CouncilGPT-2: Multi-Model Deliberation for Structured Text Generation

Stanford CS224N Default Project

Jake Bodea
Stanford University
jpbodea@stanford.edu

Abstract

Single-model language models produce brittle outputs on tasks requiring both semantic understanding and structural adherence. We investigate whether a *council* of independently fine-tuned GPT-2 models, coordinated at inference time, can improve performance on paraphrase detection and sonnet generation. We train three GPT-2 members with different random seeds and evaluate four aggregation strategies for classification (majority vote, confidence-weighted vote, unanimity-with-fallback, and logit-level debate) and three strategies for generation (token-level logit ensemble, best-of-N cross-perplexity selection, and debate-guided generation). On paraphrase detection, council debate achieves 88.98% dev accuracy with halved inter-member disagreement, a modest improvement over the best single model (88.96%). The more striking result emerges in sonnet generation: while surface-level metrics (chrF) show negligible differences across methods, the debate strategy dramatically improves structural quality—achieving 84.1% iambic pentameter adherence versus 55–68% for other methods, with mean syllable count closest to the target of 10 at 9.8 syllables per line. We argue that logit-level debate acts as an implicit regularizer for structured generation, biasing token selection toward consensus choices that respect metrical constraints.

1 Key Information

- **Mentor**: Shicheng Liu
- **External Collaborators**: None
- **Sharing project**: No

2 Introduction

Ensemble methods are a well-established technique for improving the reliability of machine learning systems [1]. In the era of large language models, however, ensemble approaches have been under-explored for autoregressive generation tasks, particularly those requiring adherence to structural constraints such as meter, rhyme, and fixed line counts.

We study this gap through the lens of the CS224N default project, extending the standard GPT-2 pipeline with a *council* inference layer inspired by the multiagent debate framework of Du et al. [2]. Rather than relying on a single model's output, we train $N{=}3$ GPT-2 models with different random seeds and aggregate their predictions at inference time. For classification (paraphrase detection), we implement four aggregation strategies culminating in a logit-level debate round. For generation (sonnet completion), we extend these ideas to the autoregressive setting with three strategies: token-level logit ensembling, best-of-N selection via cross-perplexity, and debate-guided generation.

Our key finding is that **council debate is most valuable for structured generation, not classification**. On paraphrase detection, all council variants produce marginal improvements ($<0.1\%$ accuracy). On

sonnet generation, however, debate dramatically improves structural quality: 84.1% of generated lines fall within the pentameter range (8–12 syllables) compared to 55–68% for other methods, and the average syllable count per line (9.8) is closest to the iambic pentameter target of 10. We attribute this to debate's implicit regularization effect: by blending each model's logits with its peers' preferences, debate biases generation toward tokens that the council collectively agrees on, which tend to be metrically conventional choices.

3 Related Work

Multiagent debate. Du et al. [2] proposed a multiagent debate framework in which multiple LLM instances exchange responses over several rounds, improving factuality and reasoning accuracy. Their work demonstrates that structured disagreement and revision at inference time can extract additional capability from fixed models. We adapt this idea to the logit level for computational efficiency.

Self-consistency decoding. Wang et al. [3] showed that sampling multiple chain-of-thought reasoning paths and aggregating via majority vote (self-consistency) improves reasoning performance. Our best-of-N strategy is analogous, using cross-perplexity rather than majority vote to select among candidates.

Ensemble methods. Classical ensemble methods [1] improve predictions by combining diverse models. Deep ensembles [4] extend this to neural networks, training multiple models with different initializations. Our council follows this paradigm but applies it at inference time to both classification and autoregressive generation.

Neural poetry generation. Lau et al. [5] fine-tuned language models for poetry generation and developed automatic metrics for poetic quality. Hopkins and Kiela [6] proposed methods for automatically generating rhythmic poetry. Our structural metrics (syllable counting, pentameter rate, rhyme detection) build on this line of work using the CMU Pronouncing Dictionary for phoneme-based analysis.

GPT-2. Radford et al. [7] introduced GPT-2, a transformer-based language model that demonstrated strong performance across diverse NLP tasks through unsupervised pretraining. We use GPT-2 small (117M parameters) as our base model.

4 Approach

4.1 Base Models and Council Setup

Baseline: cloze-style GPT-2. We build on the CS224N default project starter code, which provides a GPT-2 (117M parameters) backbone with pretrained weights from HuggingFace. For paraphrase detection, each input pair (s_1, s_2) is formatted as a cloze prompt: `Is ‘‘`s_1`’’ a paraphrase of ‘‘`s_2`’’? Answer ‘‘yes’’ or ‘‘no’’:` , and the model produces two logits via a linear head applied to the last token's hidden state. For sonnet generation, the model is fine-tuned for next-token prediction on Shakespeare's sonnets, and generation uses top-p nucleus sampling with temperature scaling.

Council training. We train $N{=}3$ council members by fine-tuning GPT-2 with identical hyperparameters but different random seeds $\{0, 1, 2\}$. All parameters are updated with AdamW (learning rate 10^{-5}, batch size 8). Paraphrase models are trained for 10 epochs; sonnet models for 5 epochs. The diversity required for effective ensembling arises naturally from the different weight initializations induced by distinct random seeds.

4.2 Council Inference for Paraphrase Detection

Paraphrase aggregation. Let $\boldsymbol{\ell}_i \in \mathbb{R}^2$ denote the logits of member i for a given input. We implement four aggregation strategies:

1. **Majority vote**: $\hat{y} = \text{mode}_i \ \text{argmax} \ \boldsymbol{\ell}_i$.

2. **Confidence-weighted vote**: $\hat{y} = \text{argmax} \ \frac{1}{N} \sum_{i=1}^{N} \text{softmax}(\boldsymbol{\ell}_i)$.
3. **Unanimity-with-fallback**: use the unanimous label if all members agree; otherwise fall back to confidence-weighted vote.
4. **Debate**: revise each member's logits using the peer-average signal before aggregation:

$$\boldsymbol{\ell}'_i = \boldsymbol{\ell}_i + \alpha \cdot \bar{\boldsymbol{\ell}}_{-i}, \quad \text{where } \bar{\boldsymbol{\ell}}_{-i} = \frac{1}{N-1} \sum_{j \neq i} \boldsymbol{\ell}_j \tag{1}$$

with α=0.5, followed by confidence-weighted aggregation of the revised logits.

4.3 Council Inference for Sonnet Generation

Sonnet generation aggregation. We extend the council idea to autoregressive generation with three strategies. At each time step t, let $\boldsymbol{\ell}_i^{(t)} \in \mathbb{R}^{|\mathcal{V}|}$ be the logit vector of member i over the vocabulary $\mathcal{V}$.

Token-level logit ensemble. Average the logits at each step and sample from the blended distribution:

$$x_{t+1} \sim \text{softmax}\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\ell}_i^{(t)} \ / \ \tau \right) \tag{2}$$

where τ=1.2 is the temperature. All members share the same generated prefix.

Best-of-N (cross-perplexity). Each member generates a complete sonnet independently. We select the candidate with the lowest *cross-perplexity*—the average perplexity assigned to it by all council members:

$$\hat{c} = \text{argmin}_i \ \frac{1}{N} \sum_{j=1}^{N} \text{PPL}(\text{candidate}_i, \text{model}_j) \tag{3}$$

Debate. Apply the debate revision (Equation 1) at each autoregressive step: each member's logits are shifted toward the peer mean before sampling. Each member generates independently from its revised distribution, and the best candidate is selected via cross-perplexity. This combines per-step regularization with post-hoc selection.

4.4 Structural Evaluation and Implementation

Structural metrics. To evaluate sonnet quality beyond surface-level text similarity, we implement phoneme-based structural metrics using the CMU Pronouncing Dictionary:

- **Syllable count**: phoneme-based via the `pronouncing` library, with a vowel-group heuristic fallback for out-of-vocabulary words.
- **Pentameter rate**: fraction of lines with 8–12 syllables (the range for reasonable iambic pentameter).
- **Rhyme accuracy**: fraction of correct rhyme pairs in the Shakespearean scheme (ABAB CDCD EFEF GG = 7 pairs per sonnet), detected by matching rhyming parts (phonemes from the last stressed vowel onward).
- **Completion rate**: fraction of sonnets with $\geq$11 lines generated.

Original contributions. The GPT-2 backbone, training loop, and data loading are from the provided starter code (implemented by us per the assignment). The council inference layer (`council.py`), sonnet council generation (`sonnet_council.py`), and structural metrics (`sonnet_metrics.py`) are entirely our original contributions.

5 Experiments

5.1 Data

Paraphrase detection. We use the Quora Question Pairs dataset included in the starter code: 141,506 training pairs, 20,215 development pairs, and 40,431 test pairs. Each example is a pair of questions labeled as paraphrase (1) or non-paraphrase (0).

Sonnet generation. We use 143 Shakespeare sonnets for training and 12 held-out sonnets for evaluation. Each test example provides the first three lines as a prompt; the model must generate the remaining 11 lines. Ground-truth completions from Shakespeare serve as the reference for both surface-level and structural evaluation.

5.2 Evaluation method

For paraphrase detection, we report **accuracy** and **macro F1** on the dev set. We additionally report **disagreement rate** (fraction of examples where council members predict different labels), **ECE** (Expected Calibration Error), and **Brier score** to assess consensus and calibration.

For sonnet generation, we report surface-level metrics (**chrF** [8] and **perplexity**) alongside the structural metrics described in Section 3: syllables per line, syllable deviation from 10, pentameter rate, rhyme accuracy, and completion rate.

5.3 Experimental details

All models use GPT-2 small (12 layers, 768 hidden dimensions, 12 attention heads; 117M parameters). Fine-tuning uses AdamW with learning rate 10^{-5} and batch size 8. Paraphrase models are trained for 10 epochs; sonnet models for 5 epochs. Three council members use seeds $\{0, 1, 2\}$. Generation uses temperature τ=1.2 and top-p=0.9 nucleus sampling. The debate parameter is α=0.5. Training was conducted on an NVIDIA L4 GPU via Google Cloud Platform.

5.4 Results

Paraphrase detection. Table 1 reports dev-set results for all council variants.

Table 1: Paraphrase detection results (dev set). Best in **bold**.

Method	Acc	F1	Disagree	ECE	Brier
Single (seed 0, best)	0.8896	0.8832	—	—	—
Single (seed 1)	0.8767	0.8703	—	—	—
Single (seed 2)	0.8808	0.8747	—	—	—
Council: majority	0.8879	0.8820	0.096	0.018	0.080
Council: conf.-weighted	0.8894	0.8834	0.096	0.017	0.080
Council: unanimity	0.8894	0.8834	0.096	0.017	0.080
Council: debate	**0.8898**	**0.8838**	**0.050**	0.051	0.084

All council variants perform comparably to the best single model. The debate variant achieves the highest accuracy (88.98%) and F1 (88.38%), and notably halves the disagreement rate from 9.6% to 5.0%, indicating stronger consensus. On the official test leaderboard, our final paraphrase submission placed **56th**.

Sonnet generation—surface metrics. Table 2 reports chrF, anchor perplexity, and cross-perplexity.

Table 2: Sonnet generation surface metrics. Anchor-PPL = perplexity under member 0; X-PPL = mean perplexity across all council members.

Method	chrF	Anchor-PPL	X-PPL
Single (seed 0)	40.41	140.76	—
Single (seed 1)	39.88	122.96	—
Single (seed 2)	39.36	120.54	—
Token ensemble	**40.63**	127.90	128.33
Best-of-N	39.90	102.28	102.40
Debate	39.00	**20.24**	**20.13**

ChrF differences are small ($\sim$1.5 points), indicating that council methods do not substantially improve surface-level text similarity to Shakespeare. Token ensemble achieves the best chrF (40.63), slightly

outperforming the strongest single-model baseline (40.41), but the gain is modest. On the official test leaderboard, our final sonnet submission placed **62nd**, consistent with the fact that debate is not the strongest method on the chrF-style surface metric. Debate achieves dramatically lower anchor perplexity and cross-perplexity (20.24 and 20.13 vs. roughly 100–140 for other methods), which we discuss in the analysis.

Sonnet generation—structural metrics. Table 3 reveals the most interesting result.

Table 3: Sonnet structural quality. Syl/ln = syllables per line (target: 10). SylDev = mean deviation from 10. Penta = pentameter rate (8–12 syllables). Compl = completion rate ($\geq$11 lines).

Method	Syl/ln	SylDev	Penta	Rhyme	Compl
Shakespeare (reference)	10.2	0.2	100%	83.3%	100%
Single (seed 0)	12.0	3.1	53.8%	7.1%	75.0%
Single (seed 1)	11.3	2.1	67.6%	10.7%	75.0%
Single (seed 2)	11.0	2.3	68.0%	10.7%	58.3%
Token ensemble	12.1	3.0	55.7%	7.1%	66.7%
Best-of-N	11.7	2.7	56.2%	9.5%	75.0%
Debate	**9.8**	**1.2**	**84.1%**	**14.3%**	**83.3%**

Debate is the clear winner on every structural metric. It achieves 84.1% pentameter rate versus 55–68% for other methods, the lowest syllable deviation (1.2 vs. 2.1–3.1), and the highest completion and rhyme rates. Its mean syllable count of 9.8 is the closest to the iambic pentameter target of 10. Thus, while debate does not perform especially well on the leaderboard's chrF-oriented ranking, it substantially outperforms the other methods on the sonnet-form metrics that measure meter, rhyme, and completion.

6 Analysis

6.1 Why Debate Helps Structured Generation

Why debate helps structure but not surface metrics. The debate revision $\boldsymbol{\ell}'_i = \boldsymbol{\ell}_i + \alpha \cdot \bar{\boldsymbol{\ell}}_{-i}$ acts as a *regularizer* at each generation step: it shifts each model's distribution toward tokens that the council collectively favors. For structured text like sonnets, this consensus tends to select metrically conventional words—common monosyllabic and disyllabic tokens that fit naturally into pentameter lines—over the idiosyncratic, often polysyllabic or archaic choices that individual models sometimes produce.

Token ensembling, by contrast, averages logits into a single distribution that is *more uniform* (higher entropy), which can actually increase syllable variance. Best-of-N selects among independently generated candidates but cannot influence the generation process itself, so structural quality depends entirely on the luck of individual samples. Debate is unique in that it provides per-step guidance while preserving each model's distinct generative trajectory.

The perplexity-structure connection. Debate's much lower anchor perplexity and cross-perplexity (about 20, compared with 100+ for the other methods) mean that the selected continuation is judged likely by all three council members, not just by the model that produced it. This is exactly the behavior we would expect from a consensus-based decoding rule: the council prefers candidates that remain high-probability under multiple independently trained models. For sonnets, those jointly likely candidates also tend to be the more regular ones, since irregular line lengths and degenerate continuations are less likely to receive consistently high probability across the whole council.

6.2 Qualitative Analysis and Failure Modes

Qualitative comparison. Comparing generated sonnets reveals clear differences. A single-model output (seed 0, sonnet 4) produces incoherent, structurally broken text:

> *"Seine spake he, and utter blisses add joy; / Not yet, though wrong know them'd better! / His augury admonish her purple dub tales..."*

The debate variant for the same prompt produces more coherent, metrically regular lines:

> *"Thy name is thy ghost, but thou are not to be seen. / Thy wight is not to be seen, but to be seen, / Thy sound, but thou will not live."*

While neither is high-quality poetry, the debate output maintains more consistent line length and syntactic coherence.

Failure modes. Despite debate's improvements, several failure modes persist: (1) *Repetition*: debate sometimes produces repetitive structures (e.g., sonnet 5 generates "Or where thou dost find what thou dost not see?" four times consecutively), likely because consensus reinforces already-probable continuations. (2) *Rhyme*: even with debate, rhyme accuracy remains low (14.3% vs. Shakespeare's 83.3%), suggesting that rhyme requires longer-range planning beyond per-step logit adjustment. (3) *Incomplete sonnets*: 16.7% of debate sonnets have fewer than 11 lines, indicating that the model sometimes generates an end-of-text token prematurely.

6.3 Task Comparison and Calibration

Classification vs. generation. The contrast between paraphrase and sonnet results is instructive. For classification, the council provides marginal gains because GPT-2 is already well-calibrated on a straightforward binary task—the single best model achieves 88.96%, leaving little room for improvement. For generation, however, the space of possible outputs is enormous and structural constraints are hard to enforce through standard sampling alone. The council debate mechanism provides exactly the kind of soft constraint that helps here: a bias toward consensus without eliminating diversity.

Calibration. For paraphrase detection, the debate variant trades off ECE (0.051 vs. 0.017) for lower disagreement (5.0% vs. 9.6%). The higher ECE arises because debate's logit revision changes the confidence distribution, but the reduced disagreement indicates more reliable predictions on contested examples.

7 Conclusion

We introduced CouncilGPT-2, a multi-model inference framework that applies ensemble and debate strategies to both classification and generation tasks with GPT-2. Our main finding is that **logit-level debate is most valuable for structured generation**: while classification improvements are marginal, debate dramatically improves the structural quality of generated sonnets, achieving 84.1% pentameter adherence and the closest syllable count to iambic pentameter among all methods.

We attribute debate's effectiveness to its role as an implicit regularizer: by blending each model's predictions with peer consensus at every generation step, debate biases token selection toward metrically conventional choices. This insight suggests that council debate could be broadly useful for generation tasks with structural constraints (e.g., code generation, structured data output, other poetic forms).

Limitations. Our study uses only GPT-2 small (117M parameters) with $N{=}3$ council members. The debate parameter $\alpha{=}0.5$ was not tuned. Rhyme accuracy remains poor (14.3%), suggesting that per-step logit adjustment is insufficient for long-range structural constraints. Inference cost scales linearly with the number of members.

Future work. Promising directions include: scaling to larger models and more members, learning the debate parameter α, extending debate to multi-round protocols with explicit revision, and applying the framework to other structured generation tasks such as code synthesis or constrained text generation.

References

[1] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000.

[2] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024.

[3] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[4] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413, 2017.

[5] Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958. Association for Computational Linguistics, 2018.

[6] Jack Hopkins and Douwe Kiela. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178. Association for Computational Linguistics, 2017.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[8] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics, 2015.

A Appendix: Full Structural Metrics

Table 4 provides the complete set of structural metrics including line counts and 14-line accuracy.

Table 4: Full structural metrics for sonnet generation.

Method	**Lines**	**14-line**	**Compl**	**Syl/ln**	**SylDev**	**Penta**	**Rhyme**
Shakespeare	14.0	100%	100%	10.2	0.2	100%	83.3%
Single (seed 0)	11.5	0.0%	75.0%	12.0	3.1	53.8%	7.1%
Single (seed 1)	11.8	33.3%	75.0%	11.3	2.1	67.6%	10.7%
Single (seed 2)	11.2	16.7%	58.3%	11.0	2.3	68.0%	10.7%
Token ensemble	11.5	8.3%	66.7%	12.1	3.0	55.7%	7.1%
Best-of-N	10.8	8.3%	75.0%	11.7	2.7	56.2%	9.5%
Debate	11.7	16.7%	83.3%	9.8	1.2	84.1%	14.3%