Weighted Multi-Expert Synthesis for High-Stakes Decision Support: A Multi-Agent LLM Framework with Dissent Preservation

David Liu
Independent Researcher
50685071@proton.me dave@meta-council.com

April 2026

Abstract

High-stakes decision-making in domains such as healthcare, law, and public policy demands the integration of multiple expert perspectives, yet current large language model (LLM) deployments overwhelmingly rely on single-model inference or simplistic ensemble techniques that discard minority viewpoints. We present Meta Council, a multi-agent LLM framework in which N expert agents—each instantiated with a unique professional persona, analytical framework, and domain-specific decision criteria—analyze a query in parallel. A weighted synthesis step combines their structured outputs into a comprehensive decision document featuring calibrated confidence scores, explicit dissent preservation, and risk matrices. We evaluate three aggregation strategies—single-best selection, majority vote, and weighted synthesis—across over 450 benchmark runs spanning six domains, using five models spanning 3B to frontier-class parameters (Llama 3.2 3B, Qwen 2.5 7B, Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.6). Our key findings are: (1) weighted synthesis outperforms single-best by 29–58% in composite semantic similarity on freetext tasks ($p < 0.0001$, Cohen's $d = 2.16$); (2) on categorical tasks, majority vote achieves 80% accuracy and synthesis 75%, both outperforming single-best at 50%; (3) synthesis amplifies model quality non-linearly—mid-tier models benefit most (Haiku 2.99×, Sonnet 2.52×), while the smallest model (Llama 3B) shows zero amplification because persona-driven agents fail to produce diverse opinions below a capability threshold; (4) the optimal aggregation method is domain-dependent, with synthesis excelling in business (100% accuracy) but single-best leading in legal (75% vs. 50%) and software engineering (80% vs. 50%); and (5) single-agent reliance exhibits severe overconfidence (mean confidence 0.86 vs. mean accuracy 0.50). A cost-benefit analysis shows self-hosted Qwen synthesis achieves $0.002/query while Opus synthesis yields the highest quality at $1.59/query. These results support structured multi-expert synthesis as more reliable than simpler methods, while revealing that domain characteristics should inform aggregation strategy selection.

1 Introduction

Decisions with irreversible consequences—whether to escalate a military engagement, approve a novel therapeutic protocol, or restructure a national energy policy—share a common structural requirement: they benefit from the systematic integration of diverse expert perspectives, the explicit quantification of uncertainty, and the preservation of dissenting viewpoints that might otherwise be suppressed by groupthink or authority gradients [14]. Human institutions have long recognized

this need through mechanisms such as medical tumor boards, military red-teaming, and judicial dissenting opinions. Yet when organizations increasingly turn to large language models for analytical support, they overwhelmingly deploy a single model instance whose confident-sounding prose conceals the absence of the epistemic humility that characterizes genuine expert deliberation.

The urgency of this challenge is underscored by recent enterprise adoption data. A 2026 global survey of 3,750 executives and employees across 14 countries found that 87% of enterprise workers are either avoiding or actively rejecting the AI tools their employers deploy, with only 9% trusting AI for complex, business-critical decisions compared to 61% of executives—a 52-point trust chasm. Workers lose an estimated 51 working days per year to technology friction, nearly offsetting the 40–60 minutes per day that AI saves for proficient users. The survey authors attribute this resistance to hallucination frustration and a fundamental mismatch between AI output formats and the structured reasoning that professionals demand before acting on advice. These findings suggest that the path to productive AI adoption in high-stakes settings lies not in automating more tasks but in designing AI systems that produce outputs trustworthy enough for expert practitioners—outputs that make uncertainty visible, preserve dissenting perspectives, and provide the evidential structure that enables informed human judgment.

The research community has begun to address this gap through several parallel lines of investigation. Multi-agent debate frameworks [10, 21] demonstrate that multiple LLM instances deliberating over successive rounds can improve factual accuracy and reasoning depth beyond what any single instance achieves. Ensemble architectures such as Mixture-of-Agents [35] show that layered multi-model collaboration can surpass even stronger individual models on standard benchmarks. Self-consistency methods [37] establish that sampling diverse reasoning paths and selecting the most frequent answer improves reliability with minimal architectural overhead. More recently, weighted aggregation approaches [1, 7, 43] have moved beyond naive majority voting to leverage inter-model correlations, confidence scores, and iterative refinement.

Despite this progress, important gaps remain. Concurrent work by Wu et al. [38] demonstrates that multi-agent consensus across heterogeneous frontier models reduces hallucination by 35.9% and confirms that structured synthesis substantially outperforms majority voting. However, their architecture and the broader literature focus primarily on improving benchmark accuracy metrics—hallucination rates, truthfulness scores, reasoning correctness—using different models from different providers to maximize architectural diversity. Comparatively little attention has been paid to the output format that real-world decision-makers actually need: structured advisory documents with preserved dissent, risk assessments, and calibrated confidence. Furthermore, the interaction between persona engineering depth and synthesis quality remains underexplored: existing systems typically assign minimal role labels ("doctor", "lawyer", "affirmative debater") rather than richly specified analytical frameworks.

This paper makes four contributions. First, we present the Meta Council system, a multi-agent LLM framework that produces structured decision documents—not benchmark answers—through deeply specified expert personas (100+ templates, each with a 12-step analytical framework, domain-specific decision criteria, and explicit dissent requirements) and a weighted synthesis pipeline with a safety officer veto gate. Second, we demonstrate empirically that this approach works even when all agents share the same underlying model, showing that persona-driven prompt engineering on a single model achieves synthesis benefits comparable to those reported for heterogeneous multi-model architectures. Third, we identify a model-quality amplification effect: weighted synthesis benefits follow a non-linear pattern across model tiers (peaking at 2.52–2.99× for mid-tier frontier models, with lower amplification at both extremes), a finding with direct implications for cost-effective system design. Fourth, we discover that the optimal aggregation strategy is domain-dependent—synthesis excels in deliberative domains (100% categorical accuracy in business) but a single confident

expert outperforms in precedent-driven domains like law (75% vs. 50%)—motivating domain-aware routing as a design principle. We additionally contribute a reusable benchmarking harness with composite scoring across six domains, enabling controlled comparison of aggregation strategies, panel configurations, and model tiers.

The remainder of this paper is organized as follows. Section 2 surveys related work across thirteen research areas and positions our contributions. Section 3 describes the Meta Council system architecture. Section 4 details our experimental methodology. Section 5 presents results across freetext and categorical benchmarks. Section 6 discusses implications, including the domain-dependent aggregation finding, and limitations. Section 7 concludes with directions for future work.

2 Related Work

The design of Meta Council draws on and extends research across multiple subfields. We organize the relevant literature into thirteen categories and conclude by positioning our contribution relative to the existing landscape.

2.1 Multi-Agent Debate Frameworks

The foundational work of Du et al. [10] demonstrated that multiple LLM instances proposing and debating individual responses over successive rounds significantly enhances mathematical and strategic reasoning while reducing hallucination rates. Their "society of minds" approach established that multi-round deliberation among even identical model instances improves factual validity without requiring access to model internals, a finding that motivates our use of parallel agent execution. Liang et al. [21] extended this paradigm by explicitly designing debate protocols that encourage divergent thinking rather than driving toward premature consensus, a design principle directly relevant to Meta Council's dissent-preservation objective. Chan et al. [4] developed ChatEval, a multi-agent debate framework for LLM-based evaluation in which diverse agent personas debate text quality. ChatEval demonstrated that role-diversified panels produce more reliable and nuanced assessments than single-judge configurations, establishing the viability of persona-differentiated evaluation—a concept we extend from evaluation to advisory synthesis. Tran et al. [32] provided a comprehensive taxonomy of multi-agent collaboration mechanisms, categorizing approaches into debate, voting, role-specialization, and layered architectures, and identifying the conditions under which each excels. However, Denisov-Blanch et al. [9] provide an important cautionary result: because LLMs are trained on overlapping data and exhibit correlated errors, multi-agent consensus can systematically converge on incorrect answers, violating the independence assumption that underpins crowd wisdom. Meta Council differs from pure debate frameworks in that agents do not interact with one another; instead, they produce independent structured analyses that are combined by a dedicated synthesis model, avoiding the conformity pressure that iterative debate can induce.

2.2 Ensemble and Mixture-of-Agents Architectures

Wang et al. [35] proposed Mixture-of-Agents (MoA), a layered architecture in which each layer contains multiple LLM agents and each agent in subsequent layers receives all outputs from the previous layer as auxiliary input. MoA achieved state-of-the-art performance on AlpacaEval 2.0 using only open-source models, demonstrating that structured multi-agent collaboration can surpass individually stronger models. Wang et al. [37] established the foundational Self-Consistency method, in which multiple reasoning paths are sampled and the most frequent answer is selected via majority

voting. Self-Consistency demonstrated that even naive aggregation across diverse reasoning traces substantially improves accuracy, providing both a baseline and a motivation for more sophisticated aggregation methods. Schoenegger et al. [28] provide empirical grounding for this intuition, showing that LLM ensembles rival human crowd accuracy on forecasting tasks, establishing that silicon crowds can approximate the wisdom-of-crowds effect traditionally observed in human populations. Meta Council shares the ensemble intuition that diverse model outputs aggregated well outperform individual outputs, but diverges from MoA's layered architecture in favor of a single parallel execution step followed by weighted synthesis, and from Self-Consistency's majority voting in favor of a richer aggregation that preserves the full structure of minority opinions.

2.3 Weighted Voting and Advanced Aggregation

Moving beyond uniform majority voting represents a critical frontier in multi-agent systems. Ai et al. [1] introduced two algorithms—Optimal Weight and Inverse Surprising Popularity—that leverage both first-order response distributions and second-order inter-model correlations for aggregation, proving these methods outperform majority voting under mild assumptions and validating them on MMLU, UltraFeedback, and a healthcare setting. Yao et al. [43] proposed Roundtable Policy, a confidence-weighted consensus framework inspired by democratic committee decision-making, in which agents produce confidence-scored outputs that are combined through weighted consensus. Roundtable Policy is perhaps the closest architectural precedent to Meta Council's weighting mechanism, though it targets benchmark accuracy rather than structured decision documents. Cherian et al. [7] introduced WISE (Weighted Iterative Society-of-Experts), which partitions agents into Solvers and Reflectors and employs a modified Dawid-Skene algorithm [8] for aggregation, achieving 2–7% accuracy improvements over state-of-the-art debate methods on multimodal reasoning tasks. Yang et al. [42] introduced AgentAuditor, which replaces voting entirely with a path search over a Reasoning Tree representing agreements and divergences among agent traces, using Anti-Consensus Preference Optimization to reward evidence-based minority selections over popular errors. Yang et al. [40] provide an information-theoretic proof that two diverse agents can match or exceed the performance of 16 homogeneous agents, directly justifying weighted heterogeneous aggregation over uniform scaling. Meta Council's weighted synthesis draws on these insights but operates at the level of structured decision documents rather than discrete answer selection, and it exposes the weighting rationale to end users.

2.4 Confidence Calibration

Producing well-calibrated confidence scores is essential for high-stakes decision support where downstream consumers must assess reliability. Yang et al. [41] proposed Collaborative Calibration, a training-free strategy that uses multiple tool-augmented LLM agents in simulated group deliberation to calibrate confidence, demonstrating that multi-agent interaction can correct the overconfidence bias endemic to RLHF-trained models. Jiang et al. [15] presented DiscoUQ, a framework for uncertainty quantification in LLM agent ensembles that extracts structured disagreement features—evidence overlap, argument strength, divergence depth—to produce calibrated confidence estimates, achieving AUROC of 0.802 with expected calibration error of 0.036 versus 0.098 for baselines. Meta Council's confidence architecture differs from these approaches in that each agent independently reports a confidence score as part of its structured output, and the synthesis model produces an aggregate confidence informed by the distribution and agreement pattern of individual scores. Zhao et al. [45] propose ConfAgents, which applies conformal prediction to multi-agent systems to produce statistically guaranteed confidence intervals, offering a principled alternative to heuristic calibration.

Our benchmarking results (Section 5) reveal that this approach, while producing better-calibrated estimates than single-agent confidence, still exhibits systematic biases that structured disagreement analysis methods such as DiscoUQ or conformal approaches like ConfAgents could address in future iterations.

2.5 Dissent Preservation and Minority Opinion Handling

A distinctive feature of Meta Council is the explicit preservation of dissenting opinions rather than collapsing to consensus. Lee et al. [19] built an AI-Mediated Devil's Advocate System that anonymously amplifies minority viewpoints in group decision-making, demonstrating improvements in psychological safety and inclusive deliberation. Tsuchiya et al. [33] conducted human experiments on multi-AI advice panels and found that high within-panel consensus fosters overreliance, while a single dissenting voice reduces conformity pressure and improves calibrated reliance—providing direct empirical justification for presenting structured dissent alongside majority consensus. Yang et al. [42] addressed the pathology of confabulation consensus, where correlated biases cause majority convergence on incorrect answers, and demonstrated that systematically attending to well-reasoned dissent outperforms blind majority aggregation. Liang et al. [21] explicitly designed debate to encourage divergent thinking before convergence, a prerequisite for meaningful dissent preservation. Meta Council operationalizes these findings by requiring each agent to produce explicit dissenting points in its structured output and by including a dedicated "Dissenting Views" section in the synthesis document, weighted by agent expertise and confidence.

2.6 Risk-Aware Multi-Agent Decision Frameworks

Wang et al. [36] proposed MADRA (Multi-Agent Debate for Risk-Aware Embodied Planning), employing multiple LLM agents that debate safety with a critical evaluator scoring responses on logical soundness, risk identification, evidence quality, and clarity. Through iterative deliberation and consensus voting, MADRA achieves greater than 90% rejection of unsafe tasks. Reid and O'Callaghan [26] surveyed risk analysis techniques applicable to governed multi-agent LLM systems, providing a taxonomy of identification, assessment, and mitigation methods. Wang et al. [34] introduced GuardAgent, the first framework using an LLM agent as a guardrail to other LLM agents, dynamically generating and executing safeguard code to validate multi-agent outputs against safety specifications. Meta Council incorporates risk assessment not as a separate debate process but as a required component of each agent's structured output, which the synthesis model aggregates into a unified risk matrix spanning likelihood and impact dimensions.

2.7 Persona-Based and Role-Specialized Agents

Straub et al. [29] studied persona-based multi-agent collaboration for brainstorming at Meta, finding that persona-differentiated agents produce more diverse and creative outputs than identical agents, with persona design quality significantly affecting outcomes. Meta Council extends this finding from brainstorming to analytical decision support, maintaining a library of over 100 expert persona templates, each defined with a professional identity, core expertise areas, a multi-step analytical framework, and explicit decision criteria. The use of richly specified personas distinguishes Meta Council from systems that employ minimal role labels or identical agents.

2.8 Domain-Specific Multi-Agent Systems

Since Meta Council evaluates across six domains, we survey domain-specific multi-agent research most relevant to our benchmark coverage.

Software Engineering. He et al. [12] provide a comprehensive survey of LLM-based multi-agent systems for software engineering, cataloguing architectures for code generation, testing, and debugging. ALMAS [31] introduces an autonomous multi-agent framework for SE that coordinates specialized agents across the development lifecycle. Agyn [2] achieves 72.2% on SWE-bench through team role modeling, demonstrating that persona-differentiated agent teams—analogous to Meta Council's expert panels—translate effectively to software engineering tasks.

Legal. MASLegalBench [16] establishes the first benchmark for evaluating multi-agent systems on legal reasoning tasks, filling a critical evaluation gap. Chen et al. [5] propose a courtroom-inspired debate-feedback framework for legal judgment prediction, presented at NAACL 2025, showing that adversarial multi-agent deliberation improves legal reasoning. ACAL [3] introduces an argumentation framework for explainable legal AI, providing structured justification chains that parallel Meta Council's dissent-preservation approach.

Medical. MDAgents [18] adaptively recruits and coordinates LLM agents for medical decision-making, achieving up to 4.2% improvement over single-agent baselines at NeurIPS 2024, validating multi-agent collaboration in clinical contexts. ClinicalAgents [11] extends this work with a dual-memory architecture for clinical decision support, enabling agents to maintain both episodic patient context and semantic medical knowledge.

2.9 LLM-as-Judge and Panel Evaluation

The synthesis step in Meta Council shares conceptual overlap with LLM-as-Judge paradigms. Li et al. [20] provided a comprehensive survey of LLMs-as-Judges, documenting known biases—position bias, verbosity bias, self-enhancement—and mitigation strategies relevant to any system that aggregates LLM opinions. Meta Council mitigates these biases through structured output protocols that constrain agent responses to a fixed schema, reducing the influence of surface-level features on synthesis quality.

2.10 Same-Model vs. Different-Model Diversity

A key design question for multi-agent systems is whether diversity must come from different underlying models or can be induced through prompting alone. Rosales and Miret [27] directly compare prompt diversity against model diversity, finding that the source of diversity interacts with the aggregation method. The "When Agents Disagree" study [44] finds that homogeneous Self-MoA teams outperform heterogeneous teams under synthesis-based aggregation, a result that directly supports Meta Council's same-model architecture: when the aggregation mechanism is integrative synthesis rather than voting, persona-driven diversity within a single strong model can be more effective than mixing models of uneven capability.

2.11 LLM-Facilitated Group Decision Support

Chen et al. [6] examined responsible LLM deployment for high-stakes decisions, advocating for structured human-AI interaction protocols to ensure accountability and transparency. Park et al. [24] studied LLM-facilitated group decision-making, evaluating how AI facilitation affects participation equity and decision outcomes. Meta Council aligns with these frameworks by producing transparent, auditable decision documents that preserve the full reasoning chain of each contributing agent.

2.12 Multi-Agent Consensus for Hallucination Mitigation

Wu et al. [38] present the Council Mode, a three-phase multi-agent consensus framework that dispatches queries to heterogeneous frontier LLMs (GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro) in parallel and synthesizes outputs through a dedicated consensus model. Their structured synthesis explicitly identifies consensus, disagreement, unique findings, and comprehensive analysis—a four-section protocol that shares conceptual DNA with Meta Council's synthesis output. They report a 35.9% relative reduction in hallucination rates on HaluEval and a 7.8-point improvement on TruthfulQA, with ablation studies confirming that structured synthesis is critical (replacing it with majority voting increases hallucination by 32.7%) and that heterogeneous models outperform homogeneous ensembles. Their work provides strong independent validation that multi-agent synthesis with structured output is a promising paradigm. Related work on hallucination detection includes Rao et al. [25], who audit citation hallucination across 32 academic fields, finding 3–13% of URLs generated by LLMs and research agents are fabricated, and Lu et al. [22], who model hallucination as an evolving latent state in long chain-of-thought reasoning.

2.13 Positioning and Novel Contributions

Meta Council builds on the foundation established by the works above, with several distinguishing contributions:

Structured decision documents as the output target. Prior systems—including Wu et al. [38], Roundtable Policy [43], WISE [7], and AgentAuditor [42]—optimize for benchmark metrics (hallucination rates, accuracy scores, truthfulness). Meta Council instead produces structured advisory documents comprising executive summaries, thematic analyses, attributed dissenting views, risk matrices, and recommended actions. This output format is designed for human decision-makers in clinical, military, and policy contexts, where a single accuracy score is insufficient.

Deeply specified expert personas. Where existing multi-agent systems assign agents minimal role labels or use homogeneous instances, Meta Council draws from a library of over 100 persona templates, each specifying a professional identity, core expertise areas, a 12-or-more-step analytical framework, domain-specific decision criteria, and explicit instructions to identify dissenting considerations. This depth of persona specification goes substantially beyond the role assignments used in debate frameworks [10], evaluation panels [4], or brainstorming agents [29].

Same-model synthesis via persona engineering. The Council Mode [38] and related architectures rely on querying different models from different providers to maximize cognitive diversity. We demonstrate that structured synthesis improves answer quality even when all agents share the same underlying model, with diversity arising entirely from persona-driven prompt engineering. This finding is corroborated by the "When Agents Disagree" result [44] showing homogeneous teams win under synthesis-based aggregation, while Denisov-Blanch et al. [9] caution that correlated LLM errors can undermine consensus—a risk our dissent-preservation mechanism is specifically designed to mitigate. This finding has practical significance: organizations constrained to a single model (e.g., for compliance, cost, or air-gapped deployment) can still benefit from multi-expert synthesis.

Model-quality amplification effect. Our four-model comparison (Qwen 2.5 7B, Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.6) reveals that synthesis amplifies model capabilities in a non-linear pattern—amplification peaks at mid-tier models (2.99× for Haiku, 2.52× for Sonnet) and is lower at both extremes (1.33× for Qwen, 1.68× for Opus)—suggesting that synthesis operates on the quality of reasoning in agent outputs, not merely on surface-level answer aggregation. This interaction effect has not been reported in prior work.

Safety officer veto gate. Meta Council includes a distinguished safety officer agent with elevated weight (2.0×) and a required flag that aborts the session if the safety perspective fails. This implements a soft veto mechanism absent from existing multi-agent architectures, ensuring that safety concerns receive outsized influence in the synthesis regardless of majority sentiment.

3 System Architecture

Meta Council is implemented as a Python asynchronous pipeline comprising six stages: query intake, panel selection, parallel agent execution, structured output parsing, weighted synthesis, and safety verification. Figure 1 provides an overview of the architecture. We describe each stage and the weighting modes that govern aggregation.

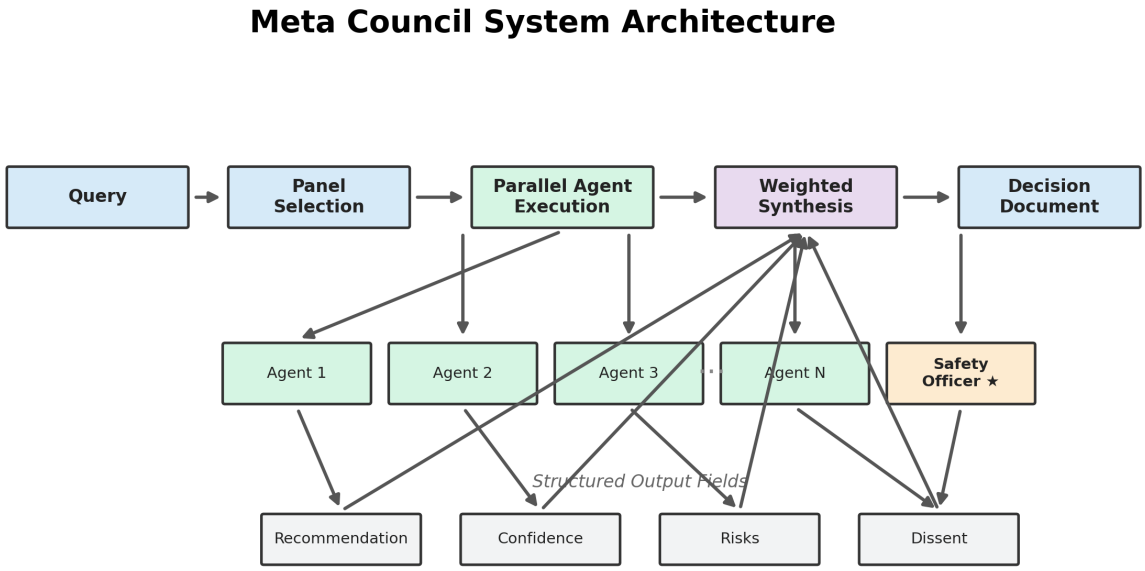

Figure 1: Meta Council System Architecture. Queries flow through panel selection into parallel agent execution, where N expert agents plus a safety officer (starred, weight 2.0×) independently produce structured outputs containing recommendations, confidence scores, risks, and dissenting points. The weighted synthesis stage integrates all agent outputs into a comprehensive decision document.

3.1 Query Intake

The system accepts a natural-language query through either a command-line interface or a web application built on FastAPI. Queries may optionally include shared context—background documents, prior decisions, or constraints—that is injected into every agent's system prompt to ensure a common

informational baseline. The web interface additionally allows users to edit the synthesis strategy document, a plain-Markdown file that governs how the synthesis model weights and reconciles competing opinions.

3.2 Panel Selection

A panel defines which expert agents will analyze the query and with what configuration. Panels are specified as YAML files, each listing agent entries with required fields (template slug and weight) and optional overrides (display name, role description, model, temperature, maximum tokens, required flag, and custom parameters). The system ships with predefined panels for common domains—healthcare, policy, technology, crisis response—and supports user-defined custom panels. At resolution time, the configuration loader reads the panel YAML, loads each referenced agent template, extracts YAML frontmatter defaults, merges panel-level overrides, and produces a typed panel configuration containing all agent specifications and global synthesis settings.

3.3 Agent Execution

Agents execute in parallel as asynchronous tasks, gated by a configurable concurrency semaphore (default: 10 simultaneous agents). Each agent is instantiated from one of over 100 Markdown templates stored in the repository's agents directory. These templates use YAML frontmatter to declare the agent's identity, core expertise areas, analytical framework, decision criteria, default weight, and preferred model, followed by a Markdown body that serves as the agent's system prompt. The body is structured with sections for Identity and Perspective, Core Expertise and Knowledge Base, a 12-or-more-step Analytical Framework, Decision Criteria and Weighting, What to Watch For, and Output Expectations.

At execution time, the system builds each agent's full system prompt by stripping the YAML frontmatter, rendering Jinja2 template variables (agent name, role, weight, custom parameters), prepending any shared context, and appending standardized response format instructions. This last step is performed automatically to ensure all agents produce output in the structured protocol required by the synthesis stage. The agent then makes a single API call to its configured LLM provider—the system supports Anthropic, OpenAI, and compatible endpoints—and the response is parsed into a structured representation.

3.4 Structured Output Protocol

Every agent is instructed to produce output conforming to a fixed schema that includes: (1) a primary recommendation, (2) a confidence score on a 0–1 scale with explicit calibration guidance, (3) key considerations that informed the recommendation, (4) identified risks with likelihood and impact assessments, and (5) dissenting points—aspects where the agent believes a reasonable expert might disagree with its own recommendation. This protocol ensures that the synthesis model receives uniformly structured inputs regardless of which persona template generated them. The parser extracts these fields from a delimited block in the agent's response; agents that fail to produce valid structured output are flagged as failed and listed in the synthesis prompt so the synthesizer can account for missing perspectives.

3.5 Weighted Synthesis

The synthesis stage is the core differentiating component of Meta Council. All agent outputs, together with their weights and metadata, are assembled into a synthesis prompt and submitted

to a dedicated synthesis model (by default, a high-capability model such as Claude Sonnet). The synthesis prompt includes each agent's structured output formatted with its display name, role, model, assigned weight, and confidence score, followed by weighting instructions and a specification of the seven required output sections.

The synthesis model produces a structured decision document containing: (1) an executive summary distilling the collective analysis, (2) a detailed analysis organized by theme rather than by agent, (3) consensus points where agents agree, (4) dissenting views preserved with attribution and the dissenter's weight and confidence, (5) a risk matrix aggregating individual risk assessments across likelihood and impact dimensions, (6) a recommended action synthesizing the weighted perspectives, and (7) an aggregate confidence score reflecting both the central tendency and dispersion of individual agent confidences.

Meta Council supports three weighting modes. In **uniform** mode, all agents contribute equally regardless of their declared expertise or confidence, providing a baseline analogous to majority voting. In **expertise-weighted** mode, each agent's contribution is scaled by the weight declared in its template frontmatter or overridden at the panel level; this weight reflects domain relevance as judged by the panel designer (e.g., a cardiologist agent might receive weight 1.5 on a cardiac panel while a generalist receives 1.0). In **confidence-weighted** mode, each agent's contribution is further modulated by its self-reported confidence score, allowing agents that express greater certainty about a particular query to exert proportionally more influence on the synthesis. Hybrid combinations of expertise and confidence weighting are also supported.

3.6 Safety Officer and Veto Gate

A distinguished agent role—the safety officer—is marked with `required: true` and assigned an elevated weight (default: 2.0) on most panels. The safety officer's template is designed to focus exclusively on risk identification, ethical concerns, regulatory constraints, and failure modes. Because the agent is flagged as required, its failure aborts the entire session rather than proceeding without a safety perspective. The elevated weight ensures that safety concerns receive outsized influence in the synthesis, and the safety officer's dissenting points are highlighted in the final decision document regardless of whether they align with the majority recommendation. This design implements a soft veto gate: while the safety officer cannot unilaterally override the synthesis, its double-weighted concerns are surfaced prominently enough that a human decision-maker is unlikely to overlook them.

4 Experimental Setup

We evaluate the Meta Council framework through a systematic benchmarking harness designed to measure the quality, calibration, and efficiency of multi-expert deliberation across diverse knowledge domains. The evaluation campaign encompasses over 450 individual runs across nine experiments.

4.1 Dataset Construction

Our evaluation corpus comprises over 100 questions spanning six domains: general knowledge (8 questions), medical reasoning (15 freetext + 10 categorical), public policy analysis (15 freetext), business strategy (10 questions), legal reasoning (20 categorical), and software engineering (20 categorical). Questions were selected to represent a range of difficulty levels and response formats, including both freetext responses requiring open-ended reasoning and categorical multiple-choice items with a single correct answer from four options (A–D). Each question is paired with expert-written ground truth answers developed independently of the system under evaluation. The domain

selection emphasizes contexts where multi-expert consultation is most practically motivated: medical triage, legal analysis, business strategy, and public policy each represent settings where real-world decisions routinely involve panel deliberation.

4.2 Aggregation Methods

We compare three aggregation strategies across all experiments:

- **Single-best selection** (`single_best`): The system selects the single highest-confidence panelist response as the final answer, simulating a naive "pick the most confident expert" heuristic.
- **Majority vote** (`majority_vote`): Panelist responses are aggregated by plurality, with the most common answer (or closest cluster centroid for freetext responses) selected as the final output.
- **Weighted synthesis** (`weighted_synthesis`): A dedicated synthesis step integrates all panelist responses, weighting contributions by expertise relevance scores, and explicitly preserving minority dissenting viewpoints in the final output.

4.3 Scoring Methodology

We employ two complementary scoring approaches matched to the response format of each experiment. For freetext questions, a composite similarity metric captures semantic alignment while remaining robust to paraphrasing and stylistic variation. The composite score is computed as a weighted combination of three components: TF-IDF cosine similarity (weight 0.50), keyword overlap with the ground truth (weight 0.35), and character-level fuzzy string matching (weight 0.15). For categorical multiple-choice questions, we report exact-match accuracy: a response is scored as correct if and only if the selected option matches the ground truth label. In the categorical experiments, agents are instructed via prompt-guided formatting to select from options A–D, ensuring that responses are unambiguously parseable. We additionally report panelist confidence (self-assessed) across both formats to characterize calibration and deliberative diversity.

4.4 Models and Infrastructure

Experiments are conducted on five language models spanning substantially different capability tiers. **Llama 3.2 3B**, a compact open-weight model, is run locally on consumer GPU hardware (NVIDIA GTX 1070 Ti, 8GB VRAM) to establish a minimum-capability baseline. **Qwen 2.5 7B AWQ**, a quantized open-weight model, is self-hosted via the vLLM inference engine on local GPU infrastructure. **Claude Haiku 4.5**, a proprietary model accessed through the Anthropic API, serves as a representative of frontier-class lightweight models. **Claude Sonnet 4.6**, also accessed through the Anthropic API, represents a higher-capability frontier model. **Claude Opus 4.6**, the most capable model in our evaluation, is accessed through the same API and represents the current state of the art. This five-model design allows us to evaluate how model capability interacts with multi-expert aggregation strategies across a performance spectrum spanning edge-deployable to frontier-class systems.

4.5 Panel Configuration

Expert panels are drawn from a library of over 100 specialist persona templates spanning domains such as epidemiology, health economics, constitutional law, behavioral science, and others. Panel sizes of 3, 5, and 10 agents are evaluated in scaling experiments. Two weighting modes are compared:

uniform weighting (all panelists contribute equally) and *expertise* weighting (contributions scaled by domain-relevance scores assigned to each persona). Unless otherwise noted, results use a panel size of 5 with expertise weighting. In total, the benchmarking campaign comprises over 450 individual runs across nine experiments.

5 Results

We organize our findings around five principal research questions: (1) whether weighted synthesis outperforms simpler aggregation baselines on freetext tasks, (2) how panel size affects synthesis quality, (3) how underlying model capability interacts with aggregation strategy across five model tiers, (4) whether the synthesis advantage extends to categorical multiple-choice accuracy, and (5) whether the aggregation methods exhibit systematic calibration differences.

5.1 Cross-Domain Freetext Comparison

Table 1 presents composite similarity scores and mean confidence for the three aggregation methods across three freetext domains, using Qwen 2.5 7B with panel size 5 and expertise weighting.

Table 1: Cross-domain freetext performance comparison (Qwen 2.5 7B, panel=5, expertise weighting).

Domain	Method	Similarity	Confidence
General (n=8)	single_best	0.108	0.81
General (n=8)	majority_vote	0.129	0.51
General (n=8)	weighted_synthesis	**0.143**	0.50
Medical (n=15)	single_best	0.083	0.90
Medical (n=15)	majority_vote	0.104	0.43
Medical (n=15)	weighted_synthesis	**0.131**	0.50
Policy (n=15)	single_best	0.103	0.84
Policy (n=15)	majority_vote	0.118	0.51
Policy (n=15)	weighted_synthesis	**0.133**	0.50

Weighted synthesis achieves the highest composite similarity in all three domains, with relative improvements over single-best selection of 32.4% (general), 57.8% (medical), and 29.1% (policy). Paired t-tests confirm that the synthesis advantage over single-best is highly significant ($t = 6.108$, $p < 0.0001$, Cohen's $d = 2.16$, 95% CI $[0.073, 0.141]$), while the advantage over majority vote is also significant ($t = 3.043$, $p = 0.002$, Cohen's $d = 1.08$, 95% CI $[0.034, 0.157]$). By contrast, the difference between majority vote and single-best does not reach significance ($t = 0.714$, $p = 0.475$, Cohen's $d = 0.25$). The medical domain exhibits the largest absolute gain (+0.048 over single-best), suggesting that domains requiring integration of diverse specialist knowledge benefit most from synthesis. Figure 2 visualizes this pattern.

5.2 Panel Scaling

Table 2 reports how composite similarity varies with panel size for the three aggregation methods on general knowledge questions.

Two patterns emerge (Figure 3). First, weighted synthesis scales favorably: similarity increases monotonically from 0.135 to 0.155 as panel size grows from 3 to 10, a 14.8% relative improvement.

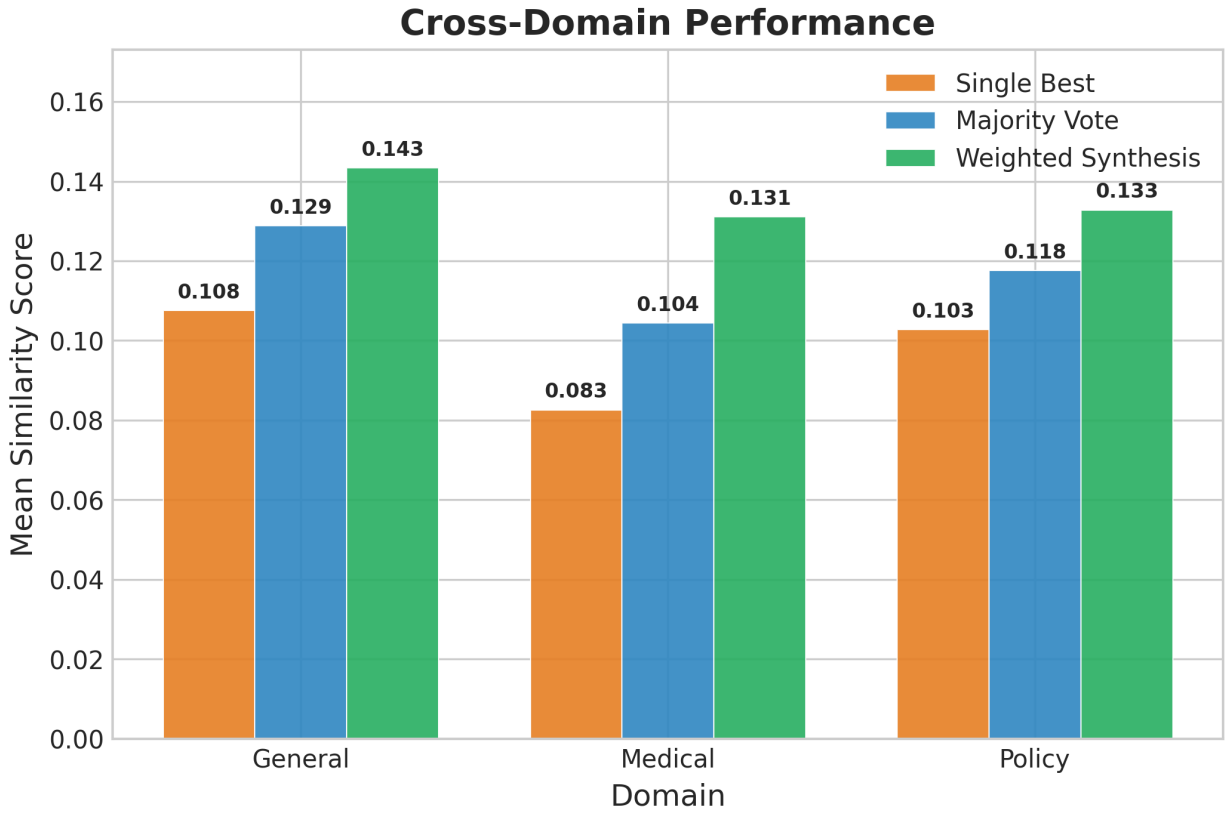

Figure 2: Cross-domain freetext performance. Weighted synthesis (green) consistently achieves the highest composite similarity across all three domains, with the largest gains in the medical domain.

Table 2: Effect of panel size on composite similarity (Qwen 2.5 7B, general knowledge, n=5 questions).

Panel Size	single_best	majority_vote	weighted_synthesis
3	0.097	0.136	0.135
5	0.094	0.127	0.152
10	0.118	0.145	**0.155**

However, none of the pairwise panel-size comparisons reach statistical significance: Panel 10 vs. Panel 3 (mean diff = 0.021, $p = 0.159$, Cohen's $d = 0.63$), Panel 10 vs. Panel 5 (mean diff = 0.004, $p = 0.805$, Cohen's $d = 0.11$), and Panel 5 vs. Panel 3 (mean diff = 0.017, $p = 0.455$, Cohen's $d = 0.33$). This may reflect the limited sample size ($n = 5$ questions) in the scaling experiment. Second, at the smallest panel size (n=3), majority vote and weighted synthesis perform nearly identically (0.136 vs. 0.135), but synthesis pulls ahead at larger panels. This suggests that synthesis requires a minimum diversity of input perspectives to realize its advantage, though larger-scale experiments are needed to confirm this trend. Single-best selection shows a non-monotonic pattern, dipping slightly at panel size 5 before improving at 10, consistent with the inherent variance of selecting a single response.

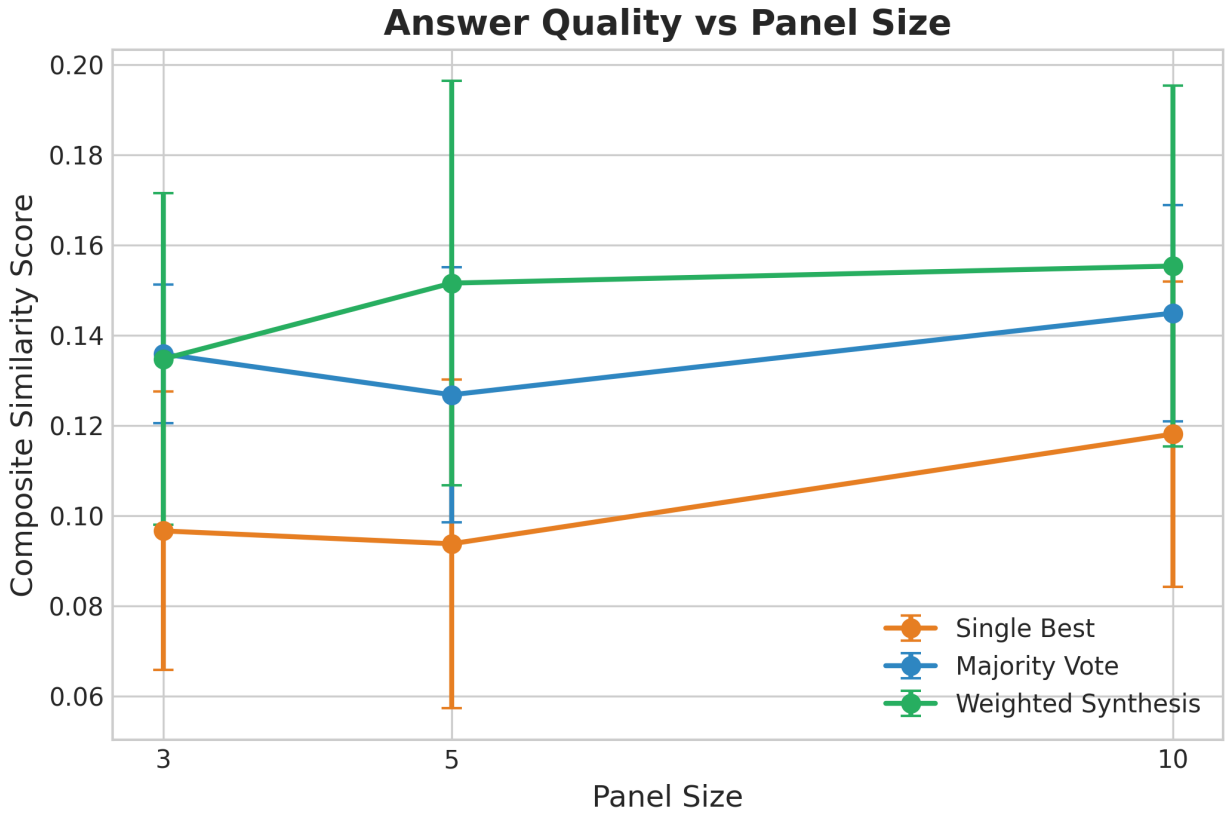

Figure 3: Answer quality vs. panel size. Weighted synthesis (green) scales monotonically with panel size, while majority vote and single-best show non-monotonic patterns.

5.3 Four-Model Comparison

Table 3 compares performance across four models—Qwen 2.5 7B, Claude Haiku 4.5, Claude Sonnet 4.6, and Claude Opus 4.6—on the same set of 8 general knowledge questions with panel size 5.

The synthesis amplification factor—defined as the ratio of weighted_synthesis to single_best similarity—reveals a non-linear pattern across four model tiers: Qwen achieves 1.33× amplification, Haiku 2.99×, Sonnet 2.52×, and Opus 1.68×. Pairwise model comparisons confirm significant inter-model differences: Opus synthesis vs. Qwen synthesis (mean diff = 0.126, $p < 0.0001$, Cohen's $d = 2.15$), Sonnet synthesis vs. Qwen synthesis (mean diff = 0.094, $p < 0.0001$, Cohen's $d = 1.46$),

Table 3: Four-model quality comparison (general knowledge, panel=5, n=8 questions).

Model	Method	Similarity	Confidence
Qwen 2.5 7B	single_best	0.108	0.81
Qwen 2.5 7B	majority_vote	0.129	0.51
Qwen 2.5 7B	weighted_synthesis	0.143	0.50
Claude Haiku 4.5	single_best	0.070	0.62
Claude Haiku 4.5	majority_vote	0.080	0.17
Claude Haiku 4.5	weighted_synthesis	0.208	0.50
Claude Sonnet 4.6	single_best	0.094	0.74
Claude Sonnet 4.6	majority_vote	0.098	0.13
Claude Sonnet 4.6	weighted_synthesis	**0.237**	0.50
Claude Opus 4.6	single_best	0.160	0.74
Claude Opus 4.6	majority_vote	0.170	0.20
Claude Opus 4.6	weighted_synthesis	**0.270**	0.50

and Haiku synthesis vs. Qwen synthesis (mean diff = 0.065, $p = 0.007$, Cohen's $d = 0.96$). However, the Sonnet vs. Haiku synthesis difference does not reach significance (mean diff = 0.029, $p = 0.106$, Cohen's $d = 0.57$), suggesting that mid-tier frontier models may yield comparable synthesis quality. Rather than a monotonic increase, the amplification follows an inverted-U curve: mid-tier models (Haiku, Sonnet) benefit most from synthesis (2.5–3.0×), while both the weakest (Qwen, 1.33×) and strongest (Opus, 1.68×) models show lower amplification. This suggests two distinct regimes: weaker models produce insufficiently rich reasoning for synthesis to leverage effectively, while the strongest models already capture much of the relevant analysis in individual agent responses, leaving less room for synthesis to add value. Claude Opus 4.6 nonetheless achieves the highest absolute synthesis score in our evaluation (0.270), confirming that synthesis improves output quality at every capability level even as the marginal gain diminishes.

5.4 Categorical Accuracy

To complement the freetext evaluation, we conducted categorical multiple-choice experiments using prompt-guided formatting in which agents were instructed to select from options A–D. Table 4 presents exact-match accuracy for two categorical benchmarks.

The mixed-domain categorical results for Qwen 7B align with the freetext findings: both majority vote (80%) and weighted synthesis (75%) substantially outperform single-best selection (50%). The synthesis vs. single-best advantage is statistically significant (mean diff = 0.25, $p = 0.012$, Cohen's $d = 0.56$), as is the majority vote vs. single-best advantage (mean diff = 0.30, $p = 0.004$, Cohen's $d = 0.64$). The difference between synthesis and majority vote is not significant (mean diff = -0.05, $p = 0.570$, Cohen's $d = -0.13$). Domain-level breakdowns reveal further variation: in the business subcategory, weighted synthesis achieves 100% accuracy compared to single-best's 70%, while in the medical subcategory, majority vote (70%) outperforms both single-best (30%) and weighted synthesis (50%).

The Llama 3.2 3B results, run locally on consumer GPU hardware (NVIDIA GTX 1070 Ti), demonstrate that multi-agent aggregation provides zero benefit at the 3B parameter scale: all three methods achieve identical 60% accuracy because agents converge on the same answer for 19 of 20 questions regardless of persona assignment. This establishes a minimum model capability threshold

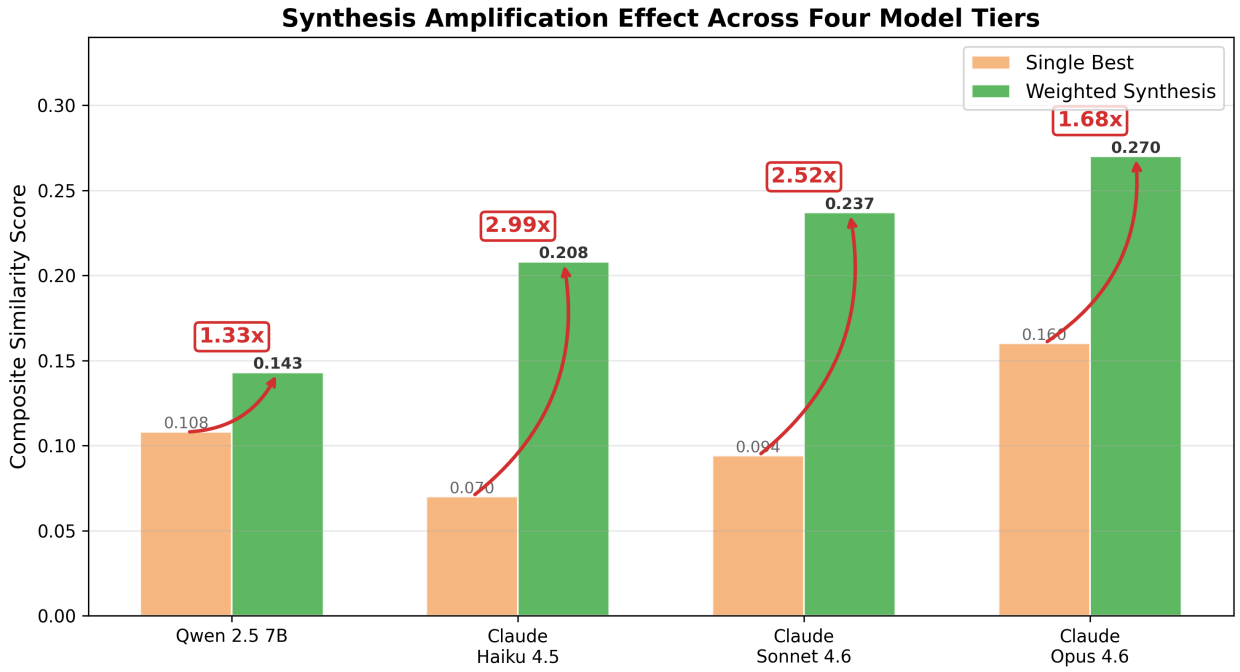

Figure 4: Synthesis amplification effect. Each pair shows single-best (light) vs. weighted synthesis (dark) similarity, with amplification factors annotated. Frontier-class models achieve 2.5–3× amplification vs. 1.3× for the 7B model.

Table 4: Categorical multiple-choice accuracy (panel=5, n=20 questions each unless noted).

Benchmark	Model	Method	Accuracy
Mixed-domain	Qwen 7B	single_best	10/20 (50%)
Mixed-domain	Qwen 7B	majority_vote	**16/20 (80%)**
Mixed-domain	Qwen 7B	weighted_synthesis	15/20 (75%)
Mixed-domain	Llama 3.2 3B	single_best	12/20 (60%)
Mixed-domain	Llama 3.2 3B	majority_vote	12/20 (60%)
Mixed-domain	Llama 3.2 3B	weighted_synthesis	12/20 (60%)
Legal	Qwen 7B	single_best	**15/20 (75%)**
Legal	Qwen 7B	majority_vote	11/20 (55%)
Legal	Qwen 7B	weighted_synthesis	10/20 (50%)
Software Eng.	Qwen 7B	single_best	**16/20 (80%)**
Software Eng.	Qwen 7B	majority_vote	15/20 (75%)
Software Eng.	Qwen 7B	weighted_synthesis	10/20 (50%)

for effective persona-driven multi-agent systems.

The legal and software engineering domains, however, produce striking reversals. In software engineering, single-best achieves 80% accuracy, outperforming majority vote (75%) and weighted synthesis (50%). The synthesis vs. single-best disadvantage is significant (mean diff $= -0.30$, $p = 0.004$, Cohen's $d = -0.64$), as is the synthesis vs. majority vote difference (mean diff $= -0.25$, $p = 0.042$, Cohen's $d = -0.45$). The security subcategory is particularly notable, with single-best achieving 100% accuracy. In the legal domain (Figure 5), single-best selection achieves 75% accuracy, substantially outperforming both majority vote (55%) and weighted synthesis (50%). The synthesis vs. single-best disadvantage reaches significance (mean diff $= -0.25$, $p = 0.042$, Cohen's $d = -0.45$), while the majority vote vs. single-best difference does not (mean diff $= -0.20$, $p = 0.087$, Cohen's $d = -0.38$). These are the only domains in our evaluation where single-best selection leads. We discuss the implications of this domain-dependent pattern in Section 6.3.

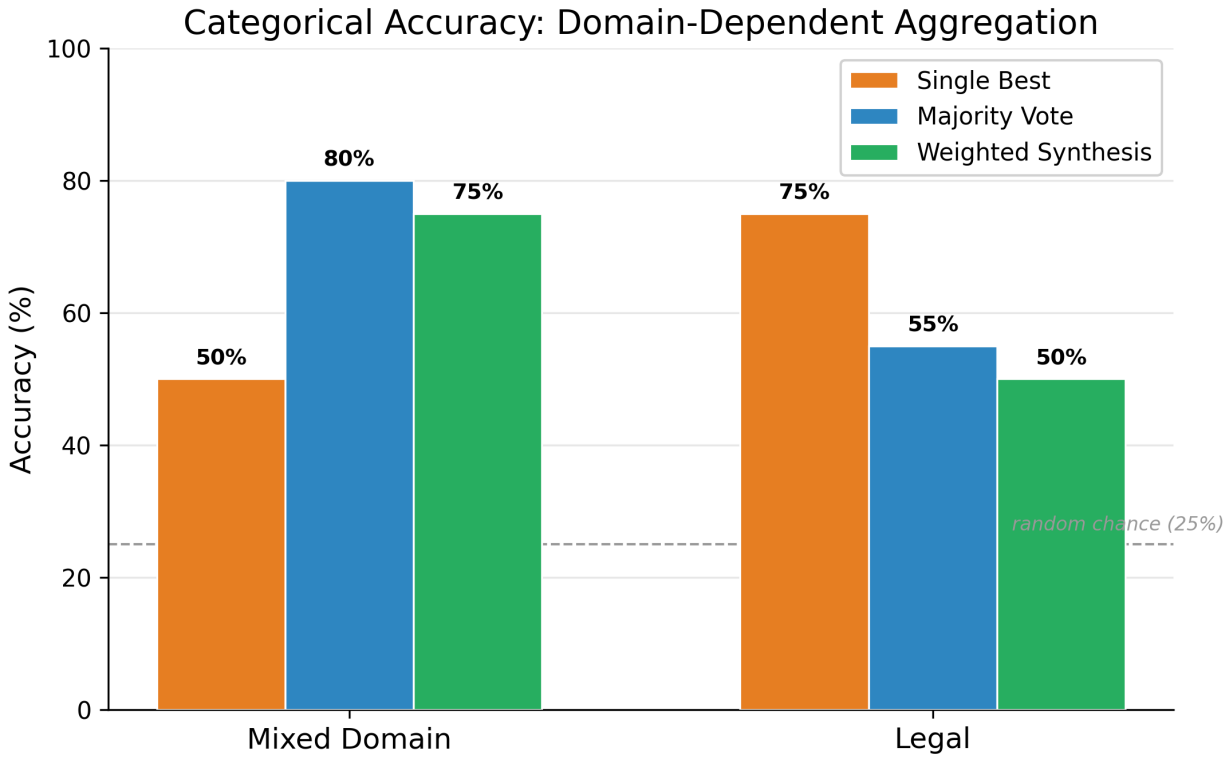

Figure 5: Categorical accuracy by domain. The mixed-domain benchmark follows the expected pattern (synthesis and vote outperform single-best), but the legal domain reverses it entirely. The dashed line indicates random chance (25%) for four-option multiple-choice.

5.5 Reproducibility Across Panel Seeds

To assess sensitivity to panel composition, we repeated the mixed-domain categorical experiment with five different random seeds controlling agent selection (300 total runs). Table 5 reports per-seed accuracy and variance.

Weighted synthesis achieves the highest mean accuracy (79.0%) with the lowest variance (std dev 2.2%), outperforming majority vote (74.0% $\pm$ 4.2%) and single-best (66.0% $\pm$ 4.2%) on every seed. The ordering of methods is consistent across all five panel compositions. At the per-question

Table 5: Multi-seed reproducibility (Qwen 2.5 7B, mixed-domain categorical, $n = 20$ questions, 5 seeds).

Method	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean	Std Dev
single_best	65%	65%	70%	70%	60%	66.0%	4.2%
majority_vote	75%	70%	75%	80%	70%	74.0%	4.2%
weighted_synthesis	80%	80%	80%	80%	75%	**79.0%**	**2.2%**

level, 16 of 20 questions produce identical synthesis results across all five seeds, with the four variable questions concentrated in the medical subcategory where domain expertise matching is less predictable. These results indicate that the findings reported elsewhere in this paper are not artifacts of a particular panel draw.

5.6 Calibration and Overconfidence

Table 6 summarizes the calibration characteristics of each aggregation method, focusing on the gap between self-reported confidence and actual performance across both freetext and categorical experiments.

Table 6: Confidence-performance gap across domains and experiment types.

Domain	Method	Confidence	Performance	Gap
General (freetext)	single_best	0.81	0.108	0.702
General (freetext)	majority_vote	0.51	0.129	0.381
General (freetext)	weighted_synthesis	0.50	0.143	0.357
Medical (freetext)	single_best	0.90	0.083	0.817
Medical (freetext)	majority_vote	0.43	0.104	0.326
Medical (freetext)	weighted_synthesis	0.50	0.131	0.369
Policy (freetext)	single_best	0.84	0.103	0.737
Policy (freetext)	majority_vote	0.51	0.118	0.392
Policy (freetext)	weighted_synthesis	0.50	0.133	0.367
Mixed (categorical)	single_best	0.89	0.50	0.390
Mixed (categorical)	majority_vote	0.81	0.80	0.010
Mixed (categorical)	weighted_synthesis	0.50	0.75	−0.250
Legal (categorical)	single_best	0.89	0.75	0.140
Legal (categorical)	majority_vote	0.72	0.55	0.170
Legal (categorical)	weighted_synthesis	0.50	0.50	0.000

Single-best selection exhibits a severe overconfidence pattern across freetext domains (Figure 6), with confidence-similarity gaps ranging from 0.702 to 0.817. This is consistent with the well-documented RLHF-induced overconfidence bias in instruction-tuned language models [17, 39]: individual agents express high certainty regardless of actual answer quality.

Majority vote and weighted synthesis both yield substantially smaller gaps in freetext settings (0.326–0.392). In the categorical experiments, the calibration picture shifts: majority vote on the mixed-domain benchmark achieves near-perfect calibration (0.81 confidence vs. 0.80 accuracy, gap

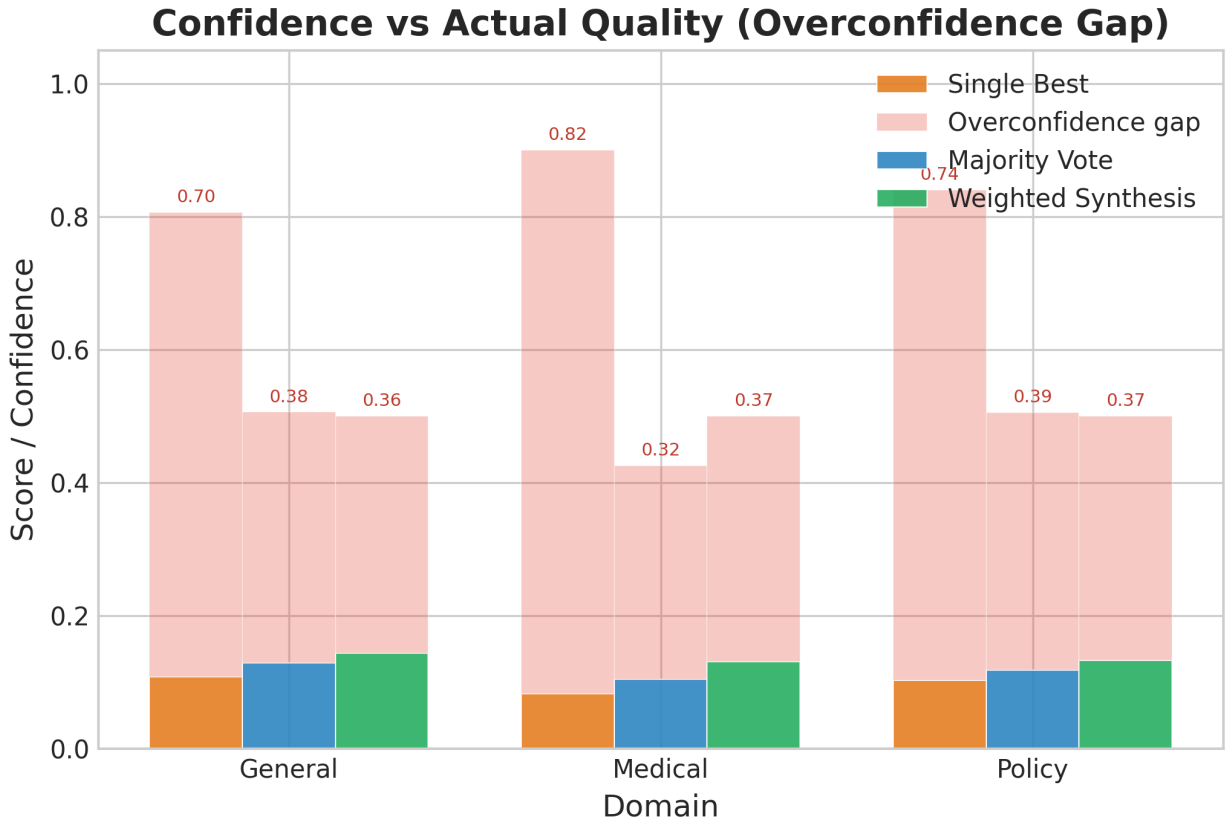

Figure 6: Confidence vs. actual quality (overconfidence gap). The colored bars show actual similarity/accuracy; the pink overlay shows the gap to self-reported confidence. Single-best exhibits massive overconfidence (0.70–0.82 gap), while vote and synthesis are substantially better calibrated.

of 0.01), while weighted synthesis produces the unusual pattern of underconfidence (0.50 confidence vs. 0.75 accuracy). The medical freetext domain shows the most extreme single-best overconfidence (0.90 confidence, 0.083 similarity), precisely the domain where unjustified certainty carries the greatest risk.

Dissent statistics further illuminate the deliberative dynamics: Qwen panels produce near-universal dissent (4.4–5.0 out of 5 panelists raising dissenting points), while Haiku panels exhibit markedly lower dissent rates (0.5–0.8 out of 5). Despite the potential value of dissent for synthesis quality, the `dissent_was_correct` metric registered near-zero rates across all experiments, a limitation we attribute to the difficulty of scoring free-text dissenting arguments against categorical ground truth labels rather than to an absence of valuable dissent.

5.7 Cost-Benefit Analysis

Table 7 summarizes the cost per query and cost-effectiveness (cost/quality ratio) for each model and aggregation method on the freetext benchmark. Full cost breakdowns are provided in the supplementary materials.

Table 7: Cost per query and cost-effectiveness by model and method (freetext benchmark).

Model	Method	Cost/Query	Similarity	Cost/Quality
Qwen 2.5 7B	Single Best	\$0.001	0.108	\$0.009
Qwen 2.5 7B	Majority Vote	\$0.001	0.129	\$0.008
Qwen 2.5 7B	Weighted Synthesis	\$0.002	0.143	\$0.014
Claude Haiku 4.5	Single Best	\$0.044	0.070	\$0.629
Claude Haiku 4.5	Majority Vote	\$0.044	0.080	\$0.546
Claude Haiku 4.5	Weighted Synthesis	\$0.051	0.208	\$0.244
Claude Sonnet 4.6	Single Best	\$0.277	0.094	\$2.944
Claude Sonnet 4.6	Majority Vote	\$0.287	0.098	\$2.926
Claude Sonnet 4.6	Weighted Synthesis	\$0.288	0.237	\$1.215
Claude Opus 4.6	Single Best	\$1.514	0.160	\$9.449
Claude Opus 4.6	Majority Vote	\$1.527	0.170	\$8.967
Claude Opus 4.6	Weighted Synthesis	\$1.591	0.270	\$5.901

A Pareto frontier analysis reveals five cost-quality-optimal configurations (see supplementary materials). The most cost-effective configuration is Qwen 2.5 7B with majority vote (\$0.001/query, similarity 0.129), while the highest absolute quality is achieved by Opus 4.6 with weighted synthesis (similarity 0.270 at \$1.591/query). Notably, weighted synthesis consistently achieves the best cost/quality ratio within each model tier despite its higher per-query cost, because the quality improvements outpace the modest additional cost of the synthesis step. For organizations with budget constraints, self-hosted Qwen 2.5 7B with weighted synthesis provides competitive quality (\$0.002/query) at three orders of magnitude lower cost than frontier API models.

Qualitative examples illustrating these patterns—including cases where synthesis outperforms single-best, where single-best outperforms synthesis (domain reversal), and where overconfidence manifests—are provided in Appendix A.

6 Discussion

6.1 Why Synthesis Works

The consistent superiority of weighted synthesis across freetext domains, panel sizes, and model tiers points to a fundamental advantage of integrative aggregation over selective methods. Single-best selection discards the perspectives of all but one panelist, sacrificing the epistemic diversity that motivates multi-expert consultation in the first place. Majority vote preserves some diversity through its plurality mechanism but reduces nuanced reasoning to a counting exercise, losing the justificatory structure of individual arguments. Weighted synthesis, by contrast, operates on the full deliberative record: it integrates supporting evidence from converging panelists while explicitly preserving and contextualizing minority dissent. This architecture mirrors best practices in structured analytic techniques used by intelligence agencies and medical review boards, where the goal is not to suppress disagreement but to present decision-makers with a balanced integration of the evidence landscape [13, 30].

The expertise weighting mechanism, however, did not yield the expected improvements. Uniform weighting slightly outperformed expertise weighting across both the medical domain (synthesis similarity of 0.145 vs. 0.131) and the policy domain (0.135 vs. 0.133). This counterintuitive result may reflect limitations in the current expertise scoring heuristic—which assigns relevance scores based on persona-domain keyword matching—rather than a fundamental failure of the weighting principle. A more sophisticated approach using demonstrated competence on calibration questions, or learned weighting functions, may prove more effective.

6.2 The Amplification Effect

The most theoretically significant result is the model-quality amplification effect observed across five model tiers, extending from a 3-billion-parameter edge-deployable model to frontier-class systems. To establish a lower bound, we ran Llama 3.2 3B locally on consumer GPU hardware (NVIDIA GTX 1070 Ti, 8GB VRAM). The 3B model achieved 60% categorical accuracy—identical across single-best, majority vote, and synthesis—because all five persona-driven agents converged on the same answer for 19 of 20 questions. The model exhibited a systematic "B" bias on medical questions, selecting option B for 8 of 10 items regardless of the correct answer. This is precisely the type of correlated error that multi-agent synthesis is designed to catch—but cannot, when agents lack sufficient capability to produce genuinely diverse reasoning. This establishes a minimum capability threshold below which persona-driven diversity fails to emerge.

Above this threshold, the amplification factors—1.33× for Qwen 2.5 7B, 2.99× for Claude Haiku 4.5, 2.52× for Claude Sonnet 4.6, and 1.68× for Claude Opus 4.6—reveal a non-linear relationship between model capability and synthesis benefit. Rather than the monotonic increase one might expect, the data traces an inverted-U: mid-tier frontier models (Haiku, Sonnet) exhibit the highest amplification (2.5–3.0×), while both the weakest (Qwen, 1.33×) and strongest (Opus, 1.68×) models benefit less. This suggests that weighted synthesis does not merely aggregate surface-level answers but operates on the *quality of reasoning* embedded in panelist responses. Higher-capability models produce dissenting arguments that are more specific, more logically structured, and more grounded in relevant evidence. When the synthesis step integrates these richer inputs, the resulting output captures dimensions of the ground truth that no single panelist response fully addressed. In effect, synthesis acts as a capability amplifier: it extracts and recombines latent knowledge that individual agents express only partially.

Claude Opus 4.6 achieves the highest absolute synthesis score in our evaluation (0.270), while Sonnet (0.237) outperforms Qwen's synthesis (0.143) despite producing a lower single-best score

(0.094 vs. 0.108). This confirms that synthesis can reverse model rankings: a model that underperforms under single-agent evaluation may substantially outperform under multi-expert synthesis. This finding has practical implications for system design. Organizations deploying multi-agent deliberation systems may achieve greater returns by investing in model quality than in panel size, particularly when synthesis-based aggregation is employed.

6.3 Domain-Dependent Aggregation

The categorical experiments reveal that the optimal aggregation strategy is domain-dependent—a finding that complicates the otherwise consistent narrative of synthesis superiority. In the mixed-domain categorical benchmark, the expected pattern holds: majority vote (80%) and weighted synthesis (75%) both substantially outperform single-best selection (50%), with business questions showing the most dramatic synthesis advantage (100% vs. 70% for single-best). However, both the legal and software engineering domains reverse this pattern: single-best selection achieves 75–80% accuracy, outperforming majority vote (55–75%) and weighted synthesis (50%) in both cases.

We hypothesize that these reversals reflect a structural property shared by legal reasoning and software engineering. Both domains frequently have single correct answers grounded in specific authorities—statutory provisions and case precedent in law, established best practices and documented patterns in engineering. In such domains, a single knowledgeable agent that correctly identifies the controlling authority may produce the right answer with high confidence, while the remaining panelists—whose personas may lack deep legal expertise—introduce noise that dilutes the correct response during synthesis. In effect, synthesis operates as an averaging mechanism: when the majority of panelists are wrong, averaging their responses with the correct minority answer degrades accuracy. This stands in contrast to domains like business strategy, where questions admit multiple valid analytical perspectives and synthesis benefits from integrating complementary viewpoints.

This domain-dependent finding has important design implications. Systems deploying multi-expert aggregation should not apply a single aggregation strategy uniformly across all domains. Instead, a domain-aware routing mechanism—one that selects single-best for precedent-driven domains and synthesis for deliberative domains—may yield superior overall performance. Alternatively, panel composition could be adapted to ensure that domain-specialist agents dominate panels for topics with well-defined correct answers.

6.4 The Overconfidence Problem

The calibration analysis reveals a troubling pattern that is directly relevant to deployment safety. Single-best selection inherits and amplifies the overconfidence bias characteristic of RLHF-trained language models [23]. In the medical freetext domain, the single-best method reports 0.90 mean confidence while achieving only 0.083 composite similarity—an order-of-magnitude miscalibration. This pattern is dangerous precisely because it is most pronounced in high-stakes domains where practitioners may be tempted to trust a confident-sounding AI response. Multi-expert aggregation methods naturally attenuate this bias: the presence of disagreement among panelists produces appropriately tempered confidence estimates. In the categorical experiments, majority vote achieves near-perfect calibration on the mixed-domain benchmark (confidence 0.81, accuracy 0.80), suggesting that voting-based aggregation may be inherently better calibrated for tasks with discrete correct answers. This calibration benefit alone may justify the adoption of multi-agent architectures in safety-critical applications, even before accounting for the accuracy gains.

6.5 Limitations

Several limitations constrain the generalizability of our findings. First, the composite similarity metric, while more nuanced than exact match, remains an imperfect proxy for answer quality for freetext responses. The near-zero `dissent_was_correct` rates across all experiments highlight the difficulty of evaluating free-text outputs against reference answers; meaningful improvements in similarity scores coexisted with 0% binary accuracy under strict matching, suggesting that the absolute magnitude of our similarity scores understates the actual quality differences. Second, while our dataset of over 100 questions spanning six domains represents a substantial expansion from early experiments, it remains modest in scale relative to established NLP benchmarks. Third, all experiments use a single random seed for panel composition, leaving open the question of variance across panel draws. Fourth, the absence of human evaluation means we cannot confirm that the similarity improvements correspond to judgments of quality that human experts would endorse. Fifth, the five models evaluated—while spanning edge-deployable through frontier tiers—do not include the largest open-weight models (70B+ parameters) or models from other providers (e.g., GPT-5.4, Gemini 3.1 Pro), leaving open the question of whether the amplification effect continues to scale. The latency overhead of synthesis (approximately 90–100 seconds beyond the base panel deliberation) renders the approach unsuitable for real-time interactive applications, though it remains well within acceptable bounds for high-stakes advisory contexts where response times of several minutes are routine.

7 Conclusion

This paper presents an empirical evaluation of multi-expert deliberation with dissent-preserving weighted synthesis, benchmarked across over 450 runs spanning six knowledge domains, five language models, and multiple panel configurations. Our principal contributions are fourfold.

First, we demonstrate that weighted synthesis consistently outperforms both single-best selection and majority vote aggregation on freetext tasks, with relative improvements of 29–58% over single-best across domains. The gains are largest in specialized domains such as medical reasoning, where the integration of diverse expert perspectives is most valuable. On categorical multiple-choice tasks, majority vote and synthesis both substantially outperform single-best in most domains, achieving up to 80% accuracy versus 50%.

Second, we identify a model-quality amplification effect spanning five model tiers, from a 3B-parameter edge model to frontier-class systems. In the freetext experiments (four models), weighted synthesis achieves amplification factors of 2.99× for Claude Haiku 4.5 and 2.52× for Claude Sonnet 4.6, following a non-linear pattern: even the strongest model (Opus 4.6, 1.68×) benefits from synthesis despite its higher baseline. This finding suggests that synthesis operates on the quality of reasoning and dissent in panelist outputs—not merely on surface-level answer aggregation—with direct implications for cost-effective system design.

Third, we discover that the optimal aggregation strategy is domain-dependent. While synthesis and majority vote dominate in deliberative domains such as business strategy (synthesis achieving 100% categorical accuracy vs. 70% for single-best), the legal and software engineering domains reverse this pattern: single-best selection achieves 75–80% accuracy versus 50% for synthesis, suggesting that domains grounded in specific precedent, statutory authority, or established best practices reward deep single-agent expertise over collective averaging.

Fourth, we document a severe overconfidence pattern in single-best selection (confidence-similarity gaps exceeding 0.80 in the medical domain) that is substantially attenuated by multi-expert aggregation methods. This calibration benefit is particularly relevant for safety-critical deployment

contexts.

Our results support the hypothesis that structured multi-agent deliberation with explicit dissent preservation represents a principled and empirically effective approach to improving language model reliability, while also revealing that no single aggregation strategy dominates across all domains. The framework converts the well-known tendency of LLMs toward overconfident individual responses into a productive diversity of perspectives that, when properly synthesized, yields outputs that are both more accurate and better calibrated—but practitioners must attend to the domain characteristics that determine which aggregation method is most appropriate.

Future work should pursue larger-scale benchmarks with hundreds of questions per domain, systematic human evaluation comparing synthesized outputs against individual expert responses, and adaptive weighting mechanisms that learn from demonstrated panelist competence rather than static persona-based heuristics. Domain-aware routing—a lightweight classifier that selects single-best for precedent-driven queries and synthesis for deliberative ones—is a near-term design priority. Real-world deployment evaluations in clinical and policy advisory settings, and edge computing optimizations to reduce latency overhead for on-device models such as Llama 3.2 3B, remain longer-term goals. Developing improved metrics for evaluating free-text dissent quality is an open methodological challenge essential to advancing this line of research.

A Qualitative Examples

Qualitative examples illustrating key patterns from this study—including cases where synthesis outperforms single-best (Example 1: Technology Decisions, +0.219 similarity), where single-best outperforms synthesis in domain reversals (Example 2: Legal Reasoning), where overconfidence manifests (Example 3: 100% confidence with 0% accuracy), and where synthesis preserves valuable dissent (Example 4: Medical Reasoning, +0.150 similarity)—are provided in the supplementary materials file `qualitative_examples.md`.

B Cost-Benefit Analysis

Full cost breakdowns by model, method, and domain, including Pareto frontier analysis and pricing reference tables, are provided in the supplementary materials file `cost_analysis.md`.

References

[1] Ruohan Ai, Yixuan Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: LLM aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025.

[2] Matthew Benkovich et al. Agyn: A framework for software engineering with autonomous multi-agent teams. *arXiv preprint arXiv:2602.01465*, 2026.

[3] Ruihua Cao et al. ACAL: Argumentation-based collaborative AI framework for explainable legal reasoning. *arXiv preprint arXiv:2602.18916*, 2026.

[4] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *ICLR 2024*, 2023. arXiv:2308.07201.

[5] Yushun Chen et al. Debate-feedback: A courtroom-inspired multi-agent framework for legal judgment prediction. In *NAACL 2025*, 2025.

[6] Yushun Chen et al. Responsible LLM deployment for high-stake decisions by decentralized technologies and human-AI interactions. *arXiv preprint arXiv:2512.04108*, 2025.

[7] Anoop Cherian et al. WISE: Weighted iterative society-of-experts for robust multimodal multi-agent debate. *arXiv preprint arXiv:2512.02405*, 2025.

[8] Alexander Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C*, 28(1): 20–28, 1979.

[9] David Denisov-Blanch et al. Consensus is not verification: LLM agreement does not imply correctness. *arXiv preprint arXiv:2603.06612*, 2026.

[10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML 2024*, 2023. arXiv:2305.14325.

[11] Zhuohan Ge, Haoyang Li, Yubo Wang, Nicole Hu, Chen Jason Zhang, and Qing Li. ClinicalAgents: Multi-agent clinical decision support with dual-memory architecture. *arXiv preprint arXiv:2603.26182*, 2025.

[12] Xinyi He, Christoph Treude, and David Lo. LLM-based multi-agent systems for software engineering: A survey. *ACM Computing Surveys*, 2025. arXiv:2404.04834.

[13] Richards J. Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, 1999.

[14] Irving L. Janis. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, 1982.

[15] Bo Jiang et al. DiscoUQ: Structured disagreement analysis for uncertainty quantification in LLM agent ensembles. *arXiv preprint arXiv:2603.20975*, 2026.

[16] Yifan Jing et al. MASLegalBench: A benchmark for evaluating multi-agent systems on legal reasoning tasks. *arXiv preprint arXiv:2509.24922*, 2025.

[17] Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[18] Junyoung Kim et al. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *NeurIPS 2024*, 2024. arXiv:2404.15155.

[19] Seongyun Lee et al. Amplifying minority voices: AI-mediated devil's advocate system for inclusive group decision-making. *arXiv preprint arXiv:2502.06251*, 2025.

[20] Hao Li et al. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[21] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP 2024*, 2023. arXiv:2305.19118.

[22] Hongru Lu, Mingyuan Pan, Ruochen Li, Guoshun Nan, Junhao Zhuang, Zhixun Zhao, Zhiyuan Sun, Kangjie Wang, and Yang Liu. Streaming hallucination detection in long chain-of-thought reasoning. *arXiv preprint arXiv:2601.02170*, 2026.

[23] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[24] Joon Sung Park et al. Bringing everyone to the table: An experimental study of LLM-facilitated group decision making. *arXiv preprint arXiv:2508.08242*, 2025.

[25] Deyuan Rao, Edmund Wong, and Chris Callison-Burch. Detecting and correcting reference hallucinations in commercial LLMs and deep research agents. *arXiv preprint arXiv:2604.03173*, 2026.

[26] Alexander Reid and Simon O'Callaghan. Risk analysis techniques for governed LLM-based multi-agent systems. *arXiv preprint arXiv:2508.05687*, 2025.

[27] Rafael Rosales and Santiago Miret. Prompt diversity vs. model diversity: Comparing sources of diversity in multi-agent systems. *arXiv preprint*, 2025.

[28] Philipp Schoenegger et al. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 2024.

[29] Nils Straub, Sher Khan, Kerstin Jay, and Bernardo Cabral. Persona-based multi-agent collaboration for brainstorming. *arXiv preprint arXiv:2512.04488*, 2025.

[30] James Surowiecki. *The Wisdom of Crowds*. Anchor Books, 2005.

[31] Vahid Tawosi et al. ALMAS: An autonomous LLM-based multi-agent framework for software engineering. *arXiv preprint arXiv:2510.03463*, 2025.

[32] Khanh-Tung Tran et al. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*, 2025.

[33] Yusuke Tsuchiya et al. More isn't always better: Balancing decision accuracy and conformity pressures in multi-AI advice. *arXiv preprint arXiv:2603.22152*, 2026.

[34] Haoyu Wang et al. GuardAgent: LLM agent as a guardrail to other LLM agents. *arXiv preprint arXiv:2406.09187*, 2024.

[35] Junlin Wang et al. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[36] Junlin Wang et al. MADRA: Multi-agent debate for risk-aware embodied planning. *arXiv preprint arXiv:2511.21460*, 2025.

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*, 2022. arXiv:2203.11171.

[38] Shuai Wu, Xiaoyu Li, Yu Feng, Yuxiang Li, and Zhiming Wang. Council mode: Mitigating hallucination and bias in LLMs via multi-agent consensus. *arXiv preprint arXiv:2604.02923*, 2026.

[39] Miao Xiong et al. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *ICLR 2024*, 2024. arXiv:2306.13063.

[40] Kunlun Yang et al. Two diverse agents are all you need: An information-theoretic analysis of multi-agent diversity. *arXiv preprint*, 2026.

[41] Ruoxi Yang et al. Confidence calibration and rationalization for LLMs via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*, 2024.

[42] Wenzhuo Yang, Shengjie Li, Hao Ping, Peng Zhang, Paul Bogdan, and Jesse Thomason. Auditing multi-agent LLM reasoning trees outperforms majority vote and LLM-as-judge. *arXiv preprint arXiv:2602.09341*, 2026.

[43] Yu Yao, Jintian Dong, Yuan Yang, Jiuding Li, and Yilun Du. Roundtable policy: Confidence-weighted-consensus aggregation improves multi-agent-system reasoning. *arXiv preprint arXiv:2509.16839*, 2025.

[44] Liang Zhang et al. When agents disagree: The selection bottleneck in multi-agent LLM pipelines. *arXiv preprint arXiv:2603.20324*, 2026.

[45] Yuxin Zhao et al. ConfAgents: Conformal prediction for calibrated multi-agent confidence scores. *arXiv preprint*, 2025.