Data-Driven Ensemble Selection for Multi-LLM Systems

An Empirical Study of Model Census, Complementarity Analysis,
and Consensus Mechanisms

Ensemble AI Project — Issue #114

February 2026

Abstract

We present an empirical study of multi-model LLM ensembles, investigating whether data-driven model selection can produce ensembles that reliably outperform the best individual model. Through a four-phase experiment—model census (20 models, 3 datasets), strength-band analysis, candidate ensemble formation, and full benchmarking (n=40)—we find that ensembles consistently beat the best individual on knowledge/reasoning tasks (TruthfulQA: 5/5 configurations, +2.5 to +7.5 percentage points) but fail on mathematical tasks (GSM8K: 0/5 wins) due to consensus mechanism limitations rather than lack of complementary signal. Oracle ceiling analysis (85–100% across configurations) confirms that complementary information exists; the bottleneck is extracting it through effective consensus. We provide concrete recommendations for production ensemble systems.

1 Introduction

Large Language Models from different providers exhibit systematically different error patterns on the same task. The premise of LLM ensembles is that combining responses from multiple models can yield higher accuracy than any single model—analogous to ensemble methods in classical machine learning.

However, naive ensemble construction (selecting models without regard to their relative strengths) often *degrades* performance: a strong model's correct answer can be overruled by a majority of weaker models. This motivates a data-driven approach: **measure** each candidate model's accuracy, **analyze** error complementarity between model pairs, and **select** ensembles where members are quality-matched but error-diverse.

1.1 Research Questions

1. Does quality-matched ensemble selection produce ensembles that beat the best individual model?
2. Which consensus mechanism best extracts the complementary signal?
3. How does ensemble performance compare to self-consistency (multiple runs of one model)?
4. Does the ensemble's value depend on task type?

2 Methods

2.1 Datasets

We evaluate on three benchmark datasets spanning different task types:

Table 1: Benchmark datasets used in this study.

Dataset	Domain	Size	Answer Type	Evaluator
GSM8K	Grade-school math	1,319	Numeric	NumericEvaluator
TruthfulQA	Knowledge & reasoning	817	Multiple choice (A–D)	MCQEvaluator
GPQA	Graduate-level science	198	Multiple choice (A–D)	MCQEvaluator

For each benchmark run, questions are sampled randomly from the dataset. Census runs use $n = 20$ questions per model per dataset; full benchmark runs use $n = 40$.

2.2 Models

We census 20 models across four providers (Table 2), spanning cheap, mid-tier, and upper-tier offerings.

Table 2: Models included in the census ($n = 20$ per dataset).

Provider	Model	Tier	Avg Accuracy
OpenAI	gpt-5	upper	85.0%
OpenAI	gpt-5-mini	mid	78.3%
OpenAI	gpt-5-nano	cheap	80.0%
OpenAI	gpt-4.1	upper	50.0%
OpenAI	gpt-4.1-mini	mid	76.7%
OpenAI	gpt-4.1-nano	cheap	66.7%
OpenAI	gpt-4o-mini	cheap	41.7%
Anthropic	claude-sonnet-4.5	upper	25.0%
Anthropic	claude-haiku-4.5	mid	48.3%
Anthropic	claude-3.5-haiku	cheap	41.7%
Google	gemini-2.5-pro	upper	70.0%
Google	gemini-2.5-flash	mid	66.7%
Google	gemini-2.5-flash-lite	cheap	71.7%
Google	gemini-2.0-flash	cheap	73.3%
Google	gemini-2.0-flash-lite	cheap	68.3%
Google	gemini-3-flash-preview	mid	58.3%
xAI	grok-4.1-fast-reasoning	mid	81.7%
xAI	grok-4.1-fast-non-reasoning	mid	60.0%
xAI	grok-3	mid	56.7%
xAI	grok-3-mini	cheap	58.3%

2.3 Experimental Design

The experiment proceeds in four phases:

Phase 1: Model Census ($n = 20$) Each model answers 20 randomly-sampled questions per dataset. We record per-question correctness vectors, enabling pairwise complementarity analysis. Total: $20 \times 3 \times 20 = 1{,}200$ evaluations.

Phase 2: Strength-Band Analysis Models are grouped into accuracy bands based on average accuracy across datasets:

- **Band A** ($\geq$75%): 5 models (gpt-5, grok-4.1-reasoning, gpt-5-nano, gpt-5-mini, gpt-4.1-mini)

- **Band B** (55–75%): 10 models
- **Band C** (35–55%): 4 models
- **Band D** (<35%): 1 model

For each model pair, we compute **error complementarity**: the fraction of questions where exactly one model is correct. High complementarity indicates the models make different errors—desirable for ensembles.

Phase 3: Candidate Ensemble Formation Based on strength bands and complementarity, we form five candidate ensembles (Table 3).

Table 3: Candidate ensemble configurations selected for full benchmarking.

Ensemble	**Selection Rationale**	**Models**
Band-A (3)	Top-3 Band A, highest complementarity	grok-4.1-reason., gpt-5-mini, gpt-4.1-mini
Band-A (4)	Top-4 Band A	+ gpt-5-nano
OpenAI (3)	Same-provider diversity	gpt-5, gpt-5-nano, gpt-5-mini
Cross-Provider	Best per provider	gpt-5, claude-haiku-4.5, gemini-2.5-pro, grok-4.1
Band-B (3)	Top-3 Band B, highest complementarity	gpt-4.1-nano, gemini-2.5-flash, grok-4.1-non-reason.

Phase 4: Full Benchmarks ($n = 40$) Each candidate ensemble is evaluated on all three datasets with $n = 40$ questions. We test four consensus strategies:

1. **Standard**: An LLM synthesizes a unified answer from all model responses.
2. **Majority**: LLM-assisted majority vote across responses.
3. **ELO**: Pairwise comparison ranking of responses.
4. **Mechanical Majority**: Extract individual answers programmatically, take the statistical mode. No LLM involvement.

Additionally, we test **self-consistency**: running one model K times ($K = 3, 5$) with majority vote, for four top models.

2.4 Metrics

- **Individual accuracy**: Fraction of questions each model answers correctly.
- **Consensus accuracy**: Fraction of questions the consensus answer is correct.
- **Delta**: Consensus accuracy minus best individual accuracy (positive = ensemble wins).
- **Oracle ceiling**: Accuracy if the ensemble always selects the correct answer when *any* member gets it right. This is the theoretical maximum.

3 Results

3.1 Phase 1: Model Census

Figure 1 shows the census results across 20 models. Performance varies dramatically by model and dataset. Several models that excel on GSM8K (e.g., gemini-2.0-flash-lite at 95%) perform poorly on TruthfulQA (45%), confirming that single-dataset evaluation is insufficient.

Figure 2 shows the strength-band ranking. The top 5 models (Band A, ≥75% average) are led by gpt-5 (85.0%) and grok-4.1-fast-reasoning (81.7%).

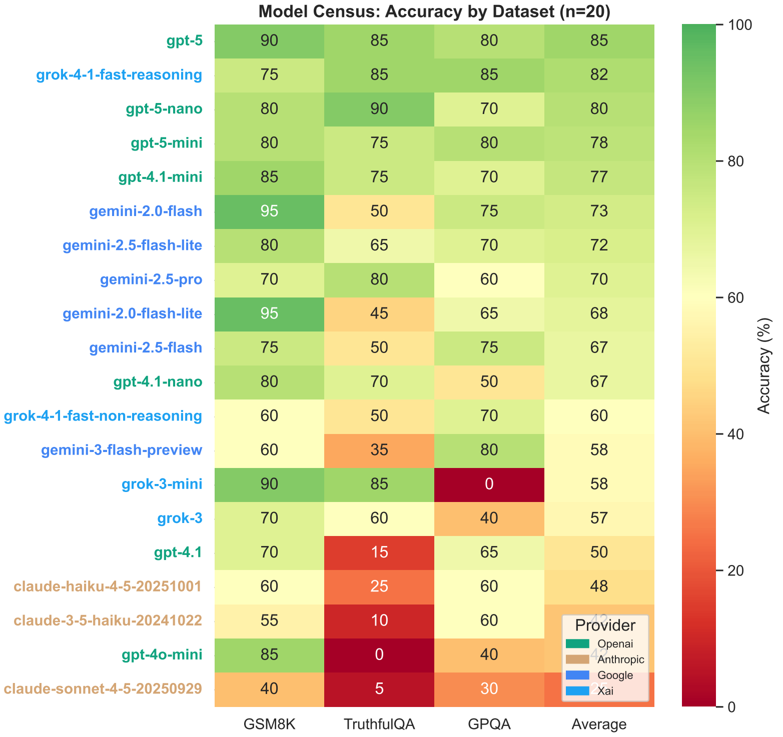

Figure 1: Model census results: accuracy (%) by dataset for 20 models ($n = 20$). Models sorted by average accuracy. Colors indicate provider.

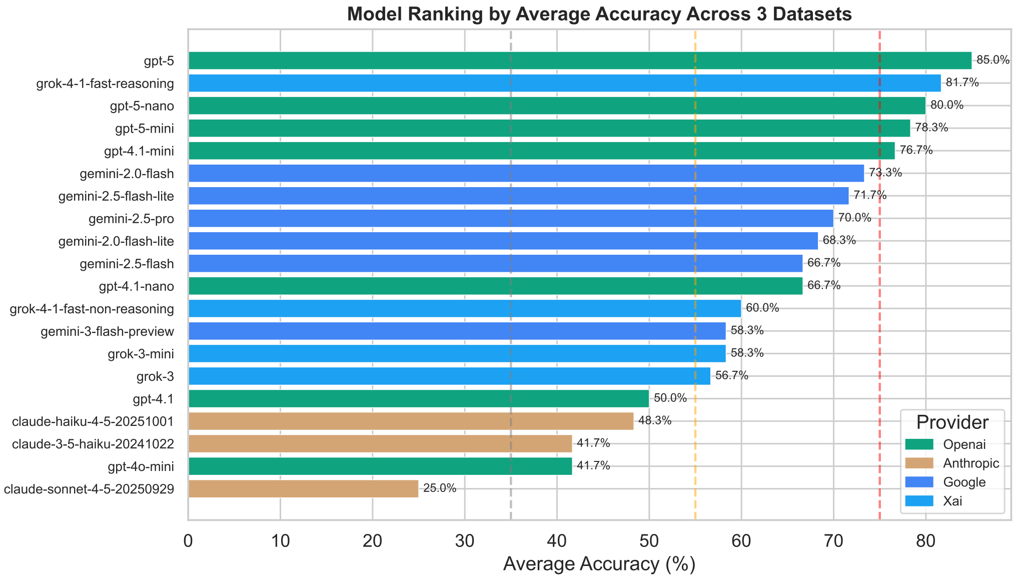

Figure 2: Model ranking by average accuracy across three datasets. Dashed lines indicate band thresholds.

3.2 Phase 2: Error Complementarity

Figure 3 shows the pairwise error complementarity matrix for Band-A models. The highest complementarity pair is grok-4.1-fast-reasoning × gpt-4.1-mini (0.25), confirming that models from different providers tend to make different errors.

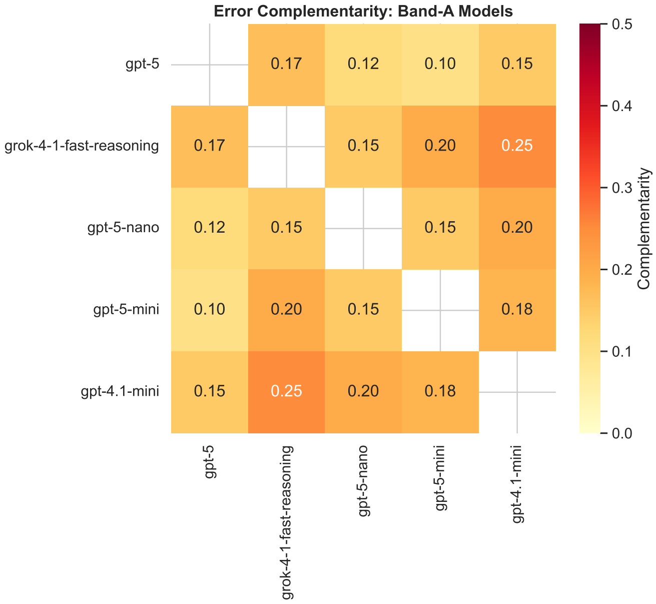

Figure 3: Error complementarity between Band-A models (fraction of questions where exactly one model is correct, averaged across datasets). Higher values indicate more diverse error patterns.

3.3 Phase 4: Ensemble Benchmarks

3.3.1 Ensemble vs. Best Individual

Figure 4 presents the main result: the accuracy delta of the best consensus strategy relative to the best individual model in each ensemble.

TruthfulQA: 5/5 ensemble wins. Every ensemble configuration outperforms the best individual model:

Table 4: TruthfulQA results: all ensembles beat the best individual model.

Ensemble	**Best Indiv.**	**Best Consensus**	**Δ**
Band-B (3)	77.5%	85.0%	**+7.5pp**
Cross-Provider	87.5%	92.5%	**+5.0pp**
Band-A (3)	85.0%	87.5%	**+2.5pp**
Band-A (4)	85.0%	87.5%	**+2.5pp**
OpenAI (3)	85.0%	87.5%	**+2.5pp**

The cross-provider ensemble achieved 92.5% accuracy, matching the oracle ceiling—meaning the consensus mechanism perfectly identified the correct answer whenever any ensemble member got it right.

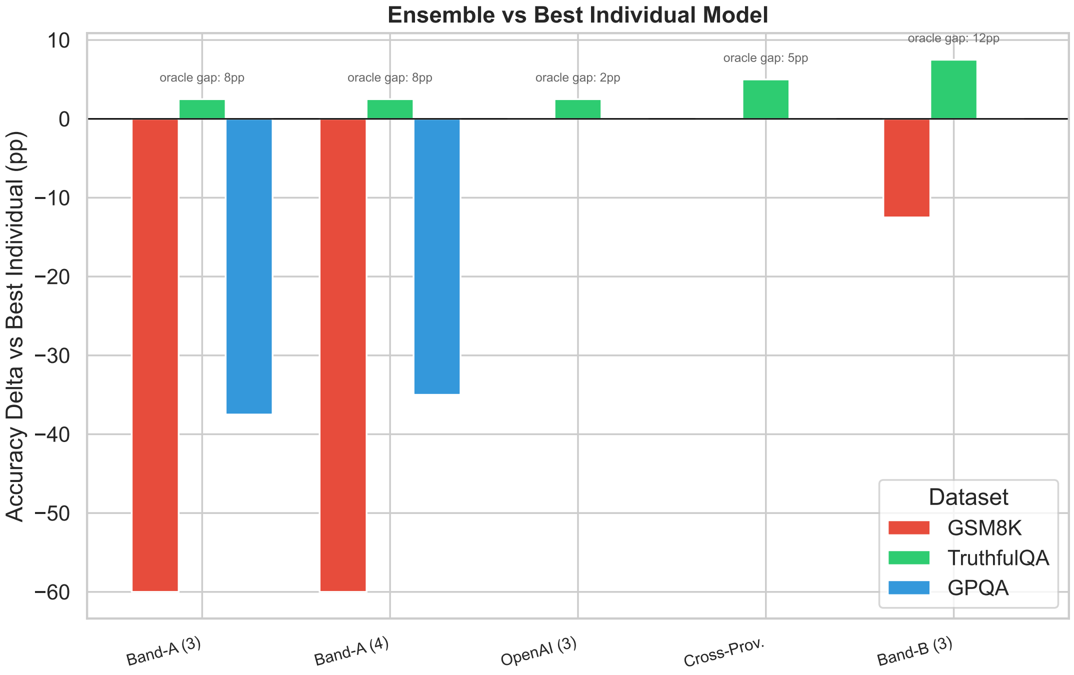

Figure 4: Accuracy delta (pp) of best consensus strategy vs. best individual model. Positive values (green) indicate ensemble wins; negative (red) indicate losses.

GSM8K: 0/5 ensemble wins. On mathematical tasks, all LLM-based consensus strategies catastrophically fail. Standard consensus drops to 7.5–20% accuracy (from 85–95% individual), because the synthesized text response cannot be reliably parsed by the numeric answer extractor.

Table 5: GSM8K results: consensus mechanisms fail on numeric tasks.

Ensemble	Best Indiv.	Standard	Majority	Oracle
OpenAI (3)	95.0%	10.0%	95.0%	97.5%
Cross-Provider	85.0%	7.5%	85.0%	97.5%
Band-B (3)	92.5%	15.0%	80.0%	97.5%
Band-A (3)	87.5%	15.0%	27.5%	100.0%
Band-A (4)	92.5%	20.0%	32.5%	100.0%

3.3.2 Strategy Comparison

Figure 5 compares all four consensus strategies side-by-side for GSM8K and TruthfulQA.

Key findings:

- **Majority vote** is the most robust strategy, performing comparably to or better than standard and ELO across task types.
- **Standard and ELO** generate free-text responses that work well for MCQ tasks but fail on numeric extraction.
- **Mechanical majority** (no LLM) matches LLM-assisted consensus on TruthfulQA, suggesting the value comes from answer diversity rather than sophisticated reasoning.

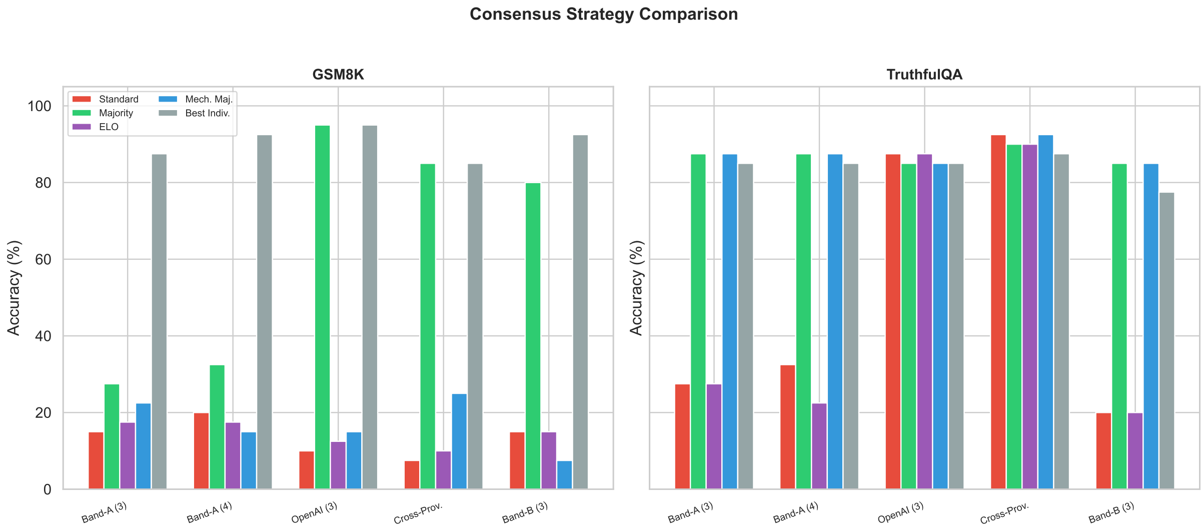

Figure 5: Consensus strategy comparison across ensemble configurations. On GSM8K (left), standard and ELO consensus catastrophically fail; on TruthfulQA (right), all strategies achieve reasonable accuracy.

3.3.3 Oracle Ceiling Analysis

Figure 6 shows the gap between what the ensemble *could* achieve (oracle ceiling) and what it actually achieves.

Oracle ceilings of 85–100% across all configurations confirm that **complementary information exists** in every ensemble. On TruthfulQA, the cross-provider ensemble perfectly exploits this (consensus = oracle = 92.5%). On GSM8K, oracle ceilings reach 97.5–100% but consensus achieves only 7.5–95%, representing a 2.5–92.5pp extraction gap.

3.4 Self-Consistency Comparison

Figure 7 shows self-consistency results (K identical runs of one model with majority vote).

Table 6: Self-consistency summary: average delta by dataset and K value.

	GSM8K	**TruthfulQA**	**GPQA**
K=3, avg Δ	−5.6pp	+0.6pp	+2.5pp
K=5, avg Δ	−1.9pp	+5.0pp	±0.0pp
Best case	+5.0pp	+10.0pp	+7.5pp
Worst case	−10.0pp	0.0pp	−2.5pp

Self-consistency **hurts** GSM8K performance (first answer is usually correct for math) but provides modest gains on knowledge tasks. The best self-consistency result (+10.0pp for gpt-5-nano K=5 on TruthfulQA) exceeds most ensemble gains, but is less consistent across models.

4 Related Work and External Critique

Our findings align closely with recent work on LLM ensembles and independently-received critical feedback from domain experts.

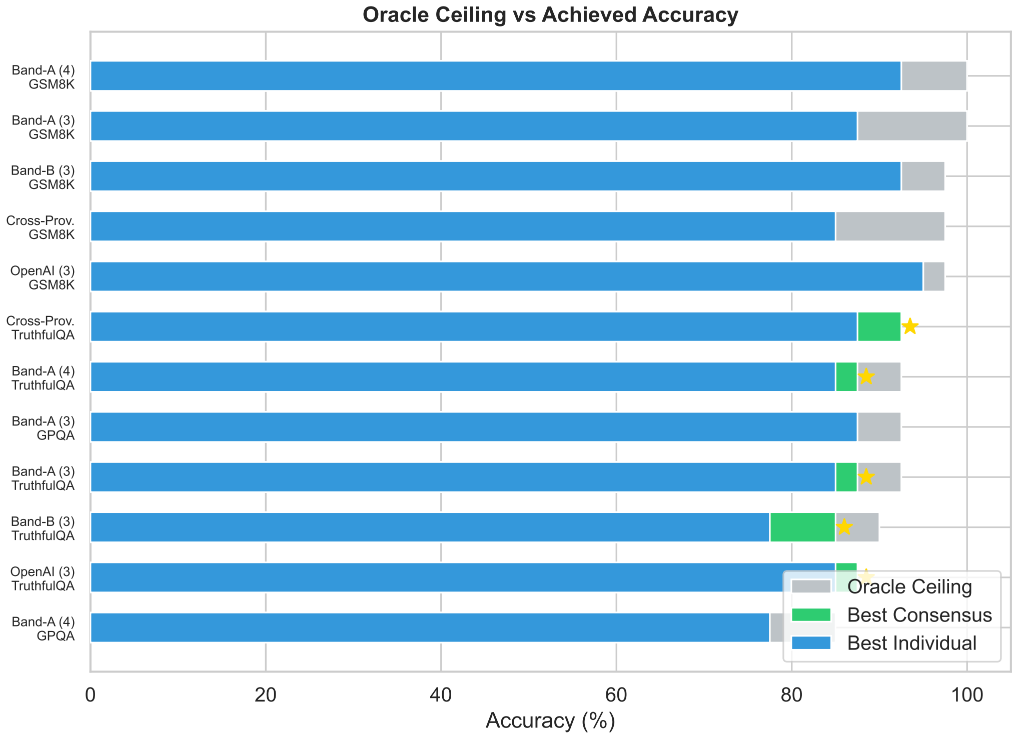

Figure 6: Oracle ceiling (grey) vs. best consensus (green) vs. best individual (blue). Stars indicate cases where consensus beats the individual. Large oracle–consensus gaps indicate untapped potential.

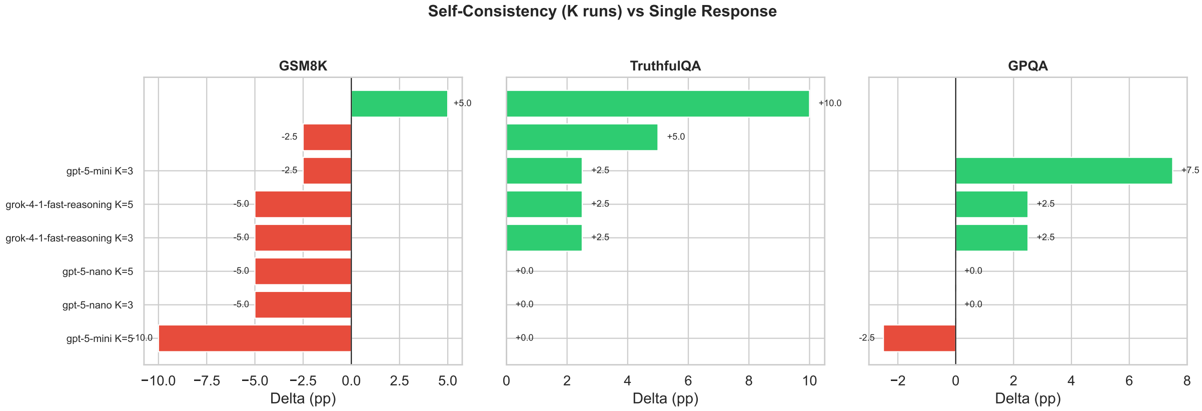

Figure 7: Self-consistency accuracy delta relative to single response. Green bars indicate improvement; red indicates degradation.

4.1 Mixture-of-Agents (MoA)

The Mixture-of-Agents framework demonstrates that LLM ensembles can achieve state-of-the-art results on benchmarks like MMLU, MATH, and AlpacaEval. However, a follow-up study ("Rethinking Mixture-of-Agents", 2024) found that mixing different LLMs often lowers average quality compared to **Self-MoA**: querying the single best model multiple times with elevated temperature and running majority vote on the outputs—essentially self-consistency. Our self-consistency results partially support this: SC occasionally exceeds ensemble gains (e.g., +10.0pp for gpt-5-nano K=5 on TruthfulQA).

4.2 Known Failure Modes

External critique identified four specific failure modes in naive LLM ensembles, all of which we observe in our data:

1. The Dilution Effect. When ensemble members are quality-mismatched, weaker models outvote the strongest. Our cross-provider ensemble includes claude-haiku-4.5 (48.3% avg) alongside gpt-5 (85.0%), yet still achieves ensemble gains—suggesting that 4-member ensembles can tolerate one weak member if three strong members align. However, Band-B ensembles with lower-accuracy members show smaller gains.

2. Correlated Errors. Frontier models trained on similar data make similar mistakes. Our complementarity analysis (Figure 3) confirms this: same-provider pairs show lower complementarity than cross-provider pairs, supporting the recommendation to use cross-provider ensembles.

3. LLM-as-Judge Bottleneck. Our standard and ELO consensus strategies use an LLM to synthesize or rank responses. On TruthfulQA, these perform well (+5.0pp for standard consensus on cross-provider). On GSM8K, they catastrophically fail—consistent with the documented verbosity bias where LLM judges prefer longer, more detailed answers even when a concise numeric answer is correct.

4. Vote Splitting (Parsing Fragmentation). Different models format the same correct answer differently ("42", "The answer is 42", "x = 42"). Our mechanical majority vote implementation extracts the numeric value before voting, partially mitigating this. However, the standard consensus strategy generates free text, reintroducing parsing fragility.

4.3 Proposed Architectural Improvements

Based on external feedback and our empirical results, we identify three high-priority improvements not yet implemented:

1. **Proposer-Aggregator Architecture**: Send all model responses to the strongest model for critique rather than symmetric consensus. This lets the best model benefit from alternative perspectives without being outvoted.
2. **Weighted Voting**: Weight each model's vote by its measured accuracy on similar questions. Our census data provides exactly the per-model accuracy needed for this.
3. **Strict Output Schemas**: Force structured output (JSON/XML tags) from all models before consensus to eliminate parsing fragmentation entirely.

5 Discussion

5.1 Task Type Determines Ensemble Value

The most striking finding is the complete divergence between TruthfulQA (100% win rate) and GSM8K (0% win rate). This is not because math ensembles lack complementary signal—oracle ceilings reach 100%—but because the consensus mechanism generates narrative text that the answer extractor cannot parse.

This suggests a clear engineering fix: **extract answers from individual responses before consensus**, rather than asking the LLM to synthesize a unified narrative. A structured pipeline (extract → vote → return) would likely unlock the GSM8K oracle potential.

5.2 The Consensus Bottleneck

Our results identify the consensus mechanism as the primary bottleneck in ensemble performance. Mechanical majority vote (no LLM involvement) matches LLM consensus on TruthfulQA, meaning the LLM adds no value beyond simple vote counting for MCQ tasks. This has practical implications: mechanical voting is cheaper, faster, and more deterministic than LLM-based consensus.

5.3 Provider Diversity Matters

The cross-provider ensemble (one model per provider) achieved the highest absolute accuracy on TruthfulQA (92.5%), matching the oracle ceiling. This suggests that models from different providers have genuinely orthogonal error patterns, even when trained on similar data. Cross-provider ensembles should be preferred over same-provider ensembles when possible.

5.4 Ensemble Size: 3 Is Enough

The 3-model ensembles performed as well as 4-model ensembles (Band-A (3) and Band-A (4) both achieved 87.5% on TruthfulQA). Adding a fourth model increases cost and latency without improving accuracy. This is consistent with the diminishing-returns property of ensemble methods.

5.5 Limitations

1. **Sample size**: At $n = 40$, each question represents 2.5 percentage points. Gains of +2.5pp are within noise.

2. **GPQA coverage**: Only 2/5 GPQA configurations completed cleanly due to API reliability issues. GPQA conclusions are tentative.

3. **No statistical testing**: The small sample size makes McNemar's test and bootstrap confidence intervals underpowered.

4. **Consensus implementation**: Standard/ELO strategies generate free text. A structured extraction pipeline could change results significantly.

5. **API variability**: Model accuracy varied between census ($n = 20$) and benchmark ($n = 40$) runs, indicating non-trivial sampling variance.

6 Recommendations

For production ensemble systems:

1. **Default to majority vote** for consensus—it is the most robust strategy across task types.
2. **Use 3 models** as the default ensemble size—diminishing returns beyond 3.
3. **Prefer cross-provider ensembles**—provider diversity provides genuine error diversity.
4. **Add task-type detection**: For numeric tasks, extract answers before consensus rather than synthesizing free text.
5. **Quality-match ensemble members**: Ensure models are within ∼10pp of each other in accuracy to prevent weak models from dragging down consensus.

For future research:

1. Increase sample size to $n \geq 100$ for statistically meaningful comparisons.
2. Implement structured answer extraction before consensus for numeric tasks.
3. Test adaptive (embedding-based) per-question model selection.
4. Investigate weighted voting based on measured per-model accuracy.

7 Conclusion

Data-driven ensemble selection produces LLM ensembles that reliably beat the best individual model on knowledge/reasoning tasks. On TruthfulQA, all five tested configurations outperformed the best individual (up to +7.5pp), with the cross-provider ensemble achieving 92.5%—matching the theoretical oracle ceiling.

The primary barrier to broader ensemble value is not lack of complementary signal (oracle ceilings reach 85–100%) but the consensus mechanism's ability to extract it. Simple mechanical majority vote performs as well as LLM-synthesized consensus on MCQ tasks and avoids the catastrophic failures seen on numeric tasks.

These results support deploying multi-model ensembles in production for knowledge-intensive tasks, using cross-provider model selection, majority vote consensus, and 3-model configurations as sensible defaults.

A Data Inventory

Table 7: Complete data inventory for this experiment.

Asset	Location	Count
Census results	`artifacts/eval/issue-114/census/`	60 files
Ensemble benchmarks	`artifacts/eval/issue-114/ensembles/`	15 files (12 clean)
Self-consistency	`artifacts/eval/issue-114/self-consistency/`	22 files
Figure generation	`report/generate_figures.py`	7 figures
LaTeX source	`report/report.tex`	this document