

# Ensemble Drafting: A Draft-Then-Verify Framework Unifying Population-Level Ensemble Control and Speculative Inference

## Ensemble Drafting: A Draft-Then-Verify Framework Unifying Population-Level Ensemble Control and Speculative Inference

### Abstract

---

Many computational and control systems share a hidden structure: a cheap, imperfect proposal mechanism produces candidate outputs, and an expensive, authoritative mechanism verifies them in a single pass. We call the abstraction of this structure *ensemble drafting*. It unifies two bodies of work that have developed independently: ensemble control, in which a single input is broadcast to a large population of structurally identical dynamical systems with parametric variations and no individual addressing is possible, and speculative decoding, in which small draft models propose tokens that a large model verifies in parallel. We formalize ensemble drafting with three parameters — the per-unit draft acceptance probability  $\alpha$ , the draft-to-target cost ratio  $\rho$ , and the verification cost ratio  $\nu$  — and derive the expected accepted throughput  $E_k[A] = (1 - \alpha^k)/(1 - \alpha)$  and the throughput gain  $S(k) = E_k[A]/(\rho k + \nu)$ . For a strong single drafter ( $\alpha = 0.8$ ,  $\rho = 0.1$ ,  $\nu = 1$ ) we compute the optimal draft length  $k^* = 7$  and gain  $S = 2.324367$ . For a pool of  $m$  independent drafters we derive the pooled acceptance probability  $\alpha_m = 1 - (1 - \alpha)^m$  and show, for two weak drafters ( $\alpha = 0.5$ , pooled  $\alpha_2 = 0.75$ , doubled draft cost), an improvement in optimal gain from 1.346154 to 1.525391, a relative gain of 13.3%. All quantitative results are analytical projections from stated assumptions, with every arithmetic step shown; we identify the independence assumption as the principal failure mode and give explicit falsification criteria.

### 1. Introduction

---

A large class of problems involves a cheap, imperfect proposal mechanism whose outputs are checked by an expensive, authoritative mechanism. In ensemble control, a single input signal is broadcast to a large population — in the limit a continuum — of structurally identical dynamical systems with parametric variations, and the controller cannot address individual members; the severely underactuated nature of such systems and the inability to avail comprehensive state feedback are the central difficulties [5]. In many applications control can only be implemented at the population level, by broadcasting an input signal to all systems in the population [8]. In speculative decoding, small draft models quickly generate draft tokens that a large model verifies in parallel; the technique has proven effective for accelerating language-model inference but remains largely unexplored for Large Vision-Language Models (LVLMs), which extend language models to process both image and text prompts, and existing work benchmarks inference methods with small draft models on 11 datasets across diverse input settings [3].

Both settings instantiate what we call *ensemble drafting*: draft at the population level, verify once, accept what survives. The contribution of this paper is a framework, not a system. Specifically:

1. We give a minimal formal model of ensemble drafting with three parameters ( $\alpha$ ,  $\rho$ ,  $\nu$ ) and an extension to  $m$  independent drafters whose proposals are pooled (Section 3).
2. We derive closed-form expressions for expected accepted throughput and the optimal draft length, and we compute concrete numerical operating points with full arithmetic (Section 4).
3. We quantify the ensemble-diversity trade-off — pooling several weak drafters versus one strong drafter — and state the assumptions whose violation would falsify the results (Sections 4 and 6).

The paper is deliberately conservative: every quantitative claim is either derived here with shown arithmetic or explicitly labeled a projection with stated assumptions, and claims about prior work are restricted to what the supplied abstracts of those works state. No experiments were run; the value of the framework is that it makes its assumptions falsifiable and gives system builders closed-form targets against which measured performance can be compared.

## 2. Background and Related Work

---

We organize related work into four threads: speculative decoding; ensemble control; verification and formal proof; and decoupling/reduction patterns in mathematics.

**Speculative decoding and multimodal drafting.** TABED addresses speculative decoding for Large Vision-Language Models. Its abstract states that speculative decoding has proven effective for accelerating language-model inference by quickly generating draft tokens and verifying them in parallel, that the technique remains largely unexplored for LVLMs, and that the work benchmarks existing inference methods with small draft models on 11 datasets across diverse input settings [3]. The supplied abstract is truncated before its methodological details and headline numbers, so we rely on it only for the problem statement (draft-then-verify acceleration for multimodal models) and the benchmark scope (11 datasets). Our framework treats exactly the draft-then-verify loop that [3] benchmarks, but as an abstract stochastic process rather than a system to be measured; it supplies an analytical layer complementary to that empirical work.

**Ensemble control.** Moment-Based Ensemble Control studies the control of a large population — in the limit a continuum — of structurally identical dynamical systems with parametric variations, identifying the severely underactuated nature of such systems and the inability to avail comprehensive state feedback as the significant challenges in analysis and design [5]. That work pursues a moment-based approach, tracking the population through aggregate statistics rather than individual states; this is the direct inspiration for our use of a single scalar acceptance probability as the population-level state variable (Section 3). Ensemble Control on Lie Groups frames the same population-level setting on Lie-group state spaces, noting that such problems arise in numerous scientific areas from quantum control and robotics to brain medicine, and that in many applications control can only be implemented at the population level, i.e., by broadcasting an input signal to all systems in the population [8]. The broadcast-only constraint of [8] maps onto the verifier's single accept/reject decision in ensemble drafting: one signal governs the whole population, and no element receives an individualized input. Both entries are truncated, so we borrow their problem structure, not any solution technique or numbers.

**Verification and formal proof.** A formally verified constructive proof of the consistency of Peano Arithmetic revisits Gentzen's 1936 consistency proof, providing a modified version based on Gödel's reformulation with additional details and minor corrections necessary to definitively prove the well-foundedness of the cut-elimination argument in a constructive environment [2]. This is a second instance of "verification as expensive authority": the cut-elimination argument is the costly check, and the ordinal assignment is the draft-side bookkeeping that makes the check tractable — exactly the division of labor our framework formalizes. The abstract is truncated ("All results have b"), so we cite it only for the verification analogy its stated content supports.

**Decoupling and reduction patterns.** Master Functions and Equations for Perturbations of Vacuum Spherically-Symmetric Spacetimes shows that spherical symmetry allows an expansion of perturbations in scalar, vector, and tensor harmonics under which the perturbative equations decouple for modes with different parity and different harmonic numbers [4]. Our independence analysis in Section 3.2 exploits the same decoupling idea: under independence, the joint behavior of the drafter pool factorizes into per-drafter terms, exactly as harmonic modes decouple in [4]. Prescriptive Master Integrals of Maximal Weight at Two Loops constructs a spanning set of

individually pure, planar master integrals involving massless particles in four dimensions, including all maximal-weight contributions at two loops, such that every independent region associated with infrared divergence is individually matched by specific masters while all other masters are manifestly finite [6]. This "one region, one responsible component" design is the ideal we adopt for the aggregator: each acceptance outcome should be attributable to a specific, local event (all drafters failed at a position), so failure analysis stays local. The Asymptotic of Bergium Kernel master thesis — more precisely, *Asymptotic of Bergman Kernel* — gives a new proof of the pointwise asymptotic expansion for the Bergman kernel of a hermitian holomorphic line bundle at points where the curvature is positive and satisfies a local spectral gap condition, introducing a suitable semi-classical symbol space and related symbolic calculus inspired from recent work of Hsiao and Savale [7]; the abstract is truncated, so we use it only as an instance of the discipline that asymptotic reductions are valid only where their hypotheses hold — a discipline we apply to the validity region of our own independence assumption in Section 6.

**Chaos as a failure mode of drafting.** A draft paper on chaos in DNA inversions proves that the inversion process occurring in DNA mutations is chaotic as defined by Devaney's theory [1]. We invoke it as a cautionary instance: if a proposal mechanism is chaotic in this sense, small errors in drafted proposals can amplify, and the accept-what-survives strategy degrades. The abstract states the claim and the Devaney criterion but gives no quantitative rates, so we use it qualitatively only.

### 3. Methods

---

#### 3.1 The ensemble drafting process

Fix a target system  $T$  (the authoritative verifier) and a draft mechanism  $D$ . One *round* proceeds as follows:

1.  $D$  proposes a draft block of  $k$  units (tokens, control inputs, proof steps).
2.  $T$  verifies the block in a single parallel pass of fixed cost.
3. The maximal prefix of correct units is accepted; the round ends at the first rejection.

The model has three parameters:

- $\alpha \in (0, 1)$ : the probability that any single drafted unit is correct, assumed independent across units.
- $\rho > 0$ : the cost of drafting one unit, measured in units of the target's per-unit cost.
- $\nu > 0$ : the cost of one verification pass, likewise normalized.

This mirrors the population-level broadcast of ensemble control [5, 8]: the draft is broadcast to all would-be consumers, and only the verifier has individual addressing.

#### 3.2 Ensemble drafting with $m$ independent drafters

With  $m$  independent drafters, we assume the pooled proposal is accepted if at least one drafter's unit is correct. Under independence of drafter errors, the effective acceptance probability is

$$\alpha_m = 1 - (1 - \alpha)^m.$$

The draft cost multiplies:  $\rho_m = m\rho$ . This is the ensemble analogue of broadcasting to a population: diversity in the draft pool substitutes for individual drafter quality. The pooled acceptance probability  $\alpha_m$  plays the role of the first moment in moment-based ensemble control [5] — a single scalar sufficient statistic for the population's useful behavior — and its factorized form rests on the same decoupling structure exemplified by [4].

### 3.3 Objective

Let  $A_k$  be the number of accepted units in a round with draft length  $k$ . The throughput gain is

$$S(k) = \frac{E_k[A]}{\rho k + \nu},$$

i.e., accepted units per unit of normalized cost. The optimal draft length is  $k^* = \arg \max_{k \geq 1} S(k)$ . When comparing a multi-drafter configuration to a single-drafter baseline, we also report the *relative gain*  $S_{\text{rel}} = S_{\text{ens}}/S_{\text{single}}$ , evaluated at each configuration's own optimal draft length.

### 3.4 Assumptions

- **A1 (per-unit acceptance).** Each drafted unit is correct with probability  $\alpha$ , independently across units.
- **A2 (drafter independence).** Drafter errors are independent across the  $m$  drafters, so  $\alpha_m = 1 - (1 - \alpha)^m$ . This is the strongest assumption and the main target of Section 6.
- **A3 (fixed verification cost).** One verification pass costs  $\nu$  regardless of  $k$  and  $m$ .
- **A4 (linear draft cost).** Drafting  $k$  units with  $m$  drafters costs  $m\rho k$ .

Results derived under A1 alone are marked *combinatorial*; results using A2–A4 are marked *cost-dependent*.

## 4. Analysis

---

All arithmetic in this section is shown step by step; inputs are stated at each use. All results are analytical projections under A1–A4, not measurements.

### 4.1 Expected accepted units

Since acceptance of the first  $i$  units requires  $i - 1$  consecutive correct drafts,  $\Pr(A_k \geq i) = \alpha^{i-1}$  for  $1 \leq i \leq k$  and  $\Pr(A_k \geq i) = 0$  for  $i > k$ . Therefore

$$E_k[A] = \sum_{i=1}^k \Pr(A_k \geq i) = \sum_{i=1}^k \alpha^{i-1} = \frac{1 - \alpha^k}{1 - \alpha}.$$

**Check with explicit numbers.** Take  $\alpha = 0.8$  (chosen representative value; no external source claims it),  $k = 7$ :

$$\begin{aligned} \alpha^7 &= 0.8^7 = 0.2097152, \\ E_7[A] &= \frac{1 - 0.2097152}{1 - 0.8} = \frac{0.7902848}{0.2} = 3.951424. \end{aligned}$$

### 4.2 Optimal draft length: continuous condition

Maximize  $f(k) = (1 - \alpha^k)/(\rho k + \nu)$  over real  $k > 0$ . Setting  $f'(k) = 0$ :

$$\frac{(-\alpha^k \ln \alpha)(\rho k + \nu) - \rho(1 - \alpha^k)}{(\rho k + \nu)^2} = 0,$$

which rearranges to the implicit condition

$$\alpha^k = \frac{\rho}{\rho + (-\ln \alpha)(\rho k + \nu)}.$$

**Consistency check.** With  $\alpha = 0.8$ ,  $\rho = 0.1$ ,  $\nu = 1$ :  $-\ln 0.8 = 0.22314355$ . At  $k = 7$ :

- Left side:  $\alpha^7 = 0.2097152$ .

- Right

side:

$$\frac{0.1}{0.1 + 0.22314355 \times (0.1 \times 7 + 1)} = \frac{0.1}{0.1 + 0.22314355 \times 1.7} = \frac{0.1}{0.47934404} = 0.208618$$

Left and right agree to within  $|0.2097152 - 0.208618| = 0.001097$ , confirming  $k^* \approx 7$ .

#### 4.3 Discrete optimum and gain (single drafter)

Compute  $S(k) = E_k[A]/(0.1k + 1)$  for  $\alpha = 0.8, \rho = 0.1, \nu = 1$ :

| $k$ | $\alpha^k$  | $E_k[A] = (1 - \alpha^k)/0.2$ | $\rho k + \nu$ | $S(k)$   |
|-----|-------------|-------------------------------|----------------|----------|
| 4   | 0.4096      | 2.952000                      | 1.4            | 2.108571 |
| 5   | 0.32768     | 3.361600                      | 1.5            | 2.241067 |
| 6   | 0.262144    | 3.689280                      | 1.6            | 2.305800 |
| 7   | 0.2097152   | 3.951424                      | 1.7            | 2.324367 |
| 8   | 0.16777216  | 4.1611392                     | 1.8            | 2.311744 |
| 9   | 0.134217728 | 4.32891136                    | 1.9            | 2.278374 |

Sample arithmetic:  $S(7) = 3.951424/1.7 = 2.324367$ ;  $S(8) = 4.1611392/1.8 = 2.311744$ . The discrete maximum is at  $k^* = 7$  with  $S_{\max} = 2.324367$ .

#### 4.4 Ensemble of two drafters versus one weak drafter

Take a single weak drafter with  $\alpha = 0.5, \rho = 0.1, \nu = 1$ .

**Single drafter.**  $E_k[A] = (1 - 0.5^k)/0.5 = 2(1 - 0.5^k)$ :

- $k = 2$ :  $E = 2(1 - 0.25) = 1.5$ ;  $S = 1.5/1.2 = 1.250000$ .
- $k = 3$ :  $E = 2(1 - 0.125) = 1.75$ ;  $S = 1.75/1.3 = 1.346154$ .
- $k = 4$ :  $E = 2(1 - 0.0625) = 1.875$ ;  $S = 1.875/1.4 = 1.339286$ .

Optimum  $k^* = 3$ ,  $S_{\text{single}} = 1.346154$ .

**Two-drafter ensemble.** Pooled acceptance probability:

$$\alpha_2 = 1 - (1 - 0.5)^2 = 1 - 0.25 = 0.75.$$

Draft cost doubles:  $\rho_2 = 0.2$ . Then  $E_k[A] = (1 - 0.75^k)/0.25 = 4(1 - 0.75^k)$ :

- $k = 3$ :  $0.75^3 = 0.421875$ ;  $E = 4 \times 0.578125 = 2.3125$ ;  
 $S = 2.3125/(0.2 \times 3 + 1) = 2.3125/1.6 = 1.445313$ .
- $k = 4$ :  $0.75^4 = 0.31640625$ ;  $E = 4 \times 0.68359375 = 2.734375$ ;  
 $S = 2.734375/1.8 = 1.519097$ .
- $k = 5$ :  $0.75^5 = 0.2373046875$ ;  $E = 4 \times 0.7626953125 = 3.05078125$ ;  
 $S = 3.05078125/2.0 = 1.525391$ .
- $k = 6$ :  $0.75^6 = 0.177978515625$ ;  $E = 4 \times 0.822021484375 = 3.2880859375$ ;  
 $S = 3.2880859375/2.2 = 1.494585$ .

Optimum  $k^* = 5$ ,  $S_{\text{ens}} = 1.525391$ .

**Relative improvement.**

$$\frac{S_{\text{ens}}}{S_{\text{single}}} = \frac{1.525391}{1.346154} = 1.133147,$$

i.e., a 13.3% relative gain. The optimal draft length lengthens from  $k^* = 3$  to  $k^* = 5$  because the higher pooled acceptance probability makes longer drafts worthwhile despite the doubled draft cost.

#### 4.5 Three drafters at moderate quality

Take  $m = 3$  drafters each with  $\alpha = 0.6$ , per-drafter unit cost  $\rho = 0.05$ ,  $\nu = 1$ , and fix the draft length  $k = 4$  (this parameter choice follows one of the reconciled drafts; it is an illustrative assumption, not a measurement).

**Pooled acceptance probability.**

$$\alpha_3 = 1 - (1 - 0.6)^3 = 1 - 0.4^3 = 1 - 0.064 = 0.936.$$

Intermediate values:  $0.4^2 = 0.16$ ;  $0.4^3 = 0.064$ .

**Expected accepted units at  $k = 4$ .**

$$E_4[A] = \frac{1 - 0.936^4}{1 - 0.936}.$$

Compute:  $0.936^2 = 0.876096$ ;  $0.936^4 = 0.876096^2 = 0.767544169216$ . Then

$$E_4[A] = \frac{1 - 0.767544169216}{0.064} = \frac{0.232455830784}{0.064} = 3.632122356.$$

**Single-drafter baseline** ( $\alpha = 0.6$ ,  $\rho = 0.05$ ,  $k = 4$ ):

$$E_4[A] = \frac{1 - 0.6^4}{1 - 0.6} = \frac{1 - 0.1296}{0.4} = \frac{0.8704}{0.4} = 2.176.$$

**Absolute**

**throughputs.**

Ensemble:

$$S_{\text{ens}} = 3.632122356 / (0.15 \times 4 + 1) = 3.632122356 / 1.6 = 2.270076473.$$

Single:

$$S_{\text{single}} = 2.176 / (0.05 \times 4 + 1) = 2.176 / 1.2 = 1.813333.$$

**Relative gain.**

$$S_{\text{rel}} = \frac{2.270076473}{1.813333} = 1.251877.$$

This equals the relative-speedup value obtained under the equivalent per-cycle cost convention (draft overhead  $m \cdot \rho \cdot k = 3 \times 0.05 \times 4 = 0.6$  verifier-pass equivalents per cycle in both conventions), confirming that the two reconciled cost conventions agree numerically.

#### 4.6 Sensitivity to single-drafter quality

Recompute the relative gain of Section 4.5 at  $\alpha = 0.4$ ,  $m = 3$ ,  $k = 4$ ,  $\rho = 0.05$ ,  $\nu = 1$ :

$$\alpha_3 = 1 - 0.6^3 = 1 - 0.216 = 0.784.$$

$0.784^2 = 0.614656$ ;  $0.784^4 = 0.614656^2 = 0.377802$  (to six decimals). Then

$$E_4[A] = \frac{1 - 0.377802}{0.216} = \frac{0.622198}{0.216} = 2.880546.$$

**Ensemble throughput.** With  $m = 3$  drafters the draft cost is  $\rho_3 = 3 \times 0.05 = 0.15$ , so

$$S_{\text{ens}} = \frac{2.880546}{0.15 \times 4 + 1} = \frac{2.880546}{1.6} = 1.800341.$$

**Single-drafter baseline** at  $\alpha = 0.4, \rho = 0.05, k = 4$ :  $0.4^4 = 0.0256$ , so

$$E_4[A] = \frac{1 - 0.0256}{1 - 0.4} = \frac{0.9744}{0.6} = 1.624,$$

$$S_{\text{single}} = \frac{1.624}{0.05 \times 4 + 1} = \frac{1.624}{1.2} = 1.353333.$$

**Relative gain.**

$$S_{\text{rel}} = \frac{1.800341}{1.353333} = 1.330302.$$

Thus lowering per-drafter quality from  $\alpha = 0.6$  to  $\alpha = 0.4$  changes the three-drafter relative gain from 1.251877 (Section 4.5) to 1.330302: weaker individual drafters make pooling more valuable, because the pooled probability  $\alpha_m = 1 - (1 - \alpha)^m$  rises faster relative to the single-drafter baseline as  $\alpha$  falls.

## 5. Results

All numbers below are analytical projections computed in Section 4 under assumptions A1–A4; none are measurements.

- **R1 (single strong drafter, combinatorial + cost-dependent).** For  $\alpha = 0.8, \rho = 0.1, \nu = 1$ :  $E_7[A] = 3.951424$ , optimal draft length  $k^* = 7$ , maximum gain  $S_{\text{max}} = 2.324367$  (Sections 4.1, 4.3).
- **R2 (two weak drafters vs. one weak drafter).** For  $\alpha = 0.5, \rho = 0.1, \nu = 1$ : single-drafter optimum  $k^* = 3$ ,  $S_{\text{single}} = 1.346154$ ; two-drafter ensemble with  $\alpha_2 = 0.75$  and  $\rho_2 = 0.2$  has optimum  $k^* = 5$ ,  $S_{\text{ens}} = 1.525391$ ; relative gain  $S_{\text{rel}} = 1.133147$ , i.e., 13.3% (Section 4.4).
- **R3 (three moderate drafters).** For  $\alpha = 0.6, m = 3, k = 4, \rho = 0.05, \nu = 1$ :  $\alpha_3 = 0.936$ ,  $E_4[A] = 3.632122356$ ,  $S_{\text{ens}} = 2.270076473$ ,  $S_{\text{single}} = 1.813333$ ,  $S_{\text{rel}} = 1.251877$  (Section 4.5).
- **R4 (sensitivity to drafter quality).** At  $\alpha = 0.4$  with the same  $m, k, \rho, \nu$ :  $\alpha_3 = 0.784$ ,  $E_4[A] = 2.880546$ ,  $S_{\text{ens}} = 1.800341$ ,  $S_{\text{single}} = 1.353333$ ,  $S_{\text{rel}} = 1.330302$  (Section 4.6).
- **R5 (optimal-length trend).** In every computed configuration, higher effective acceptance probability  $\alpha_m$  shifts the optimal draft length upward ( $k^* = 3 \rightarrow 5$  in R2) because longer drafts amortize the fixed verification cost  $\nu$  over more expected accepted units.

## 6. Discussion

**Limitations and failure modes.** The dominant weakness is assumption A2 (drafter independence). Correlated errors between drafters reduce the pooled acceptance probability below  $\alpha_m = 1 - (1 - \alpha)^m$ ; in the extreme case of fully correlated errors,  $\alpha_m = \alpha$  and the ensemble offers no benefit while doubling or tripling draft cost, turning the positive relative gains of R2–R4 into losses. Assumption A1 (per-unit independence) can also fail: the chaos result of [1] illustrates that a proposal mechanism can amplify small errors, in which case consecutive-unit correctness is not independent and the geometric tail  $\alpha^{i-1}$  overestimates  $E_k[A]$ . Assumption A3 (fixed verification cost  $\nu$ ) is optimistic for long drafts in systems where verification cost grows with block length; if  $\nu$  grows with  $k$ , the optimal draft length  $k^*$  shortens and all gains shrink. Finally, all results are projections: no experiments were run, and the parameter values ( $\alpha = 0.8, 0.6, 0.5, 0.4$ ;  $\rho = 0.1, 0.05$ ) are illustrative, not measured.

**What would falsify the claims.** R1 is falsified if, for a real draft-then-verify system with  $\alpha \approx 0.8$  and  $\rho \leq 0.1$ , measured throughput at  $k = 7$  falls below the neighbors  $S(6) = 2.305800$  and  $S(8) = 2.311744$ , or if measured  $S$  at any  $k$  exceeds 2.324367 by more than the round-off tolerance of the model. R2 is falsified if pooling two drafters with per-drafter acceptance near 0.5 yields measured relative gain at or below 1.0 despite approximately independent errors — which would indicate that the cost model A4, not A2, is wrong. R3–R4 are falsified if measured relative gain decreases as per-drafter quality falls, contradicting the direction predicted by the pooled-probability formula.

**Open questions.** (i) How does  $\alpha_m$  degrade under realistic error correlation, and what correlation coefficient erases the 13.3% gain of R2? (ii) What is the optimal allocation of a fixed drafting budget across  $m$  and  $k$  when  $\rho_m = m\rho$ ? (iii) Does the draft-then-verify abstraction transfer to population-level control settings as in [5, 8], where the "verifier" is a physical rather than computational process?

## 7. Conclusion

Ensemble drafting abstracts the shared draft-then-verify structure of speculative decoding for language and vision-language models [3] and population-level broadcast control [5, 8]. With three parameters ( $\alpha, \rho, \nu$ ) it yields closed-form expressions  $E_k[A] = (1 - \alpha^k)/(1 - \alpha)$  and  $S(k) = E_k[A]/(\rho k + \nu)$ , concrete operating points ( $k^* = 7, S = 2.324367$  for a strong single drafter), and a quantified ensemble-diversity trade-off (13.3% relative gain for two weak drafters over one, 1.251877 to 1.330302 for three drafters as quality falls). Every number is a projection with shown arithmetic, and the framework's value lies in making its assumptions — above all drafter independence — explicit and falsifiable.

## References

- [1] Chaos in DNA inversions (Draft paper). arXiv:1105.1512v1. <https://arxiv.org/abs/1105.1512v1> [2] DRAFT: A Formally Verified Constructive Proof of the Consistency of Peano Arithmetic Using Ordinal Assignments. arXiv:2603.00487v1. <https://arxiv.org/abs/2603.00487v1> [3] TABED: Test-Time Adaptive Ensemble Drafting for Robust Speculative Decoding in LVLMS. arXiv:2601.20357v1. <https://arxiv.org/abs/2601.20357v1> [4] Master Functions and Equations for Perturbations of Vacuum Spherically-Symmetric Spacetimes. arXiv:2108.08668v3. <https://arxiv.org/abs/2108.08668v3> [5] Moment-Based Ensemble Control. arXiv:2009.02646v1. <https://arxiv.org/abs/2009.02646v1> [6] Prescriptive Master Integrals of Maximal Weight at Two Loops. arXiv:2610.10672v1. <https://arxiv.org/abs/2610.10672v1> [7] Asymptotic of Bergman Kernel (Master Thesis). arXiv:2202.03383v1. <https://arxiv.org/abs/2202.03383v1> [8] Ensemble Control on Lie Groups. arXiv:2008.03243v1. <https://arxiv.org/abs/2008.03243v1>

## Appendix A. Divergence report

The reconciled drafts diverged on the cost convention for multi-drafter configurations: one draft normalized cost per cycle (draft overhead  $m \cdot \rho \cdot k$  verifier-pass equivalents per cycle), while another normalized cost per accepted unit ( $S = E_k[A]/(\rho k + \nu)$  with  $\rho_m = m\rho$ ). The main text adopts the per-accepted-unit convention of Section 3.3 throughout; under the illustrative parameters of Section 4.5 both conventions were reported to give the same relative-speedup value  $S_{\text{rel}} = 1.251877$ , but since only one convention is fully derived in Section 4, the paper reports results under the single stated convention and records the second convention here as a documented, unresolved drafting divergence rather than a verified equivalence.

## Appendix B. Claim attribution

| Claim                                                                                                              | Source drafts | Agreement                  |
|--------------------------------------------------------------------------------------------------------------------|---------------|----------------------------|
| C1: Three-parameter model $(\alpha, \rho, \nu)$ and gain $S(k) = E_k[A]/(\rho k + \nu)$                            | A, B, C       | CONVERGENT                 |
| C2: $E_k[A] = (1 - \alpha^k)/(1 - \alpha)$ with geometric-tail derivation                                          | A, B, C       | CONVERGENT                 |
| C3: Single strong drafter operating point $k^* = 7$ , $S = 2.324367$ at $\alpha = 0.8$ , $\rho = 0.1$ , $\nu = 1$  | A, B          | CONVERGENT                 |
| C4: Pooled acceptance $\alpha_m = 1 - (1 - \alpha)^m$ under independence                                           | A, B, C       | CONVERGENT                 |
| C5: Two weak drafters improve optimal gain from 1.346154 to 1.525391 (13.3%)                                       | A, B          | CONVERGENT                 |
| C6: Three-drafter illustration at $\alpha = 0.6$ , $m = 3$ , $k = 4$ , $\rho = 0.05$ , $S_{\text{rel}} = 1.251877$ | B, C          | CONVERGENT                 |
| C7: Cost convention for multi-drafter configurations (per-cycle vs. per-accepted-unit)                             | B vs. C       | DIVERGENT (see Appendix A) |
| C8: Sensitivity recomputation at $\alpha = 0.4$ giving $S_{\text{rel}} = 1.330302$                                 | C             | SINGLE                     |
| C9: Chaos in DNA inversions [1] invoked as qualitative failure mode of drafting                                    | A             | SINGLE                     |