrieval or autonomous policy ideation. Our system focuses strictly on the controlled drafting and synthesis of the memo based on the provided source materials and constraints.

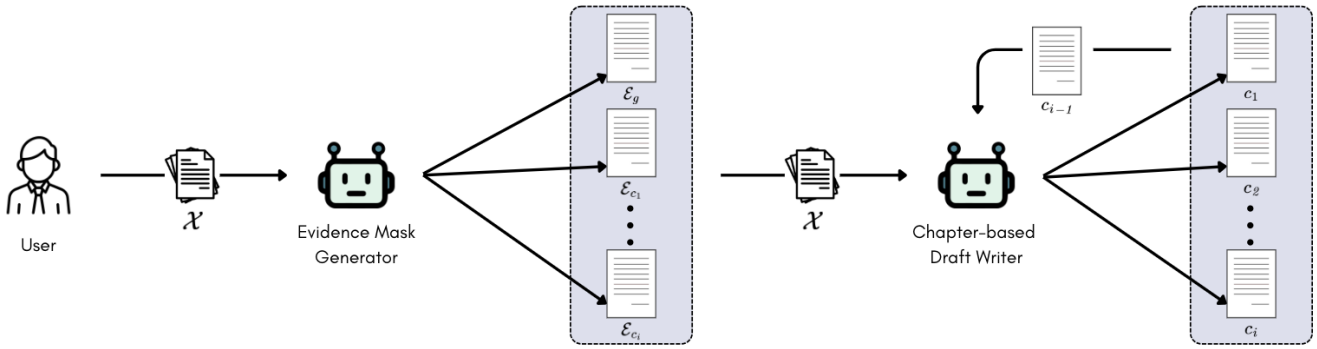

Figure 1: Process Diagram of the Policy Memo Generator Pipeline

To operationalize this task, the following section will detail our technical pipeline, which decomposes the drafting process into a structured, multi-agent workflow.

4.2 Pipeline

To manage the complexity of this task, we rejected single-shot generation in favor of a specialized agentic architecture. By mimicking the professional workflow of rigorous source curation followed by systematic outlining, our solution consists of two specialized agents that collaborate to transform the memo metadata ($\mathcal{I}_{meta}$), required structure ($\mathcal{S}$), and raw reference documents ($\mathcal{R}$) into a polished memo $\mathcal{M}$:

Evidence Mask Generator (Module 1) Functioning as a research analyst, this agent processes the set of reference documents $\mathcal{R}$ to identify and extract passages that are strictly relevant to the memo's background context and policy recommendations. Rather than producing new summaries, the module operates in a purely extractive manner and outputs two structured artifacts.
First, it produces a global relevance set $\mathcal{E}_g$, consisting of verbatim passages that are broadly relevant across the memo. Second, it generates chapter-specific relevance sets $\mathcal{E}_{c_i}$ for each section $i \in \mathcal{S}$. These chapter-specific outputs act as relevance masks that select verbatim source passages pertinent to the respective chapter.
Importantly, $\mathcal{E}_{c_i}$ does not constitute a chapter-level summary. Instead, it represents a collection of unaltered excerpts drawn from the source documents, filtered solely based on their relevance to chapter i.

Chapter-based Draft Writer (Module 2) Acting as the primary drafter, this module composes the memo sequentially. While all chapter agents receive the full set of metadata $\mathcal{I}_{meta}$ and the required structure $\mathcal{S}$, the injection of source content is conditional. Chapters requiring evidentiary support are provided with the global summary $\mathcal{E}_g$ and the specific chapter mask $\mathcal{E}_{c_i}$. Conversely, chapters designated as purely directive rely solely on the provided user instructions, receiving neither the global summary nor external reference injections. Crucially, every agent receives the context of all previously drafted chapters $\mathcal{D}_{c_i}$ to ensure narrative coherence and logical progression across the document.

4.3 Architectural Strategy

Having established the two-stage pipeline, it is critical to understand how specific engineering choices within these modules address the failure modes identified in Section 3.3. Our design strategy maps each identified challenge: Hallucination, Context Bleeding, Structural Failure, and Narrative Misalignment, to a specific architectural countermeasure implemented in the pipeline.

4.3.1 Countering Hallucination via Closed-Loop Retrieval. To address the risk of factual hallucination, we explicitly rejected the RAG (Retrieval-Augmented Generation) approach that queries open external databases. Instead, the Evidence Mask Generator (Module 1) is restricted strictly to the user-provided Reference Set ($\mathcal{R}$).

4.3.2 Eliminating Context Bleeding via Hierarchical Masking. Standard context windows treat all input data as equally relevant, leading to "Context Bleeding". We addressed this by implementing the hierarchical masking strategy described in Module 1. Rather than feeding the entire corpus to the writer, the system creates specific "Evidence Masks" ($\mathcal{E}_{c_i}$) for each chapter. For instance, when the agent drafts the "Policy Recommendations," it is blinded to unrelated information, ensuring that historical context remains in the "Background" chapter and does not contaminate the forward-looking analysis.

4.3.3 Enforcing Structure via Agentic Separation. To solve structural hallucination, we decomposed the writing task. A single agent trying to write a full memo often loses track of formatting constraints. By assigning a distinct "Chapter Agent" to each section of the structure S (Module 2), we reduce the cognitive load on the model. Each agent is responsible for only one segment (e.g., "Background"), making strict adherence to the requested format significantly more robust than in a monolithic generation pass. Crucially, our pipeline enforces a functional constraint where the Writer Agent is only invoked for chapters explicitly defined in the user's input structure S. This leaves the agent no computational room to invent unrequested structural elements, effectively rendering structural hallucination impossible by design.

4.3.4 Correcting Narrative Flow via Sequential Chaining. To counter narrative misalignment, we implemented a sequential context chain. While the Evidence Masks isolate data, the narrative flow must be continuous. Therefore, each Chapter Agent receives the text of the immediately preceding chapter (c_{i-1}) as a read-only input. This allows the model to use transitional language (such as "As mentioned in the previous section...") and maintain the "inverted pyramid" style, preventing the disjointedness often seen in parallelized generation.

4.4 UI Solution

To ensure accountability in the sensitive domain of policy decision-making, we implemented a four-step user interface designed to maintain a strict human-in-the-loop workflow. As illustrated in Figure 2, the interface guides the user through a linear process that prioritizes user control over full automation.

The process begins with the Configuration phase, where the user defines the input set $\mathcal{X}$. This includes uploading reference documents $\mathcal{R}$ and specifying metadata $\mathcal{I}_{meta}$ such as the memo's title, occasion, purpose, and the addressee's specific stance. Additionally, the user defines the required structure $\mathcal{S}$ and annotates the focus of each chapter. In the second stage, Policy Specification, the user explicitly inputs the preliminary policy options and/or recommendations. Users may optionally provide argumentation, but the primary goal is to ground the AI's generation in specific, human-defined policy choices.

Once the pipeline generates the initial text, the workflow moves to the Refinement stage. Here, the user reviews the memo $\mathcal{M}$ chapter by chapter. The interface allows for granular editing where the user can leverage the AI to rewrite specific sections, ensuring the draft accurately reflects the intended nuance using the full context of the provided inputs. Finally, the Export stage allows the user to apply institutional templates and finalize the document for distribution.

With the architectural pipeline and user interface established, the following section will outline the rigorous experimental framework and the specific datasets used to evaluate the system's performance against real-world policy standards

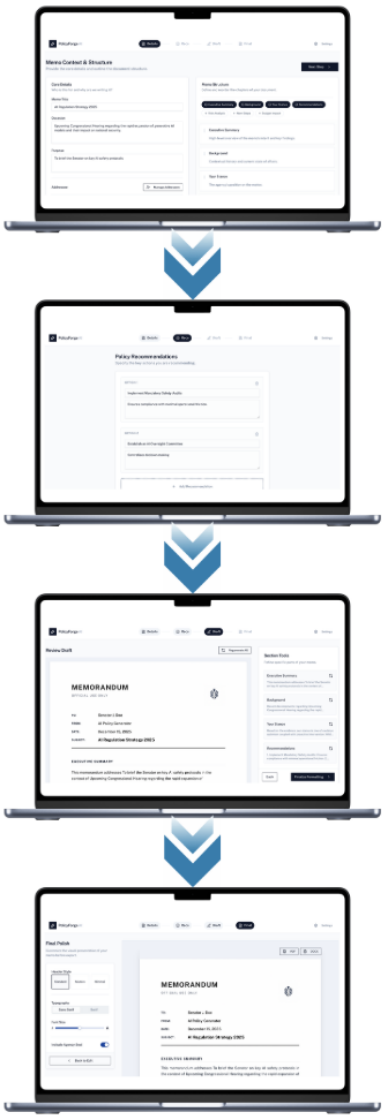

Figure 2: Vertical Process Diagram of the User Interface

5 Experimental Setup

To facilitate rigorous manual review and evaluation, we restricted our experimental scope to policy memos of approximately 3–4 pages. However, it is important to note that our underlying pipeline, utilizing section masks and iterative drafting, is architecturally agnostic to length and designed to scale to long-form reports.

We selected the memo series "Memos to the President" from The Brookings Institution (2009) as our source dataset. While we evaluated archives from the Harvard Kennedy School, the Economic Policy Institute, and the Council on Foreign Relations, the Brookings series offered the highest degree of structural consistency and comparable length across documents. From this series, we selected two memos to serve as a stylistic "Gold Standard" representing the target tone and format (Bradford and Unger, 2008, Prasad, 2009).

For our specific test case, we selected the memo titled "Fix the Tax System" (Harris and Gale, 2008). This document was chosen because it relies on a discrete set of accessible references which we explicitly provided as the input set $\mathcal{R}$, and possesses a clear argument structure that aligns with our pipeline's design. To validate the efficacy of our multi-agent approach, we compared our generated output against two baselines: the original expert-written text and a draft generated by a standard Zero-Shot Prompt. The Zero-Shot baseline received the identical input set $\mathcal{X}$ (metadata, structure, and references) but attempted to gather the evidence and generate the full memo in a single context window without the iterative chapter-based architecture.

Having established the comparative baselines and the target source material, the following section will present the quantitative metrics and qualitative expert feedback used to measure the efficacy of our multi-agent approach.

6 Evaluation

6.1 Evaluation Methodology

To assess the efficacy of our pipeline, we employed a mixed-methods approach combining automated similarity metrics with qualitative expert review.

6.2 Automated Metrics Results

We utilized ROUGE-L, a metric based on the Longest Common Subsequence that captures sentence-level structural similarity and vocabulary overlap, to measure the similarity between the generated drafts and the Gold Standard memos. As detailed in Table 2, we performed two types of analysis: a General Stylistic Fit (comparing against unrelated memos to measure format adherence) and a Content Accuracy Check (comparing against the specific target memo the system was tasked to recreate).

Table 2: ROUGE-L F-Measure Comparison

Comparison Target	Zero-Shot	Pipeline	**Imp.**
General Style Checks			
vs. Financial Memo	0.1149	**0.1259**	+9.6%
vs. Global Dev Memo	0.1211	**0.1259**	+4.0%
Content Accuracy Check			
vs. **Target (Tax) Memo**	0.1562	**0.1936**	**+23.9%**

Our multi-agent pipeline consistently outperformed the Zero-Shot baseline across all metrics, exhibiting the most pronounced advantage in the Content Accuracy Check with a score of 0.1936, representing a 23.9% improvement over the baseline. To rigorously validate whether this score represented true semantic synthesis or merely generic formatting compliance, we calculated a "Human Baseline" (or "noise floor"). This baseline was derived by calculating the ROUGE-L similarity between two unrelated expert-written memos from the Brookings Institution (e.g., comparing the Tax Memo against a Financial Stability Memo). This comparison yielded a score of 0.1524, quantifying the intrinsic similarity that any two professional policy memos would share simply by adhering to the same institutional template, headers, and formal tone.

Interpreting our experimental results against this noise floor reveals a critical distinction in model behavior. The Zero-Shot baseline achieved a score of 0.1562, which is negligibly higher than the Human Baseline of 0.1524. This proximity suggests that the Zero-Shot model largely succeeded only in mimicking the superficial structure of the document without accurately capturing the unique semantic arguments of the source material. In contrast, our Pipeline's score of 0.1936 significantly clears the Human Baseline by approximately 27%. These metrics indicate that the multi-agent system effectively incorporated domain-specific vocabulary and concepts from the provided reference set, distinguishing the output from generic policy language and aligning it more closely with the target draft.

6.3 Qualitative Review

We separated the qualitative evaluation into two tiers: a general inspection of structural compliance performed by the authors, and an overall style evaluation performed by a domain expert.

6.3.1 General Compliance & Readability. We first assessed the generated drafts against the strict constraints defined in the user prompt (max. 1200 words, specific chapter structure). As summarized in Table 3, the Multi-Agent approach demonstrated superior adherence to constraints compared to the Zero-Shot baseline.

Table 3: Qualitative Comparison of Generated Drafts

Criteria	Zero-Shot Baseline	Multi-Agent Pipeline
Length	**Fail:** Generated 1940 words (exceeded limit by 740 words).	**Pass:** Generated 1390 words (barely exceeded limit).
Structure	**Fail:** Added unrequested chapters ("Recommendations in Context").	**Pass:** Adhered strictly to the annotated structure $\mathcal{S}$.
Style	**Poor:** Included raw source notes in text; used awkward "Pros/Cons" bullet points.	**Good:** Integrated reasoning into coherent paragraphs.
Flow	**Disjointed:** Interrupted by raw source citations.	**Coherent:** Smooth narrative flow without meta-commentary.

The Zero-Shot model struggled significantly with formatting constraints. It over-generated text (1940 words vs. the 1200-word limit) and hallucinated structural elements, adding a "Recommendations in Context" chapter that was not requested in the input $\mathcal{S}$. Stylistically, it failed to separate "research notes" from "draft text," often leaving raw source citations and "Pros/Cons" lists directly in the final output. In contrast, the final pipeline produced a clean, narrative text that strictly followed the structural template with a professional writing style.

6.3.2 Expert Content Evaluation. To evaluate the utility of our generated artifacts, we commissioned a blind review by an experienced policy memo writer (MPA Student, Harvard Kennedy School). The expert was asked to compare the Zero-Shot baseline against the Final Pipeline output with a specific focus on their suitability for high-level decision-makers.

The review confirmed a distinct preference for the Final Pipeline-generated memo, which was characterized as "notably better" and significantly more "expert-like" than the baseline. Two key differentiators were highlighted:

Structural Integrity: The expert noted that the Pipeline draft adhered strictly to the expected professional format. Unlike the Zero-Shot baseline, which struggled with pacing and organization, the Pipeline output maintained a logical progression that facilitated rapid information retrieval.

Decision-Maker Stylistics: The expert emphasized that the stylistic tone of the Pipeline draft was "much more aligned with the

expectations of decision-makers, such as politicians and professors." While the Zero-Shot model often defaulted to a generic style, the Final Pipeline successfully adopted the executive tone and writing style required for policy memo's.

6.4 Iterative Refinements and Ablation Study

The streamlined architecture presented in Section 4.2 is the result of an iterative experimental process. While the final system relies on a two-stage workflow, we initially designed and tested different pipeline settings including a more complex four-fold pipeline that included dedicated Review and Editor agents. Based on empirical performance during the drafting of short-form memos, we simplified the system to the current architecture to prioritize efficiency and coherence.

6.4.1 Component Ablation: The Evaluator Paradox. In our initial prototype, we hypothesized that a recursive "Review and Refine" loop would strictly improve output quality by simulating a human editorial process. We implemented a dedicated Reviewer Agent to critique the draft and an Editor Agent to integrate that feedback. However, our ablation experiments revealed a counter-intuitive "Evaluator Paradox." The Reviewer Agent consistently assigned inflated qualitative scores to the initial drafts, failing to discriminate between "acceptable" and "excellent" prose. Because the Writer Agent was already strictly constrained by the inputs, the initial drafts rarely contained the gross errors the Reviewer was designed to catch.

Consequently, the feedback loop rarely triggered substantive revisions. Instead, it introduced significant latency and increased token costs without yielding a significant improvement. We attribute this partially to the experimental scope: for memos of 3-4 pages (which then were separated between separate agents), modern LLMs can maintain sufficient coherence in a single pass without needing an external critic. We posit that for longer documents (e.g., 20+ page reports), the "drift" in quality would likely be higher, making a Reviewer Agent necessary. However, for the specific design goal of this paper, we excluded these modules to optimize for system speed and cost-efficiency.

6.4.2 Successful Architectural Shifts. In contrast to the removal of the review loop, architectural changes focused on the drafting phase yielded immediate gains. The most significant performance improvement resulted from the transition to Sequential Chapter Drafting. Moving from generating the full memo in one pass to a modular, chapter-by-chapter approach was essential for allowing the system to strictly adhere to the user-defined structure S and maintain logical progression. This was complemented by the implementation of Hierarchical Evidence Masking, where we replaced a single "General Relevance" search with a two-tiered strategy consisting of a Global Evidence Mask ($\mathcal{E}_g$) for high-level context and specific Chapter Evidence Masks ($\mathcal{E}_{c_i}$). This refinement proved critical for optimizing token usage within the context window and preventing "context bleeding," ensuring that specific chapters, such as "Policy Recommendations," were grounded exclusively in relevant data without being distracted by unrelated background information.

7 Discussion

7.1 Strengths of the Approach

Our evaluation highlights several key architectural strengths that distinguish our pipeline from standard LLM interaction models. First, the decoupled nature of the pipeline, which separates evidence extraction from drafting, renders the system highly adaptable and modular. While tested on tax policy, this architecture is domain-agnostic and could be used for a variety of different policy topics. Second, another significant technical achievement was the elimination of "context bleeding." By utilizing a dual-layer masking strategy consisting of global and chapter-specific masks, we ensured that the agent focused exclusively on relevant information for each section. This prevented the common failure mode where models conflate background facts with policy recommendations. Finally, unlike fully autonomous agents that attempt to independently solve policy problems, our system enforces a strict division of labor that maintains human sovereignty. By requiring the user to specify the required user inputs such as policy options and references in the metadata, we ensure that the strategic direction remains firmly in human hands, with the AI serving purely as a drafting engine.

7.2 Limitations

Despite these strengths, the current iteration faces distinct constraints. The system acts as a synthesis engine rather than an analytical one, meaning it relies entirely on the quality of the user-provided reference set and metadata. If the user uploads irrelevant or biased sources, the model will produce a polished but substantively flawed memo, reinforcing the garbage in, garbage out principle. Additionally, we strictly evaluated the system on memos of three to four pages.

While the modular architecture is theoretically capable of scaling to longer reports, this capability remains untested in our current experimental setup.

Furthermore, our pipeline currently suffers from multimodal blindness, generating text only. This represents a significant deviation from the Gold Standard defined in Section 3.2, where our expert interview highlighted that visualizations are increasingly employed to support rapid decision-making. Consequently, users must currently manually generate and insert these visual elements post-generation to meet professional standards.

Finally, regarding the architectural evolution, we initially hypothesized that a recursive review loop would improve quality, but our ablation study revealed an Evaluator Paradox. The Reviewer Agent consistently assigned inflated qualitative scores to initial drafts and failed to discriminate between acceptable and excellent prose. Because the feedback loop introduced latency without yielding significant improvement for memos of this length, we removed these modules to prioritize system speed.

7.3 Ethical Considerations

The automation of government document drafting introduces a set of ethical risks that require careful governance. First, "LLMs are made to produce text that is convincing", which may lead both the author of the policy memo and the addresses to accept outputs as accurate without sufficient fact-checking (Coeckelbergh, 2025).

In fact, an LLM is designed not to provide the truth but to create a coherent text solely "through estimating the likelihood that a particular word will appear next, given the text that has come before" (Hicks et al., 2024). As a result misleading or incomplete inputs can easily propagate into misleading outputs, even in cases where a human reader could correctly interpret the underlying source material.

Furthermore, a growing amount of literature has documented systemic biases in LLM outputs, particularly with respect to race and gender (Salinas et al., 2025). In fact, models seem to frequently favor white over black individuals and men over women (Salinas et al., 2025). In a political and administrative context, such biases are especially problematic. If policy drafts are shaped by latent demographic biases, foundational principles such as equality before the law are undermined, and decision-making risks becoming skewed toward discriminatory outcomes rather than societal welfare maximization. These risks are only exacerbated by the high stakes inherent in public policy, particularly in domains such as security or defense. Relying upon the output of an LLM might seem highly inappropriate for such cases and also raise serious liability issues (Gabison and Xian, 2025).

Moreover, policy drafting frequently involves confidential or classified data. The use of commercial, cloud-hosted LLMs therefore poses significant data leakage risks, implying that any deployable system should rely exclusively on locally hosted models to ensure that sensitive information does not leave secure environments.

Finally, transparency is essential in a political context. As AI-generated text can be virtually indistinguishable from human writing, the undisclosed use of automated drafting tools risks eroding public trusts. Memos generated with AI assistance should thus include clear disclosure indicating the use of automated systems.

7.4 Future Research Directions

While this study established a robust baseline for automated drafting, several avenues remain for expanding the system's agentic capabilities.

First, future work could address the reliance on user-provided sources by integrating an investigator agent capable of performing open-ended research. This agent would need to autonomously query external databases, verify source credibility, and curate the reference set prior to the drafting phase.

Second, future iterations should address the current multimodal blindness by integrating agents capable of parsing raw data and generating charts or figures directly within the memo output.

Third, future research should test this pipeline on significantly longer documents of over 20 pages. In this context, the Evaluator Paradox observed in short-form drafting may disappear. As document complexity increases, the re-introduction of Reviewer and Editor agents may become necessary to maintain narrative coherence over long horizons.

Finally, we propose extending the system to generate autonomous policy options. While our current approach keeps ideation strictly human, future iterations could employ a debate mechanism similar to the agentic collaboration seen in recent editorial review frameworks or multi-perspective simulations. Agents could argue distinct viewpoints to synthesize novel policy solutions. Crucially, such a feature would necessitate a rigorous user interface design that allows human experts to quickly evaluate and distinguish AI-generated options from human-derived ones, ensuring that automated recommendations are clearly alienated and validated before inclusion in the final draft.

8 Conclusion

In this paper, we demonstrated that a structured, multi-agent pipeline significantly outperforms standard zero-shot prompting in the generation of professional policy memos. By decomposing the complex task of memo writing into discrete stages of evidence extraction and sequential chapter drafting, we addressed critical limitations in current LLM capabilities regarding structural adherence and context management.

Our experimental results validate the efficacy of this modular approach. The implementation of hierarchical evidence masking and sequential chapter composing allowed our pipeline to achieve a 23.9% improvement in content accuracy (ROUGE-L) compared to the baseline, successfully clearing the statistical "noise floor" of generic formatting. Qualitatively, blind expert review confirmed that our system produced drafts that adopted the requisite executive tone, whereas the baseline frequently suffered from pacing issues and hallucinated structural elements.

Crucially, our findings emphasize the importance of a human-in-the-loop design. By ensuring that the user retains control over the source material, policy stance, and argumentation structure, we mitigate the risks of hallucination and bias while leveraging the AI's speed in synthesis and drafting. While the current system is limited to text generation and relies on the quality of user-provided references, the architecture is designed to be scalable. Future iterations could address the integration of multimodal data visualization, the processing of longer reports, and the generation of autonomous policy options via agentic debate mechanisms. We conclude that specialized prompt engineering solutions can effectively reduce the administrative burden of policy memo drafting, allowing experts to focus on strategic analysis rather than formatting and composition.

Acknowledgments

The authors utilized the Gemini (Pro 3) large language model for editorial assistance, including the generation of structured outlines, refinement of LaTeX formatting, and optimization of prose for clarity and coherence. All research, experimental design, prompt engineering methodology, and data analysis were performed independently by the authors.

References

Goshi Aoki. 2024. Large Language Models in Politics and Democracy: A Comprehensive Survey. arXiv:2412.04498 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2412.04498

Boston University School of Public Health. n.d.. *Policy Memos.* Boston University. https://www.bu.edu/sph/students/on-campus-students/academic-accommodations-and-support/communication-resources/policy-memo/ Webpage providing guidance on writing policy memos; Boston University School of Public Health communication resources.

James D. Boys and Michael F. Keating. 2009. The policy brief: Building practical and academic skills in international relations and Political science. *Politics* 29, 3 (9 2009), 201–208. doi:10.1111/j.1467-9256.2009.01356.x

Colin I. Bradford and Noam Unger. 2008. Memo to the President: Redefine America8217;s Global Development Cooperation. (12 2008). https://www.brookings.edu/articles/memo-to-the-president-redefine-americas-global-development-cooperation/

Mark Coeckelbergh. 2025. LLMs, Truth, and Democracy: An Overview of risks. *Science and Engineering Ethics* 31, 1 (1 2025), 4. doi:10.1007/s11948-025-00529-0

Garry A. Gabison and R. Patrick Xian. 2025. Inherent and emergent liability issues in LLM-based agentic systems: a principal-agent perspective. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*. Association for Computational Linguistics, 109–130. doi:10.18653/v1/2025.realm-1.9

Andrew Gao. 2023. Implications of ChatGPT and large language Models for environmental policymaking. *SSRN Electronic Journal* (1 2023). doi:10.2139/ssrn.4499643

Ben Harris and William G. Gale. 2008. Memo to the President: Fix the tax system. (12 2008). https://www.brookings.edu/articles/memo-to-the-president-fix-the-tax-system/https://www.brookings.edu/articles/memo-to-the-president-fix-the-tax-system/

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology* 26, 2 (6 2024). doi:10.1007/s10676-024-09775-5

Paul Karp. 2025. AI-tainted Deloitte report was worse than previously thought. *Australian Financial Review* (6 Nov 2025). https://www.afr.com/politics/ai-tainted-deloitte-report-was-worse-than-previously-thought-20251106-p5n863

Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–24. doi:10.1145/3706598.3713865

Leadership for Educational Equity. n.d.. Guide to Writing an Effective Policy Memo. Organizational guide. Nonpartisan guidance document accessed via wearelee.org.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–35. doi:10.1145/3613904.3642697

Song Mao, Lejun Cheng, Pinlong Cai, Guohang Yan, Ding Wang, and Botian Shi. 2025. DeepWriter: A Fact-Grounded Multimodal Writing Assistant Based On Offline Knowledge Base. arXiv:2507.14189 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2507.14189

Massachusetts Institute of Technology. 2004. Writing Effective Policy Memos. Course handout, 11.479 – Water & Sanitation Infrastructure Planning. https://dspace.mit.edu/bitstream/handle/1721.1/36824/11-479Spring-2004/NR/rdonlyres/Urban-Studies-and-Planning/11-479Spring-2004/9CE4ACA2-EC3D-4C1D-91CC-27971E27DCF5/0/pmwriting.pdf Internal course materials, Spring 2004.

OECD. 2025. *Governing with Artificial Intelligence*. doi:10.1787/795de142-en

Andrew Pennock. 2011. The case for using policy writing in undergraduate political science courses. *PS Political Science Politics* 44, 1 (1 2011), 141–146. doi:10.1017/s1049096510002040

Eswar Prasad. 2009. Memo to the President: Restore global financial stability. (1 2009). https://www.brookings.edu/articles/memo-to-the-president-restore-global-financial-stability/

Dhruv Sahnan, David Corney, Irene Larraz, Giovanni Zagni, Ruben Miguez, Zhuohan Xie, Iryna Gurevych, Elizabeth Churchill, Tanmoy Chakraborty, and Preslav Nakov. 2025. Can LLMs Automate Fact-Checking Article Writing? arXiv:2503.17684 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2503.17684

Alejandro Salinas, Amit Haim, and Julian Nyarko. 2025. What's in a Name? Auditing Large Language Models for Race and Gender Bias. arXiv:2402.14875 [cs.CL] https://arxiv.org/abs/2402.14875

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6252–6278. doi:10.18653/v1/2024.naacl-long.347

The Brookings Institution. 2009. Memos to the President. https://www.brookings.edu/tags/memos-to-the-president/

University of Chicago Harris School of Public Policy. n.d.. How to Write a Policy Memo That Matters. Instructional guide. Policy memo writing guide used in public policy instruction.

University of Southern California. 2025. *Writing a Policy Memo*. University of Southern California. https://libguides.usc.edu/writingguide/assignments/policymemo USC Libraries research guide on how to write a policy memo; last updated 2025.

Maggie Wang, Ella Colby, Jennifer Okwara, Varun Nagaraj Rao, Yuhan Liu, and Andrés Monroy-Hernández. 2025. PolicyPulse: LLM-Synthesis Tool for Policy Researchers. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM, 1–17. doi:10.1145/3706599.3720266

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, daiyuan li, Yu Hu, and Mingkui Tan. 2024a. Generating Long-form Story Using Dynamic Hierarchical Outlining with Memory-Enhancement. arXiv:2412.13575 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2412.13575

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, and Min Yang. 2024b. AutoPatent: A Multi-Agent Framework for Automatic Patent Generation. arXiv:2412.09796 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2412.09796

Weimin Wu, Alexander C. Furnas, Eddie Yang, Gefei Liu, Akhil Pandey Akella, Xuefeng Song, Dashun Wang, and Han Liu. 2025b. Sci2Pol: Evaluating and Fine-tuning LLMs on Scientific-to-Policy Brief Generation. arXiv:2509.21493 [cs.CE] https://arxiv.org/abs/2509.21493

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Juanzi Li, and Roy Ka-Wei Lee. 2025a. SuperWriter: Reflection-Driven Long-Form Generation with Large Language Models. arXiv:2506.04180 [[cs.CL](http://cs.cl/)] https://arxiv.org/abs/2506.04180

Ruibin Xiong, Yimeng Chen, Dmitrii Khizbullin, Mingchen Zhuge, and Jürgen Schmidhuber. 2025. Beyond Outlining: Heterogeneous Recursive Planning for Adaptive Long-form Writing with Language Models. arXiv:2503.08275 [[cs.AI](http://cs.ai/)] https://arxiv.org/abs/2503.08275

Matt Ziegler, Sarah Lothian, Brian O'Neill, Richard Anderson, and Yoshitaka Ota. 2025. AI language models could both help and harm equity in marine policymaking. *npj Ocean Sustainability* 4, 1 (6 2025). doi:10.1038/s44183-025-00132-7

Technical Appendix

A Experimental User Input Definitions

The following section defines the specific user parameters used for the "Fix the Tax System" test case evaluated in this paper. All of this user inputs were created from the authors based on the original policy memo "Fix the Tax System" by Harris and Gale (2008).

Memo Title
Fix the Tax System
Addressee
President Barack Obama
Addressee's Stance
President-elect Obama campaigned on providing broad-based tax relief for middle-class families, restoring fairness in the tax code, and returning to fiscal responsibility. He emphasized support for small businesses, job creation in the United States, and reforming the tax system to address both urgent and long-term structural problems. Obama also acknowledged that repairing the health care and energy systems requires more rational and efficient tax rules. At the same time, he recognized the need to address expiring tax cuts, the growth of the AMT, and the nation's large long-term fiscal gap through a balanced mix of tax reform and responsible revenue generation.
Occasion
This memo is written in the style of the Brookings transition memos prepared for President-elect Barack Obama in late 2008. These memos were created during a moment of political transition and economic uncertainty, when the incoming administration required clear, timely guidance on urgent policy challenges.
Purpose
The memo's purpose is to provide President Obama with concise, evidence-based, and actionable policy recommendations. It should translate complex issues into clear options, highlight trade-offs, and offer feasible steps the administration can take immediately.
Target Length
Max. 1200 words
Structure
(1) **The Situation:** A concise, analytical summary that describes the current context or crisis, explains why the issue matters now, identifies key pressures, risks, and opportunities, and acknowledges relevant political, economic, or institutional constraints. The tone is factual, situational, and problem-defining.
(2) **Your Stance:** A short section that recalls the president's campaign commitments or expressed priorities and connects these priorities to the current policy challenge. This situates the memo within Obama's agenda.
(3) **Recommendations:** The core of the memo should contain 3–6 high-level recommendations, each with a brief headline (e.g., "Reform Energy Taxation") and 2–4 sentences explaining why it matters. Brookings emphasizes clarity, sequencing, and feasibility.
Preliminary Policy Options
(1) Price Carbon Through a Carbon Tax or Cap-and-Trade System
(2) Reform the Tax Treatment of Employer-Provided Health Insurance
(3) Broaden the Tax Base and Address Long-Term Fiscal Imbalances
(4) Integrate Corporate and Individual Capital Income Taxation for New Investment
(5) Strengthen Enforcement and Eliminate Corporate Tax Shelters and Inefficient Subsidies
(6) Enhance Retirement Saving Through Automatic Enrollment and Universal Matching Incentives
(7) Convert Deductions and Exemptions Into Refundable Credits Where Appropriate
(8) Reform the Mortgage Interest Deduction Into a Refundable First-Time Homebuyer Credit
(9) Eliminate the State and Local Tax Deduction as Part of AMT Reform
(10) Simplify Tax Filing Through Return-Free Filing and Consolidation of Overlapping Credits
(11) Improve IRS Administrative Capacity and Compliance Enforcement
Sources
- https://www.finance.senate.gov/imo/media/doc/051308wgtest.pdf
- https://www.brookings.edu/articles/the-tax-system-too-complex-unfair-and-outdated/
- https://www.brookings.edu/articles/metro-raise-boosting-the-earned-income-tax-credit-to-help-metropolitan-workers-and-families/
- https://www.brookings.edu/articles/facing-the-music-the-fiscal-outlook-at-the-end-of-the-bush-administration/
- https://www.brookings.edu/articles/fixing-the-tax-system-support-fairer-simpler-and-more-adequate-taxation/
- https://www.urban.org/sites/default/files/publication/32101/411777-back-from-the-grave.pdf
- https://www.urban.org/sites/default/files/publication/31986/411749-An-Updated-Analysis-of-the-Presidential-Candidates-Tax-Plans-Updated-September--.PDF
- https://www.urban.org/sites/default/files/publication/51871/411194-The-Expanding-Reach-of-the-Individual-Alternative-Minimum-Tax.PDF
- https://www.hamiltonproject.org/wp-content/uploads/2023/01/Achieving_Progressive_Tax_Reform_in_an_Increasingly_Global_Economy.pdf
- https://www.hamiltonproject.org/wp-content/uploads/2023/01/An_Economic_Strategy_to_Address_Climate_Change_and_Promote_Energy_Security.pdf

B Final Pipeline

This section details the prompts and outputs for the final modular solution.

Note on Prompt Representation: *While our actual implementation relies on a modular codebase (available for review in the supplementary materials), the prompts presented in this appendix have been consolidated for readability. To clearly demonstrate the context received by the model, dynamic user inputs are displayed here as integrated text within the prompt if possible, rather than as abstract code templates.*

B.1 Module 1: Information Extraction

System Prompt (Information Extraction)

You are a Source Summarizing and Relevance-Masking Assistant.
SUMMARY RULES
- Summary MUST reduce length aggressively (target 10% of original).
- Preserve information that could affect policy reasoning.
- NO external knowledge, NO interpretation, NO paraphrasing of meaning.
- Remove narrative filler, redundancy, rhetorical framing, and examples.
- Combine internally redundant arguments within the same source.
- Do NOT merge arguments across different sources.

RELEVANCE MASK RULES (Relevance Component)
- ONLY extract verbatim passages from the source text.
- Include only: causal mechanisms, core findings, essential arguments, quantitative evidence, policy-relevant facts.
- Exclude: narrative background, minor examples, rhetorical content, descriptive filler.
- Maintain chronological order within the source.

GLOBAL SUMMARY LOGIC Generate ONE single global summary.
- This summary is NOT chapter-specific.
- All chapters with uniqueness = 1 must refer to the same global summary.

OUTPUT FORMAT: Your entire output MUST follow EXACTLY this format:
OUTPUT FORMAT START
GLOBAL SUMMARY
<text of the source summary>
CHAPTERS
<Chapter Number>. <Chapter Title>:
- <verbatim relevant passage>
- <verbatim relevant passage>
- ...
<Chapter Number>. <Chapter Title>:
- <verbatim relevant passage>
- ...
OUTPUT FORMAT END

STRICT FORMATTING RULES:
(1) Output EXACTLY one global summary under the '# GLOBAL SUMMARY #' header.
(2) Output ONLY chapters with uniqueness = 1. Omit all others completely.
(3) Each chapter (where NNN is the number of the chapter) MUST start with: <NNN>. <Chapter Title>:
(4) Maintain numbered order of the chapters.
No text may appear after this final marker (### OUTPUT FORMAT END ###).

User Prompt (Information Extraction)

Process the following source:
TITLE:
[source_title]
TEXT (delimited by ###):
###
[source_text]
###
CHAPTER_LIST:
[Dynamic Block: Iterates through included chapters]
[Chapter Number]. [Chapter Title]:
- Description: [Description]
[If Recommendation Chapter]:
- Preliminary Policy Options:
- [Option Text]
Return the output according to the rules in the system message.

B.2 Module 2: Chapter-based Writer

B.2.1 Recommendations Chapter.

System Prompt (Special Recommendations Writer)

You are a [Chapter Name]-Chapter writing assistant. Your task is to generate a final written chapter for this section. You must transform the user's preliminary policy ideas into a coherent, fully developed, evidence-based text.
RULES:
- The user's preliminary policy options are authoritative directions.
- You must NOT question, reject, critique, or override any user-provided idea.
- You must ONLY expand, deepen, and articulate the user's ideas.
- Do NOT add external knowledge, assumptions, or invented facts.

OUTPUT FORMAT: ### CHAPTER_BEGIN ###
<chapter text>
CHAPTER_END
NOTES ON FORMAT:
- Output must contain ONLY the chapter text between the markers.
- Do NOT add section headers, metadata, or commentary outside the markers.
- Do NOT include the previous chapters in the output.

[Conditional: If Pros/Cons exist]
PROS AND CONS RULE:
- The user may specify pros and cons for each preliminary policy option.
- These points are STRICT and BINDING in meaning.
- You must include ALL user-provided pros and cons.
- You may lightly reword them ONLY for clarity and readability, but their meaning must remain unchanged.
- You must NOT add, remove, reorder, reinterpret, or invent any pros or cons.

User Prompt (Special Recommendations Writer)

Generate the [Chapter Name]-chapter for the memo titled <[Title]>.
Context:

- Requested by: President of the United States (Barack Obama)
- Occasion: Brookings transition memos (late 2008). Moment of political transition and economic uncertainty.
- Purpose: Provide concise, evidence-based, actionable recommendations. Translate complex issues into clear options.
- Memo Title: <Fix the Tax System>

The POLICY OPTIONS chapter should be approximately 300 words long.

EVIDENCE RULE:

- Use the RELEVANCE MASK strictly as the factual evidence base.
- Draw ONLY on information contained within the relevance mask.
- If needed, use the SOURCE SUMMARIES only to clarify meaning, never to introduce new claims.

RECOMMENDATION RULE:

- If the user marks an option as recommended, indicate this by adding '(recommended)' directly after the option title.
- Do NOT write 'not recommended', or any other negative thing about non-recommended options.

WRITING AND TONE:

- Make use of a professional tone.
- The chapter contains a header for each single policy option (if a specific option is recommended it is indicated by saying 'recommended' in parantheses after the header, like: ## 1. Option Title (recommended)).
- Each policy option header is followed by a paragraph describing the option [and a few sentences about its pros and cons - if applicable].
- Don't extend the text with unnecessary narratives, framing or redundancies.
- Hence the ouptut is nothing else then the headers for each option and a text about that respective option.

Here are the preliminary policy options provided by the user (delimited by ###): Option A: <Price Carbon Through a Carbon Tax or Cap-and-Trade System>
Option B: <Reform the Tax Treatment of Employer-Provided Health Insurance> (recommended)
Option C: <Broaden the Tax Base and Address Long-Term Fiscal Imbalances>
... [Options D through I omitted for brevity] ...
Option J: <Simplify Tax Filing Through Return-Free Filing...>
Option K: <Improve IRS Administrative Capacity...>
###
[Optional: Previous Chapter Block]
PREVIOUS CHAPTER (Use this to align the text when writing the chapter): ### PREVIOUS_CHAPTER_BEGIN ###
[Last Chapter Text]
PREVIOUS_CHAPTER_END
Use the SOURCE SUMMARIES (provided after the RELEVANCE MASK) ONLY when further context is strictly necessary to interpret a relevance-mask passage, but NEVER to introduce new claims or arguments.
RELEVANCE MASK:
###
[Dynamic list of relevant passages by Source]
###
SOURCE SUMMARIES:
###
[Dynamic list of summaries by Source]
###

B.2.2 *Standard Prose Chapters.*

System Prompt (Standard Prose Chapters)

You are a Policy Memo Chapter Writing Assistant. Your task is to write one standalone chapter text.

HARD RULES:

- Write ONLY the requested chapter text.
- Do NOT include chapter titles, numbers, headings, or lists of chapters.
- Do NOT reference other chapters explicitly.
- Do NOT introduce structure beyond normal prose.
- Use ONLY the provided relevance mask as factual input.
- NO external knowledge or assumptions.

OUTPUT FORMAT: ### CHAPTER_BEGIN ###
<plain prose text only>
CHAPTER_END

User Prompt (Standard Prose Chapters)

You are writing ONE standalone policy memo chapter titled "[Chapter Title]".

Context:

- Requested by: President of the United States (Barack Obama)
- Occasion: Brookings transition memos (late 2008). Moment of political transition and economic uncertainty.
- Purpose: Provide concise, evidence-based, actionable recommendations. Translate complex issues into clear options.
- Memo Title: <Fix the Tax System>
- Target total length: 1200 words

MEMO STRUCTURE:

(1) **The Situation** (Uniqueness: 1): Summarizes core issue and main recommendation.
(2) **Your Stance** (Uniqueness: 0): Defines policy problem. User Comment: Recall President's campaign commitments, link priorities to current challenge, clarify guiding principles.
(3) **Recommendations** (Uniqueness: 1): Presents evidence, trade-offs, constraints.

[Optional: Previous Chapter Block]
LAST CHAPTER (Style reference: for tone only, do NOT copy):
[Last Chapter Text]
[Condition: If Uniqueness = 0]
User-provided guidance (this replaces the relevance mask):
- [Chapter Comment]
[Condition: If Uniqueness = 1]
Use the following relevance mask as your primary source:

```
### RELEVANCE_MASK_BEGIN ###
[Dynamic list of relevant passages]
### RELEVANCE_MASK_END ###
TASK:
• Write ONLY the prose of this chapter.
• Do NOT include headings, chapter numbers, or lists
  of chapters.
• Do NOT reference other chapters explicitly.
Output format: ### CHAPTER_BEGIN ###
<text>
### CHAPTER_END ###
```

B.3 Final Pipeline Output Memo

The text below represents output memo generated by the final pipeline.

Model Output (Final Pipeline Result)

The Situation

The tax system sits at a crossroads, faced with expiring relief, mounting complexity, and a growing fiscal gap that threatens long-term stability. Nearly all of the 2001 and 2003 tax cuts are scheduled to expire at the end of 2010, a change that would raise marginal rates on many households and complicate planning for families and small businesses just as the economy seeks stability. At the same time, the alternative minimum tax is expanding its reach, with tens of millions of taxpayers likely to be affected as inflation is not indexed and temporary relief fades. In the near term, the deficit picture reinforces the urgency of reform: unified deficits are in the hundreds of billions, and longer-run projections show escalating health, retirement, and entitlement costs consuming a growing share of GDP unless revenues are broadened or spending growth is restrained. The estate tax remains a focal point of reform discussions, as it is highly progressive but scheduled to sunset in 2010 and return in 2011 with a much lower exemption, reshaping incentives at the very top of the income distribution. The magnitude of the challenge is not purely technical: the reform effort will redistribute benefits and costs, requiring credible revenue offsets and careful sequencing to avoid sudden dislocations. Politically, both parties recognize the need to broaden the base and simplify, but meaningful reform will demand political courage, clear tax-forward revenue strategies, and a design that preserves middle-class relief while funding essential priorities. In parallel, policymakers will consider targeted improvements—such as expanding the earned income tax credit or restructuring deductions into more neutral instruments—to advance fairness and growth without compromising fiscal durability.

Your Stance

The incoming administration confronts a tax system that shapes families' decisions, firms' investments, and the government's ability to fund essential priorities. The president-elect's commitments define the core direction: broad-based tax relief for middle-class families, restored fairness in the tax code, and a return to fiscal responsibility, anchored in support for small businesses and job creation at home. He argues that reforming the tax system is not a luxury but a prerequisite for health care and energy strategy, because rational, efficient tax rules are necessary to finance reforms while avoiding excessive bottom-line costs on working Americans.

This stance recognizes that any credible reform must address both immediate pressures and long-term structural needs. Expiring tax cuts threaten planning and economic security; the growth of the alternative minimum tax threatens fairness and predictability; and the nation's large long-term fiscal gap requires a balanced mix of tax reform and responsible revenue generation. Taken together, these factors demand a base-broadening, rate-smart approach that preserves middle-class relief while eliminating or curbing provisions that distort work, saving, and investment.

The guiding principle is fairness paired with growth: modernize the code so relief goes where it is most needed, simplify and neutralize incentives that create complexity, and design revenue steps that are credible and transparent. This means steady, responsible steps now to set a durable tax framework, while maintaining the capacity to fund health care, energy, and other priorities. The president-elect's stance is clear: tax reform that does not tilt the balance away from middle-class relief, but that does repair the system, reduce uncertainty, and restore fiscal sustainability through a disciplined, reform-minded agenda.

Recommendations

1. Price Carbon Through a Carbon Tax or Cap-and-Trade System

To align climate, energy security, and fiscal strategy, price carbon emissions through either a carbon tax or a cap-and-trade system. Both approaches create clear market signals that incentivize reductions in fossil-fuel use and spur innovation in lower-emission technologies. Revenue generated should be used to address distributional consequences associated with higher energy prices, supporting households and communities most affected. The administration should also consider additional market measures to render oil prices commensurate with their broader costs, while recognizing that carbon pricing itself moves oil closer to its true social cost. Unilateral U.S. action can counter energy-security concerns and lay groundwork for international cooperation as part of a broader strategy.

2. Reform the Tax Treatment of Employer-Provided Health Insurance

Reform the tax treatment of employer-provided health insurance as part of a broader health-care reform effort. This alignment seeks tax treatment that supports efficient health financing and fairness, reducing distortions in employer-based coverage while maintaining access and affordability. Framing the reform within a comprehensive health-reform package ensures coherence with broader goals and helps sustain momentum for stable, long-term policy changes that finance reform without imposing unnecessary burdens on workers or employers.

3. Broaden the Tax Base and Address Long-Term Fiscal Imbalances

Broaden the tax base as a central step toward long-term fiscal credibility. A base-broadening, rate-smart approach should accompany measures to address growing structural imbalances, preserving middle-class relief while reducing provisions that distort work, saving, and investment. By expanding the tax base, the administration can improve revenue stability, simplify the system, and provide a more durable platform for financing priorities while maintaining essential fairness and growth incentives.

4. Integrate Corporate and Individual Capital Income Taxation for New Investment
Integrate corporate and individual taxation for new investment to ensure that capital income is taxed once and at the full rate. Integration should apply primarily to income stemming from new corporate investment, with the understanding that dividends and capital gains from new investments would be taxed consistently with the corporate level. The reform should be accompanied by stronger anti-sheltering measures and a concerted effort to reduce or eliminate corporate subsidies that distort investment choices, aligning incentives with productive, long-run growth.
5. Strengthen Enforcement and Eliminate Corporate Tax Shelters and Inefficient Subsidies
Strengthen enforcement and eliminate corporate tax shelters and inefficient subsidies. This entails beefing up anti-sheltering provisions, closing loopholes, and removing subsidies that erode the base. The aim is to curb aggressive tax avoidance, improve compliance, and ensure that corporate tax rules support a fairer and more straightforward system that better reflects true economic activity.
6. Enhance Retirement Saving Through Automatic Enrollment and Universal Matching Incentives
Enhance retirement saving by implementing automatic enrollment and universal matching incentives. Automatic enrollment reduces barriers to participation, while universal matching enhances the incentive to save. Together, these measures are designed to expand saving participation across income groups and improve long-term financial resilience for individuals and families.
7. Convert Deductions and Exemptions Into Refundable Credits Where Appropriate
Convert deductions and exemptions into refundable credits where appropriate to improve revenue adequacy, distributional outcomes, and incentives. Credits reduce tax liability directly and predictably, offering clearer benefits to households at varying income levels. This conversion can simplify the code and ensure that relief is targeted where it is most effective, while maintaining important policy goals tied to education, retirement, and family supports.
8. Reform the Mortgage Interest Deduction Into a Refundable First-Time Homebuyer Credit
Reform the mortgage interest deduction into a refundable first-time homebuyer credit. This shift preserves support for homeownership while transforming the mechanism into a direct, timely benefit for new buyers, improving targeting and reducing distortions associated with housing incentives.
9. Eliminate the State and Local Tax Deduction as Part of AMT Reform
Eliminate the state and local tax deduction as part of AMT reform. This step aims to simplify tax filing and reduce cross-state distortions, aligning the tax base with a more coherent structure. Reforming the SALT deduction within AMT considerations helps streamline calculations and promote greater fairness across taxpayers in different jurisdictions.
10. Simplify Tax Filing Through Return-Free Filing and Consolidation of Overlapping Credits
Simplify tax filing through return-free filing and consolidation of overlapping credits. Return-free filing could be achieved for as many as 50 million taxpayers with relatively minor changes in the code, reducing administrative burdens for filers. Consolidating overlapping credits further reduces complexity, making compliance easier and more transparent while preserving essential benefits.
11. Improve IRS Administrative Capacity and Compliance Enforcement
Improve IRS administrative capacity and compliance enforcement. Strengthening the IRS's operational capabilities enhances voluntary compliance, reduces evasion, and supports more predictable revenue. A modernized administration ensures that a simpler, fairer tax system functions as intended, sustaining the administration's reform agenda over time.

B.4 Final Pipeline Evaluation

The Final Pipeline represents the culmination of the iterative development process. Table 4 presents the ROUGE-L performance metrics for this finalized approach.

Table 4: ROUGE-L Performance of Final Pipeline

Comparison Target	Precision	Recall	F-Measure
Baseline (1205_taxation_memo)	0.2261	0.1693	0.1936
Reference: 1210_global_development	0.1348	0.1182	0.1259
Reference: 1211_financial_memo	0.1435	0.1122	0.1259

Quantitative Analysis. The Final Pipeline achieved the highest performance across all iterations. Against the target gold standard (*1205_taxation_memo*), it reached an F-Measure of 0.1936, significantly outperforming both the Zero-Shot and Preliminary approaches (detailed in subsequent sections). Notably, the Precision score of 22.61% indicates a strong alignment with the specific vocabulary and phrasing used by domain experts in the target memo.

Qualitative Assessment. Qualitative inspection of the generated output confirms a substantial leap in document quality compared to earlier iterations:

- **Professional Style Flow:** The text exhibits a high degree of professional polish. Unlike previous iterations, the narrative follows a clear "inverted pyramid" structure, placing the most critical conclusions up front. The reading flow is clean, logical, and appropriate for a high-level policy audience.
- **Structural Adherence:** The model adhered completely to the specified structure. It successfully generated the distinct chapters without the structural hallucinations (e.g., invented sections) observed in the Zero-Shot baseline.
- **Absence of Artifacts:** The output is entirely free of the meta-commentary and raw source notes that plagued earlier versions.
- **Length Management:** While the output total of 1,390 words slightly exceeded the strict target of 1,200 words, it represents a massive improvement over the uncontrolled output of the Zero-Shot model (which exceeded 1,900 words). The deviation is within an acceptable margin for a complex policy analysis, ensuring depth without excessive verbosity.

C Zero-Shot Baseline

The Zero-Shot baseline utilizes a single interaction where all context, constraints, and source data are provided in one turn.

C.1 Zero-Shot Prompts

Note on Prompt Representation: *While our actual implementation relies on a modular codebase (available for review in the supplementary materials), the prompts presented in this appendix have been consolidated for readability. To clearly demonstrate the context received by the model, dynamic user inputs are displayed here as integrated text within the prompt, rather than as abstract code templates.*

System Prompt (Zero-Shot)

You write a policy memo.
EVIDENCE RULES (HARD):

- Use ONLY the provided SOURCES as factual input.
- No external knowledge, no assumptions, no invented facts.
- If a claim is not supported by SOURCES, omit it.

STRUCTURE RULES (HARD):

- Follow the provided memo structure and chapter order exactly.
- Chapters with Uniqueness=1: base content strictly on source-derived evidence.
- Chapters with Uniqueness=0: use ONLY the user comment for that chapter; do not use sources.
- Do not reference other chapters explicitly inside chapter prose.

POLICY OPTIONS CHAPTER (CONDITIONAL):

- If structure indicates recommendations/options, use provided preliminary options.
- Treat preliminary options as authoritative; expand without rejecting.
- For each option, write a header and short paragraph.
- If marked (recommended), include '(recommended)' in header.

STYLE RULES (HARD):

- Professional, neutral, concise, analytic.
- No rhetorical framing, no filler, no storytelling.

LENGTH RULE:

- Target total memo word count provided.
- Allocate chapter lengths proportionally.

OUTPUT FORMAT (STRICT): ### POLICY_MEMO_BEGIN ###
<Chapter Title>
<chapter text>
...
POLICY_MEMO_END

User Prompt (Zero-Shot)

The following CONTEXT paragraph contains all the contextual information on the policy memo.
CONTEXT:

- Requested by: President of the United States (Barack Obama)
- Occasion: Brookings transition memos (late 2008). Moment of political transition and economic uncertainty.
- Purpose: Provide concise, evidence-based, actionable recommendations. Translate complex issues into clear options.
- Memo Title: <Fix the Tax System>
- Target total length: 1200 words

MEMO STRUCTURE:

(1) **The Situation** (Uniqueness: 1): Summarizes core issue and main recommendation.
(2) **Your Stance** (Uniqueness: 0): Defines policy problem. User Comment: Recall President's campaign commitments, link priorities to current challenge, clarify guiding principles.
(3) **Recommendations** (Uniqueness: 1): Presents evidence, trade-offs, constraints.

PRELIMINARY POLICY OPTIONS: ###
Option A: <Price Carbon Through a Carbon Tax or Cap-and-Trade System>
Option B: <Reform the Tax Treatment of Employer-Provided Health Insurance> (recommended)
Option C: <Broaden the Tax Base and Address Long-Term Fiscal Imbalances>
... [Options D through I omitted for brevity] ...
Option J: <Simplify Tax Filing Through Return-Free Filing...>
Option K: <Improve IRS Administrative Capacity...>
###
SOURCES: ###
SOURCE 1: <Extracted_Sources$text[1]>
SOURCE 2: <Extracted_Sources$text[2]>
###

C.2 Zero-Shot Output Memo

Model Output (Zero-Shot Output Memo)

The Situation

The United States faces a tax system that is increasingly misaligned with its economic and social objectives. Foundational analyses by The Hamilton Project (Achieving Progressive Tax Reform in an Increasingly Global Economy) argue six guiding principles for reform: fiscal responsibility that addresses both spending and taxes; a stronger case for progressivity as inequality rises; a tax system that effectively collects owed revenue; strengthening business taxation; simplification of individual taxes; and using the tax code to advance social policy in a well-designed way.

Yet the fiscal outlook over the 2000s shows a reversal from prior surpluses to deficits, with a unified budget deficit projected in the mid-2000s and long-run fiscal pressures driven by rapidly rising health spending and aging demographics. The 2007-2008 analysis highlights that large, persistent deficits threaten confidence and limit room for new policy.

In parallel, long-standing concerns about the tax system's performance persist: high- and rising-income inequality has eroded progressivity in practice, and the tax code remains highly complex, with expensive noncompliance and a substantial "tax gap" estimated in recent years. Compounding these structural

issues, the tax code has failed to fully restructure in light of globalization and financialization, contributing to perceptions of inefficiency and unfairness.
Several Brookings analyses urge a broad-based reform agenda anchored in base broadening, simplification, and refundable credits that deliver targeted social benefits while preserving revenue and growth. Analyses of reform options also show that returning to a more stable and predictable fiscal path would require a combination of higher revenue and disciplined spending, recognizing the long-run demand placed on health and retirement programs.
The main takeaway from these sources is clear: comprehensive, evidence-based reform is needed to modernize the tax system, reduce complexity, broaden the base, and fund essential priorities while ensuring fairness and economic growth.
Main recommendation (as reflected in the evidence): pursue comprehensive, base-broadening tax reform that simplifies the code, eliminates or reformulates distortive provisions (notably the AMT and targeted deductions), strengthens enforcement to reduce the tax gap, integrates business and individual taxation where appropriate, expands refundable credits to deliver social goals, and uses revenue from pricing externalities (as discussed in climate/energy policy) to protect low- and middle-income households. This approach is designed to preserve growth, reduce inequality, and ensure adequate public investment in health, retirement, education, and infrastructure, while maintaining credibility with financial markets and the broader public.
Notes on foundations of this conclusion:

- The Hamilton Project identifies six principles for progressive tax reform, including fiscal responsibility, stronger progressivity, and simplification, and argues that reform should address both taxes and spending to restore fiscal balance and growth (Source 1).
- Subsequent analyses emphasize that deficits and debt growth threaten confidence, demand discipline, and long-run growth, underscoring the need to raise revenue in a way that supports shared growth (Sources 5, 1).
- The Tax Policy Center work highlights the substantial, and growing, tax expenditure base and the potential gains from converting many deductions into refundable credits, as well as the benefits of simplifying tax filing and eliminating the AMT through revenue-neutral reform (Sources 3, 4, 7, 9).
- Analyses of alternative tax plans emphasize distributional effects and the importance of balancing efficiency with equity through base broadening, refundable credits, and targeted subsidies (Source 3).
- A climate-and-energy lens suggests using price signals (carbon tax or cap-and-trade) to address externalities, while recycling revenues to reduce regressive impacts and invest in R&D and security (Source 2).

Uniquely important policy emphasis from these sources: reform should be comprehensive, evidence-based, and designed to be revenue-neutral in a way that preserves overall fiscal health and substantive social objectives; simplification, base broadening, and refundable credits are central pillars; and enforcement capacity must be strengthened to close the tax gap.

Your Stance

This section articulates policy problem framing in a way consistent with President Obama's priorities and the administration's broader agenda, as described by the user.
The core problem is a tax system that—while capable of raising sufficient revenue in principle—has grown too complex, too opaque, and too favorable to high-income households and certain corporate arrangements. The guiding goals are to (a) restore fairness by broadening the base and making the system more progressive; (b) simplify compliance to reduce administrative burdens; (c) ensure revenue adequacy to fund essential priorities—health care, retirement security, education, and growth-oriented investments; and (d) maintain economic efficiency by avoiding distortions that misallocate capital and labor.
The policy approach should reflect President Obama's broader agenda of opportunity, prosperity, and growth through inclusive, growth-oriented reforms. In this frame, the administration should pursue reform that: (1) broadens and modernizes the tax base; (2) reduces complexity and administrative burden; (3) aligns tax incentives with social objectives (health, saving, education, housing) through refundable credits; and (4) uses prudent revenue-raising measures to support health care, debt reduction, and investments in innovation and energy security.
The stance here emphasizes moving away from selective credits and deductions toward a streamlined, fair, and investable tax structure that supports middle-class opportunity while addressing long-run fiscal imbalances.

Recommendations
The following policy options are presented as a framework for action. Each option includes a concise rationale, key trade-offs, and likely stakeholder considerations.
Option A: Price Carbon Through a Carbon Tax or Cap-and-Trade System
Implement a market-based price on carbon and, optionally, on oil, with revenue recycled through progressive cuts or credits to households and targeted investments in energy efficiency and R&D. A carbon price provides a uniform incentive to reduce emissions and can be designed to be revenue-neutral or partly revenue-recycling to offset distributional impacts.
Pros: aligns climate policy with market incentives; can be designed to be progressive with rebates or credits; supports R&D and energy transition.
Cons: design complexity, potential political resistance, and concerns about competitiveness and global leakage. (Pros/cons integrated)
Option B: Reform the Tax Treatment of Employer-Provided Health Insurance (recommended)
Replace the employer-sponsored health insurance (ESI) tax exclusion with refundable credits or "pay-or-play" mechanisms to reduce distortion, and provide targeted subsidies for those without employer coverage. This aligns incentives toward universal coverage and curbs coverage-based distortions. The design could include a universal mortgage/health subsidy blend and phased implementation to protect employer-sponsored coverage where beneficial.
Pros: improves coverage incentives, expands access, reduces distortions in labor markets; can be financed by broader base broadening elsewhere.

Cons: risk of employers dropping coverage if subsidies don't align with market costs; design must avoid regressive effects and ensure affordability. (Pros/cons integrated)

Option C: Broaden the Tax Base and Address Long-Term Fiscal Imbalances

Adopt base-broadening reforms to reduce reliance on highly distortionary tax preferences, shifting toward a simpler tax base with refundable credits and fewer exemptions. This would be paired with targeted spending restraint and growth-enhancing investments.

Pros: greater efficiency, more stable revenue; reduces tax avoidance opportunities.

Cons: potential political resistance from groups currently benefiting from deductions; requires careful revenue-neutral balancing and program protections. (Pros/cons integrated)

Option D: Integrate Corporate and Individual Capital Income Taxation for New Investment

Adopt a comprehensive business tax reform that taxes new capital income at the enterprise level and aligns treatment of corporate and noncorporate investment; consider a formulary or BEIT-type approach with a phase-out of deferrals.

Pros: reduces sheltering, simplifies, and could raise revenue while avoiding double taxation on new investment.

Cons: implementation complexity; potential transitional issues for existing investments; political economy challenges in reallocating base. (Pros/cons integrated)

Option E: Strengthen Enforcement and Eliminate Corporate Tax Shelters and Inefficient Subsidies

Enhance IRS enforcement resources, close major corporate shelters, and repeal or reform subsidies that distort investment and tax outcomes.

Pros: closes tax gaps, improves fairness, reduces aggressive tax planning; could yield significant revenue and reform credibility.

Cons: enforcement reforms require sustained funding; some reforms could result in higher compliance costs for firms; near-term political battles. (Pros/cons integrated)

Option F: Enhance Retirement Saving Through Automatic Enrollment and Universal Matching Incentives

Implement automatic enrollment in retirement plans with universal matching incentives; simplify saving incentives by consolidating into a universal saver's credit with automatic enrollment features.

Pros: increases participation rates, reduces retirement-risk for middle- and lower-income workers; improves long-term saving.

Cons: administrative costs; potential burdens on employers; design must ensure equity and avoid unintended consequences. (Pros/cons integrated)

Option G: Convert Deductions and Exemptions Into Refundable Credits Where Appropriate

Convert select deductions into refundable credits (or cap deductive benefits) to improve progressivity and simplify administration; target credits to health, education, and retirement saving.

Pros: simpler system, better targeting for low-income households; reduces tax avoidance incentives.

Cons: potential revenue impact and design complexities in credit structure; need careful sequencing with AMT considerations. (Pros/cons integrated)

Option H: Reform the Mortgage Interest Deduction Into a Refundable First-Time Homebuyer Credit

Replace the mortgage interest deduction with a refundable first-time homebuyer credit that scales with home-purchase activity and is easier to administer.

Pros: simplifies homeownership incentives, targets subsidies to new buyers, potential revenue efficiency.

Cons: transition costs and potential housing market distortions; needs careful calibration to avoid price spikes. (Pros/cons integrated)

Option I: Eliminate the State and Local Tax Deduction as Part of AMT Reform

Reform AMT by removing or consolidating the State and Local Tax deduction, aligning with base broadening and simplification goals.

Pros: reduces complexity and enhances progressivity; aligns with AMT reform objectives.

Cons: potential state revenue impacts; political feasibility concerns as states push back. (Pros/cons integrated)

Option J: Simplify Tax Filing Through Return-Free Filing and Consolidation of Overlapping Credits

Move toward return-free filing for simple households and consolidate education, retirement, and family-related credits into a unified framework; improve information reporting to enable pre-filled returns.

Pros: far lower compliance costs; simplified administration; faster refunds.

Cons: requires substantial data-sharing and privacy safeguards; implementation risk and transition costs. (Pros/cons integrated)

Option K: Improve IRS Administrative Capacity and Compliance Enforcement

Strengthen IRS infrastructure, modernize information systems, increase audit capacity for high-income and corporate taxpayers, and improve compliance enforcement where needed for better revenue collection and fairness.

Pros: reduces tax gap, improves enforcement equity, and increases revenue credibility.

Cons: requires sustained funding and political will; potential perception of overreach if not well-targeted. (Pros/cons integrated)

3. Recommendations in Context

The set of options above provides a spectrum from carbon-price-based climate policy alignment (A) to comprehensive base-broadening and simplification (C, D, E, G, J, K). The Obama-era analyses emphasize that a revenue- and growth-friendly reform is feasible if designed with (a) broad base, (b) simplification, (c) refundable credits to deliver social benefits, and (d) credible enforcement to close the tax gap (supported by S1, S6, S7, S9).

The recommended path prioritizes (i) a bold but credible base-broadening strategy that also replaces distortive deductions with refundable credits, (ii) essential health and retirement policy alignment through ESI reform and automatic-saving enhancements, (iii) a credible AMT reform that preserves progressivity and reduces administrative complexity, and (iv) a price-mechanism approach for carbon as a long-run, efficiency-enhancing complement to tax reform, with revenue recycling to protect vulnerable populations (as advanced in Option A).

Stakeholder considerations include health-insurance stakeholders, corporate interests, homeowners, and low- and middle-income families; each option requires careful design to minimize adverse

distributional effects and to preserve incentives for growth and innovation.
The overarching objective is to deliver a simpler, fairer, and more adequate tax system that supports President Obama's priorities for shared growth, opportunity, and fiscal sustainability, while giving the administration a feasible, evidence-based toolkit to address long-run fiscal and social needs.

C.3 Zero-Shot Baseline Evaluation

To assess the baseline performance, we evaluated the output of the Zero-Shot prompt against the target gold standard memo (*1205_taxation_memo*) and two additional gold standard memos from the same domain. Table 5 presents the ROUGE-L metrics, indicating a low lexical overlap with the professional standard.

Table 5: ROUGE-L Performance of Zero-Shot Prompt

Comparison Target	Precision	Recall	F-Measure
Baseline (1205_taxation_memo)	0.1515	0.1612	0.1562
Reference: 1210_global_development	0.1091	0.1360	0.1211
Reference: 1211_financial_memo	0.1091	0.1212	0.1149

Quantitative Analysis. The quantitative results highlight the limitations of the unguided Zero-Shot approach. The F-Measure of 0.1562 against the target memo indicates a significant divergence from the professional standard.

Qualitative Assessment. Beyond the quantitative metrics, a qualitative inspection reveals critical failure modes in the Zero-Shot approach:

- **Length Constraint Violation:** The Zero-Shot model demonstrated a complete inability to adhere to negative length constraints. Despite a strict target of maximum 1,200 words, the model generated approximately 1,940 words.
- **Structural Hallucinations:** The model failed to strictly adhere to the requested chapter structure. For instance, it invented an unrequested section titled *"3. Recommendations in Context"*, disrupting the prescribed flow of the memo.
- **Formatting Violations:** Despite specific instructions regarding the presentation of policy options, the model generated explicit "Pros:" and "Cons:" blocks for every option. This violated the formatting constraints, which required a seamless narrative integration of trade-offs rather than bulleted lists.
- **Meta-Level Style Drift:** The writing style frequently broke the "fourth wall," shifting from a direct policy voice to meta-commentary about the writing process itself, which is inappropriate for a professional policy memo.
- **Raw Artifact Retention:** The output failed to synthesize the source material cleanly, explicitly listing "Notes on foundations of this conclusion" within the main text, rather than integrating them into the prose.

D Preliminary Pipeline

The Preliminary Pipeline represents the initial multi-step approach. Unlike the Final Pipeline, this version is linear rather than modular: it generates a relevance mask, reduces length of the source summaries, expands policy options, and then attempts to write the entire memo in a single pass.

D.1 Prelininary Prompts

D.1.1 Relevance Masking.

System Prompt (Relevance-Masking Assistant)

```
You are a Relevance-Masking Assistant. Your task is
to extract ONLY policy-relevant passages from each
provided source, without paraphrasing.
Definition of a relevant passage:
• Contains core findings, causal mechanisms, main
  arguments, or policy-significant facts.
• Contains quantitative information that affects
  conclusions.
• Contains evidence-based implications (not opinions).
• Exclude narrative background, descriptive filler,
  examples, definitions, or minor details.
Redundancy Rule: If two passages across different
sources express the same argument, choose the
clearer one. Still list the redundant source(s) under
'supporting_sources'.
STRICT Output Format: ### RELEVANCE_MASK_BEGIN ###
<SOURCE_TITLE_1>
[1] <verbatim relevant passage>
supporting_sources: none
[2] <verbatim relevant passage>
supporting_sources: source2
... ### RELEVANCE_MASK_END ###
```

User Prompt (Input)

```
Process the following sources: [Insert Formatted
Sources]
CHAPTER_LIST : [InsertMemoStructure]
Return the output according to the rules in the system
message.
```

D.1.2 Source Summarization.

System Prompt (Source-Summarizing Assistant)

```
You are a Source-Summarizing Assistant. Your ONLY task
is to compress long policy-relevant sources into a
structured, fully faithful summary. Your objective is
to REDUCE TOKEN LENGTH while preserving ALL information
that could influence downstream policy reasoning.
STRICT RULES:
• Summarize ONLY the text inside the delimiters.
• NO interpretations, assumptions, or external
  knowledge.
• Compress wording aggressively (target 30% of
  original length).
• Merge redundant ideas within one source; do NOT merge
  across sources.
```

OUTPUT FORMAT (MANDATORY): ### SOURCE SUMMARY BEGIN ###
<SOURCE i>
TITLE: <title>
CONTENT:
- Bullet points containing ALL substantive information...
</SOURCE i+1>
SOURCE SUMMARY END

D.1.3 Policy Options Expansion (Conditional). This step runs only if the user provides preliminary policy options.

System Prompt (Policy-Options Expansion Assistant)

You are a Policy-Options Expansion Assistant. Your task is to transform the user's preliminary policy ideas into fully-developed, evidence-based policy options.
CORE PRINCIPLES:
- The user's preliminary options are authoritative. Do NOT question or override them.
- Use the RELEVANCE MASK strictly as the factual evidence base.
- If the user marks an option as (recommended), preserve this marker.
- Include ALL user-provided pros and cons (light rewording allowed for clarity).

OUTPUT FORMAT (STRICT): ### POLICY_OPTIONS_BEGIN ###
[1] <Option Title>
description: <2-3 sentences>
pros: ...
cons: ...
POLICY_OPTIONS_END

User Prompt (Policy-Options Expansions Assistant)

Generate structured policy options for the memo titled <[Title]>.
Context: [Addressee, Occasion, Purpose]
Treat the following preliminary policy options as authoritative directions: ### [Insert Preliminary Options] ###
ALL reasoning must be grounded in the RELEVANCE MASK (provided below). [Insert Output from Step 1] [Insert Output from Step 2]

D.1.4 Full Memo Generation. Unlike the Final Pipeline, this step attempts to write the entire memo in a single context window using the artifacts generated in Steps 1, 2, and 3.

System Prompt (Policy Memo Writing Assistant)

You are a Policy Memo Writing Assistant. Objective: Produce concise, fact-based policy memos for the [Addressee Role].
Primary Input Rule:
- Your main evidence foundation is the RELEVANCE MASK.
- You MAY consult the SOURCE SUMMARY to resolve ambiguity, but never to add new claims.
- Any claim not grounded in the relevance mask must be excluded.

Policy Options Input Rule: The Policy Options section must be based EXCLUSIVELY on the provided POLICY_OPTIONS block. Do not add, remove, or merge options.
Style: Plain, neutral, concise, analytical, evidence-based reasoning.

User Prompt (Policy Memo Writing Assistant)

This memo is requested by the [Addressee] in preparation for [Occasion]. It will present the [Purpose] and is titled: <[Title]>.
The memo must follow exactly the structure (delimited by ###): ### [Insert Structure] ###
The memo must incorporate the following Policy Options: [Insert Output from Step 3]
ALL reasoning must be grounded in the RELEVANCE MASK (provided below): [Insert Output from Step 1]
SOURCE SUMMARY: [Insert Output from Step 2]

D.2 Preliminary Output

The text below represents the output generated by the Preliminary Pipeline.

Model Output (Preliminary Pipeline Output Memo)

The Situation
Description: The United States faces a moment of economic uncertainty and structural fiscal challenge that requires urgent, targeted reform of the tax system. The tax code is too complex, unfair, and outdated, with large "tax gaps" from underreporting and shelters that erode revenue and legitimacy. The current regime is increasingly unable to deliver predictable revenue for priorities while maintaining pro-growth incentives. The AMT's expanding reach threatens a growing share of middle- and upper-middle-income filers, complicating planning and fairness. Long-run deficits are driven by past tax cuts and rising health costs, creating a substantial fiscal gap that must be addressed through credible reforms that balance revenue, growth, and equity. Climate policy is intertwined with tax policy: pricing carbon or oil correctly is essential to incentivize reductions, support energy security, and fund investments in basic research and transformative technologies. The administration could also reshape social policy through the tax code (e.g., refundable credits) to strengthen work incentives and reduce poverty, with a focus on metropolitan labor markets. Reform options must be feasible in a constrained political environment, with attention to enforcement, base broadening, and revenue adequacy. Constraints include expiring tax cuts, the growth of the AMT, and the nation's long-term fiscal gap, requiring a balanced mix of reform and revenue generation. Opportunities exist to modernize enforcement, integrate capital and labor taxation, and align incentives with health, education, retirement, and energy goals.
Notes from the relevance mask indicate: (a) the AMT's growing reach and complexity; (b) a three-part climate strategy emphasizing price mechanisms, redirected research support, and international leadership; (c) the case for broad base reform to

improve fairness and revenue adequacy; (d) targeted EITC expansion in metropolitan areas as a tool for poverty reduction and labor participation; (e) the estate tax as a backstop for progressivity; (f) the potential to simplify and reform through refundable credits and base broadening; (g) enforcement strengthening as a critical element of tax reform.

Your Stance
Description: Recalls the president's campaign commitments or expressed priorities and connects these priorities to the current policy challenge. This situates the memo within Obama's agenda.
User Comment: President-elect Obama campaigned on providing broad-based tax relief for middle-class families, restoring fairness in the tax code, and returning to fiscal responsibility. He emphasized support for small businesses, job creation in the United States, and reforming the tax system to address both urgent and long-term structural problems. Obama also acknowledged that repairing the health care and energy systems requires more rational and efficient tax rules. At the same time, he recognized the need to address expiring tax cuts, the growth of the AMT, and the nation's large long-term fiscal gap through a balanced mix of tax reform and responsible revenue generation.

Recommendations
Description: Follow the POLICY OPTIONS INPUT RULE given below.
[1] Price Carbon Through a Carbon Tax or Cap-and-Trade System
description: Adopt a price on carbon through either a carbon tax or cap-and-trade (upstream pricing) to reflect private-sector costs and incentivize reductions. Revenue can be used to offset burdens on households or reduce deficits, with revenue-neutral or revenue-enhancing options and distributional safeguards for low- and middle-income families. The approach should be complemented by increased basic energy research and international engagement to address climate change and energy security.
pros:
-
-
cons:
-
-

[2] Reform the Tax Treatment of Employer-Provided Health Insurance
description: Reform the tax treatment of employer-provided health insurance by reforming the ESI exclusion and broader health subsidies within the tax system. This could involve converting deductions to refundable credits or subsidies to ensure more uniform and progressive treatment, integrated with broader health reform measures such as exchanges and mandates. The reform aims to simplify tax incentives for health coverage while maintaining coverage outcomes.
pros:
-
-
cons:
-
-

[3] Broaden the Tax Base and Address Long-Term Fiscal Imbalances
description: Broaden the tax base to improve fairness, revenue adequacy, and long-run fiscal balance, addressing deficits built by past tax cuts and rising health costs. This entails reducing reliance on specialized deductions, integrating capital and labor taxation more effectively, and strengthening enforcement and reporting. The reform should align with progressivity goals and support sustainable growth.
pros:
-
-
cons:
-
-

[4] Integrate Corporate and Individual Capital Income Taxation for New Investment
description: Integrate corporate and individual capital income taxation for new investment to reduce double taxation and sheltering and improve transparency. The reform would align taxed income with real economic activity, potentially incorporating forms like BEIT or FORMULARY APPORTIONMENT, and strengthen base broadening to support investment incentives. It complements other base-broadening and enforcement efforts for a fairer tax system.
pros:
-
-
cons:
-
-

[5] Strengthen Enforcement and Eliminate Corporate Tax Shelters and Inefficient Subsidies
description: Strengthen tax enforcement to improve revenue collection, compliance, and fairness, while eliminating corporate tax shelters and inefficient subsidies that distort investment and performance. The policy would enhance reporting, close loopholes, and reduce sheltering in corporate tax practices to improve overall system integrity.
pros:
-
-
cons:
-
-

[6] Enhance Retirement Saving Through Automatic Enrollment and Universal Matching Incentives
description: Enhance retirement saving through automatic enrollment and universal matching incentives to boost participation and retirement readiness. The approach should be integrated with tax reform to improve incentives for saving, address health and retirement subsidies within a simplified tax code, and ensure revenue adequacy.
pros:
-
-
cons:
-
-

[7] Convert Deductions and Exemptions Into Refundable Credits Where Appropriate

description: Convert most deductions and exemptions into refundable credits where appropriate to simplify the code and improve progressivity. This shift should be accompanied by consolidating and reforming health, education, and retirement subsidies, and moving toward return-free filing for eligible taxpayers.
pros:
-
-
cons:
-
-

[8] Reform the Mortgage Interest Deduction Into a Refundable First-Time Homebuyer Credit
description: Reform the mortgage interest deduction by replacing it with a refundable first-time homebuyer credit to preserve homeownership incentives while improving budgetary efficiency. The credit would be mechanism to support homebuyers and simplify administration within a broader reform framework.
pros:
-
-
cons:
-
-

[9] Eliminate the State and Local Tax Deduction as Part of AMT Reform
description: Reform the AMT to address its expanding reach and alignment with regular tax rules; as part of AMT reform, consider removing the deduction for state and local taxes to widen the base and reduce shielding, while preserving progressivity.
pros:
-
-
cons:
-
-

[10] Simplify Tax Filing Through Return-Free Filing and Consolidation of Overlapping Credits
description: Simplify tax filing by moving toward return-free filing for eligible taxpayers and consolidating overlapping credits to reduce complexity and compliance costs. The policy would complement other base-broadening reforms and enforcement enhancements to improve revenue collection and ease for filers.
pros:
-
-
cons:
-
-

[11] Improve IRS Administrative Capacity and Compliance Enforcement
description: Improve IRS administrative capacity and compliance enforcement by investing in human capital, information technology, and data analytics; strengthen enforcement and reporting to improve revenue collection, fairness, and compliance across the tax system.
pros:
-
-
cons:
-
-

D.3 Preliminary Pipeline Evaluation

The Preliminary Pipeline, which introduced "Relevance Masking" and "Source Summarization" before a single-pass generation step, was evaluated against the same gold standards. Table 6 summarizes the ROUGE-L performance.

Table 6: ROUGE-L Performance of Preliminary Pipeline

Comparison Target	Precision	Recall	F-Measure
Baseline (1205_taxation_memo)	0.2330	0.1470	0.1803
Reference: 1210_global_development	0.1496	0.1105	0.1271
Reference: 1211_financial_memo	0.1488	0.0980	0.1182

Quantitative Analysis. The Preliminary Pipeline demonstrated a measurable improvement over the Zero-Shot baseline (see Table 5). Notably, the Precision against the target memo increased from 15.15% (Zero-Shot) to 23.30%, and the F-Measure improved from 0.1562 to 0.1803. This suggests that the "Relevance Masking" step (Step 1) successfully filtered out irrelevant noise, ensuring that the generated text utilized vocabulary and concepts more closely aligned with the professional standard.

Qualitative Assessment. Despite the improved grounding in source material, the qualitative inspection reveals that the single-pass generation strategy still suffers from significant structural and cognitive failures:

- **Length Compliance:** A major success of this iteration was adherence to length constraints. The output totaled 1,286 words, closely aligning with the 1,200-word target, a significant improvement over the unconstrained 1,940 words produced by the Zero-Shot model.
- **Formatting Breakdown:** The model failed to process the internal structure of the chapters cleanly. It explicitly retained prompt artifacts in the final text, such as including the label *"Description:"* before chapter texts and leaving the "pros/cons" bullet points entirely empty. This indicates that while the model "knew" the format existed, it lacked the attention span in a long context window to populate it content-wise.
- **Meta-Commentary and Leakage:** The output contained significant "leakage" of the prompt instructions into the final prose. For instance, in the *Your Stance* chapter, the model pasted the raw user instructions ("Recalls the president's campaign commitments...") rather than integrating them into the narrative. Similarly, raw "Notes from the relevance mask" were explicitly printed in the text, rather than being synthesized.
- **Stylistic Improvement:** While the writing style was notably more concise and analytical than the "wall of text" produced by the Zero-Shot baseline, it remained mechanically stiff. The text read more like a structured list of inputs than a cohesive policy narrative, struggling to transition

smoothly between the extracted evidence and the policy arguments.