Critical Comparison: Automated af Proof Attempt vs. Official Solution for Problem 3

First Proof Benchmark — Markov Chain with
ASEP Polynomial Stationary Distribution

Comparative Analysis Report
First Proof Project

February 2026

Abstract

Problem 3 of the First Proof benchmark (posed by Lauren Williams) asks whether there exists a *nontrivial* Markov chain on the set of compositions $S_n(\lambda)$ of a restricted partition λ whose stationary distribution is the ratio F^*_μ/P^*_λ of interpolation ASEP to interpolation Macdonald polynomials at $q = 1$. The official solution, by Ben Dali and Williams, constructs the *interpolation* t-PushTASEP — a modification of the ordinary t-PushTASEP with a new "Step 2" reinsertion phase governed by site-dependent probabilities $\mathfrak{p}_k$ and $\mathfrak{q}_k$. The automated `af` proof attempt also answered YES, but proposed the *ordinary* (inhomogeneous multispecies) t-PushTASEP of Ayyer–Martin–Williams as the Markov chain, and attempted to bridge the gap via a "ratio identity" $f^*_\mu/P^*_\lambda = f_\mu/P_\lambda$. This report critically compares the two approaches, evaluating what the `af` attempt got right and what it missed.

Contents

1 Problem Statement

Let $\lambda = (\lambda_1 > \cdots > \lambda_n \geq 0)$ be a partition with distinct parts that is *restricted* (unique part 0, no part 1). Let $S_n(\lambda)$ be the set of compositions that are permutations of the parts of λ. For $\mu \in S_n(\lambda)$, let $F^*_\mu(x_1, \ldots, x_n; q, t)$ be the interpolation ASEP polynomial and $P^*_\lambda(x_1, \ldots, x_n; q, t)$ the interpolation Macdonald polynomial.

Conjecture 1.1 (Williams)**.** *Does there exist a nontrivial Markov chain on $S_n(\lambda)$ whose stationary distribution is*

$$\pi(\mu) = \frac{F^*_\mu(x_1, \ldots, x_n;\ q{=}1,\ t)}{P^*_\lambda(x_1, \ldots, x_n;\ q{=}1,\ t)} \qquad \textit{for } \mu \in S_n(\lambda),$$

where "nontrivial" means the transition probabilities are not described using the polynomials F^*_μ?

The answer is YES. Both the official solution and the `af` attempt agree on this. They disagree, however, on which Markov chain achieves this stationary distribution and how to prove it.

2 The Official Solution (Ben Dali–Williams)

2.1 The Interpolation t-PushTASEP

The official solution constructs a *new* Markov chain: the **interpolation t-PushTASEP** with content λ. This is *not* the same as the ordinary t-PushTASEP of Ayyer–Martin–Williams [1]. It is a three-step process on configurations $\mu \in S_n(\lambda)$, viewed as particles on a ring:

(Step 0) A position j is selected with probability P_j proportional to a product involving $(x_k - t^{-(n-2)})$ and $(x_k - t^{-(n-1)})$ factors.

(Step 1) The particle at position j (label a) is activated and travels clockwise, displacing weaker particles via the standard t-PushTASEP geometric rule: with probability $t^{k-1}/[m]_t$ it moves to the k-th of the m weaker particles. The cascade continues until a vacancy is reached. Position j becomes vacant.

(Step 2) The vacancy at position j (now labeled $a := 0$) travels to position 1 and proceeds clockwise through positions $1, \ldots, j-1$, interacting with each particle b at site k via site-dependent probabilities:

$$\mathfrak{p}_k = \frac{t^{-n+1}(1-t)}{x_k - t^{-n+2}}, \qquad \mathfrak{q}_k = \frac{(1-t)\, x_k}{x_k - t^{-n+2}}.$$

The vacancy either skips or displaces each particle it encounters, with probabilities depending on the relative labels.

The key innovation is Step 2: a reinsertion phase that has no analogue in the ordinary t-PushTASEP. This step is what makes the chain's stationary distribution involve *interpolation* polynomials rather than the homogeneous ones.

2.2 Key Ideas in the Proof

The proof of Theorem 2.2 (that the stationary probability of μ is F^*_μ/P^*_λ) proceeds through several stages:

1. **Two-line queue encoding (Proposition 3.6).** Step 2 of the interpolation t-PushTASEP is encoded by *unsigned two-line queues* $\bar{Q} \in \bar{\mathcal{G}}^\rho_\nu$. The weight of $\bar{Q}$ (defined via ball weights and pairing weights) divided by a normalizing product equals the transition probability $\mathbb{P}^{(2)}_j(\rho, \nu)$.

2. **Signed-to-unsigned reduction (Lemma 3.2).** Signed two-line queues from [2] are related to the unsigned versions by summing over all sign assignments. The sign cancellations produce the ball weights in Equation (6) of the official solution.

3. **Factorization via Corollary 3.5.** Using the strong factorization property of interpolation Macdonald polynomials at $q = 1$ (from Dołęga [4] and [2]), the authors establish:

$$F^*_\rho(\mathbf{x}; 1, t) = F^*_\eta(\mathbf{x}; 1, t) \cdot e^*_k(\mathbf{x}; t),$$

   where $\eta = \rho^-$ (parts decremented by 1) and k is the number of nonzero parts. This uses Theorem 3.4, an interpolation analogue of [1, Theorem 4.18].

4. **Stationarity via balance equation (Proposition 3.7).** Combining [1, Lemma 5.4] (for Step 1 transition probabilities) with Proposition 3.6 (for Step 2), the full transition probability decomposes as

$$\mathbb{P}(\mu, \nu) = \sum_{\rho \in S_n(\lambda)} \frac{a^\mu_\rho \, c^\rho_\nu}{e^*_{n-1}}.$$

   Then from the expansion $F^*_\nu = \sum_\eta F^{*\eta}_\nu \cdot F^*_{\eta^-}$ (from [2, Theorem 1.15 and Lemma 5.6]) and Corollary 3.5, one obtains:

$$F^*_\nu(\mathbf{x}; 1, t) = \sum_\eta F^*_\eta(\mathbf{x}; 1, t) \cdot \mathbb{P}(\eta, \nu),$$

   proving that $(F^*_\mu)_\mu$ is a left eigenvector of the transition matrix with eigenvalue 1.

3 The af Automated Proof Attempt

3.1 Overview

The `af` attempt over three adversarial sessions produced a 9-node proof tree (root plus 8 children). It answered YES and proposed the **ordinary** (inhomogeneous multispecies) t-PushTASEP of Ayyer–Martin–Williams [1] as the Markov chain.

The proof strategy was:

1. **Node 1.1:** Define the state space $S_n(\lambda)$.

2. **Node 1.2:** Establish notation and the relationship between interpolation polynomials f^*_μ and homogeneous polynomials f_μ.

3. **Node 1.4:** Construct the ordinary t-PushTASEP.

4. **Node 1.5:** Cite AMW24 Theorem 1.1 that the ordinary t-PushTASEP has stationary distribution f_μ/P_λ.

5. **Node 1.6 (Crux):** Establish the "ratio identity" $f^*_\mu/P^*_\lambda = f_\mu/P_\lambda$.

6. **Node 1.3:** Derive positivity and normalization as a corollary.

7. **Node 1.7:** Prove nontriviality via a sparsity argument.

8. **Node 1.8:** Conclude.

3.2 The Ratio Identity (Node 1.6)

The crux of the `af` attempt is the claim that for all $\mu \in S_n(\lambda)$:

$$\frac{f^*_\mu(\mathbf{x}; 1, t)}{P^*_\lambda(\mathbf{x}; 1, t)} = \frac{f_\mu(\mathbf{x}; 1, t)}{P_\lambda(\mathbf{x}; 1, t)}. \tag{1}$$

The proposed proof had three steps:

1. Both f^*_μ and f_μ satisfy the same Hecke relations (BDW25 Proposition 2.10).
2. The AMW24 balance equation proof for f_μ "transfers" to f^*_μ because it uses only these Hecke relations.
3. Perron–Frobenius uniqueness forces proportionality: $f^*_\mu = C \cdot f_\mu$ for a constant C independent of μ.

3.3 Session History

The `af` attempt went through three sessions:

- **Session 1:** Proof tree registration (no verification).
- **Session 2:** Adversarial verification raised 97 challenges (36 critical, 49 major, 12 minor). Five systemic problems were identified: wrong arXiv reference, fabricated Ferrari–Martin proof method, logical fallacy in Node 1.6, circular dependency, and notation inconsistency.
- **Session 3:** All 8 leaf nodes were rewritten, resolving all 97 challenges. Node 1.6 was completely rewritten with the three-step Hecke argument. Node 1.3 was restructured as a corollary. No further verification was performed.

The proof tree ended with 9 pending nodes, 0 validated, 0 refuted, and 0 open challenges.

4 Critical Comparison

4.1 The Fundamental Difference: Which Markov Chain?

The most significant difference between the two approaches is the choice of Markov chain:

	Official Solution	**`af` Attempt**
Markov chain	Interpolation t-PushTASEP	Ordinary t-PushTASEP
Dynamics	Steps 0 + 1 + 2 (novel Step 2)	Standard t-PushTASEP
Stationary dist.	F^*_μ/P^*_λ directly	f_μ/P_λ (then ratio identity)
Key novelty	New chain with reinsertion	Ratio identity

The official solution constructs a *genuinely different* Markov chain whose stationary distribution is *directly* F^*_μ/P^*_λ, without needing to pass through the homogeneous polynomials. The `af` attempt, by contrast, tried to reuse the *same* chain (the ordinary t-PushTASEP) and bridge the gap algebraically.

This is a crucial distinction. The authors' commentary on AI-generated solutions explicitly notes that a common LLM failure mode was "to change the problem to a related but different, and already-solved problem, namely, to replace interpolation ASEP and interpolation Macdonald polynomials by ASEP and Macdonald polynomials. In this case the solution to this problem is the t-PushTASEP and was given in a paper by Ayyer, Martin, and Williams." The `af` attempt falls squarely into this category.

4.2 Is the Ratio Identity True?

The `af` attempt's entire strategy depends on the ratio identity (1). Does the official solution validate it?

Yes, but indirectly. The official solution's Theorem 3.4 establishes that

$$\frac{G^*_\eta(\mathbf{x};t)}{P^*_\lambda(\mathbf{x};1,t)} = \frac{F^*_\eta(\mathbf{x};1,t)}{P^*_\kappa(\mathbf{x};1,t)},$$

where $G^*_\eta = \sum_{\rho:\phi(\rho)=\eta} F^*_\rho$. In the special case where λ is restricted and $\phi : i \mapsto \max(i-1,0)$, the map ϕ is injective on $S_n(\lambda)$, so $G^*_\eta = F^*_\rho$ for a unique ρ. This, combined with the factorization $P^*_\lambda/P^*_\kappa = e^*_k$ (Corollary 3.5), gives relationships between interpolation polynomials at *different* partitions.

However, the official solution does *not* establish the ratio identity (1) in the form claimed by the `af` attempt (i.e., $f^*_\mu/P^*_\lambda = f_\mu/P_\lambda$ at the *same* partition λ). Instead, the official solution avoids this identity entirely by constructing a different Markov chain that directly has F^*_μ/P^*_λ as its stationary distribution.

That said, the ratio identity (1) is in fact a *consequence* of both solutions being correct: if the ordinary t-PushTASEP has stationary distribution f_μ/P_λ (AMW24 Theorem 1.1) and the interpolation t-PushTASEP has stationary distribution F^*_μ/P^*_λ (Ben Dali–Williams Theorem 2.2), and if the two chains happen to have the *same* stationary distribution, then the ratio identity holds. But this would require showing that the ordinary t-PushTASEP and the interpolation t-PushTASEP have the same stationary distribution, which is by no means obvious given their different dynamics.

In summary: the ratio identity is a nontrivial claim that the official solution neither proves nor needs. The `af` attempt's strategy of reducing to this identity was the wrong approach.

4.3 Did the `af` Attempt Identify the Right Techniques?

Technique	**Official**	`af`
Multiline queues	Central	Mentioned
Signed two-line queues	Central	Not used
Unsigned two-line queues	Central	Not used
Hecke relations (BDW25)	Not used	Central
AMW24 Thm 1.1	Used (Step 1)	Central
Factorization (P^*_λ)	Central	Not used
Perron–Frobenius	Not used	Central
BDW25 Thm 1.15 ($f^* = F^*$)	Used	Used

The official solution relies heavily on the *combinatorial* machinery of two-line queues (both signed and unsigned) and the factorization of interpolation Macdonald polynomials. The `af` attempt instead relied on the *algebraic* machinery of Hecke operators and Perron–Frobenius uniqueness. These are fundamentally different proof strategies.

Notably, the official solution does use AMW24, but only for the Step 1 transition probabilities ([1, Lemma 5.4]), not for its stationarity theorem. The `af` attempt treated AMW24 Theorem 1.1 as the centerpiece, which reflects the misidentification of the Markov chain.

5 Evaluation of the `af` Ratio Identity (Node 1.6)

The `af` attempt's Node 1.6 claimed the ratio identity (1) via three steps. We evaluate each in light of the official solution.

5.1 Step 1: Hecke Relations

The claim that both f^*_μ and f_μ satisfy the same Hecke relations (BDW25 Proposition 2.10) is **correct**. This is a published result. However, the official solution does not use Hecke relations at all, suggesting they are not the right tool for this problem.

5.2 Step 2: Balance Transfer

The claim that the AMW24 balance equation proof "transfers" from f_μ to f^*_μ because both satisfy the same Hecke relations is **unverified and likely false as stated**. The argument asserts that the AMW24 proof uses *only* the Hecke relations (a)–(c), and therefore applies verbatim to f^*_μ.

There are several problems:

1. The AMW24 proof of stationarity is specific to the *ordinary* t-PushTASEP. Even if f^*_μ satisfies the same Hecke relations, the relevant balance equation involves the *transition matrix* of the chain. If the chain is the ordinary t-PushTASEP, then $\sum_\mu f^*_\mu \cdot Q_{\mu,\nu} = 0$ would mean that f^*_μ is *also* a stationary measure of the ordinary t-PushTASEP. By Perron–Frobenius uniqueness, this would force $f^*_\mu = C \cdot f_\mu$, i.e., the interpolation polynomials are proportional to the homogeneous ones.

2. But f^*_μ and f_μ are *not* proportional in general: f_μ is the top-degree homogeneous component of f^*_μ (BDW25 Theorem 2.3), and f^*_μ contains lower-degree correction terms. Hence $f^*_\mu \neq C \cdot f_\mu$ for any constant C (unless the lower-degree terms vanish, which they do not in general).

3. This means Step 2 of the `af` argument is **false**: the AMW24 proof does *not* transfer from f_μ to f^*_μ for the ordinary t-PushTASEP, because f^*_μ is not in the left null space of the ordinary t-PushTASEP's generator Q.

The official solution confirms this diagnosis: a *different* Markov chain (the interpolation t-PushTASEP, with its additional Step 2) is needed to have F^*_μ as stationary weights. The ordinary t-PushTASEP only gives f_μ as stationary weights.

5.3 Step 3: Perron–Frobenius Uniqueness

The Perron–Frobenius argument is **correct in principle** but is applied to a false premise (Step 2). If f^*_μ were indeed in the left null space of the ordinary t-PushTASEP's generator, then Perron–Frobenius would force proportionality. But since f^*_μ is *not* in that null space, the argument is vacuous.

5.4 Verdict on Node 1.6

The ratio identity as formulated by the `af` attempt is **false**. The interpolation polynomials f^*_μ are not proportional to the homogeneous polynomials f_μ (they differ by lower-degree terms), and the ordinary t-PushTASEP does not have f^*_μ/P^*_λ as its stationary distribution. The `af` attempt's HANDOFF.md itself flagged this as a "high risk" node, and this assessment was correct: the node contains a fundamental error.

6 What the `af` Attempt Got Right

Despite the fundamental error in the choice of Markov chain, the `af` attempt made several correct identifications:

1. **The answer is YES.** The `af` attempt correctly identified that a nontrivial Markov chain exists.

2. **The t-PushTASEP family is the right family of chains.** The official solution's interpolation t-PushTASEP is a modification of the ordinary t-PushTASEP. The `af` attempt correctly identified the t-PushTASEP as the relevant dynamical framework.

3. **Correct use of published literature.** After Session 3 corrections, the `af` attempt correctly cited AMW24 Theorem 1.1, BDW25 Proposition 2.10, and other results. The notation table (Node 1.2) was comprehensive.

4. **The nontriviality argument (Node 1.7) is sound.** The sparsity argument — exhibiting a zero in the transition matrix — is a valid proof of nontriviality. It would work for the interpolation t-PushTASEP as well, since the interpolation chain also has sparse transition matrices (Step 1 only moves particles clockwise).

5. **Clean logical structure after Session 3.** The restructuring of Node 1.3 as a corollary of 1.5 + 1.6 was a good logical move that broke the circularity. The dependency chain 1.1 → 1.2 → 1.4 → 1.5 → 1.6 → 1.3 → 1.7 → 1.8 is clean.

6. **The adversarial process caught real errors.** Session 2 correctly identified the logical fallacy in the original Node 1.6 ("both sum to 1, therefore equal"), the fabricated Ferrari–Martin proof method, and the circular dependency. The adversarial framework worked as intended.

7 What the `af` Attempt Missed

1. **The need for a new Markov chain.** The most critical miss. The problem asks for a Markov chain whose stationary distribution involves *interpolation* polynomials, not homogeneous ones. The ordinary t-PushTASEP gives the homogeneous ratio f_μ/P_λ, not the interpolation ratio F^*_μ/P^*_λ. A new chain — the interpolation t-PushTASEP with its Step 2 reinsertion — is required.

2. **The combinatorial machinery of two-line queues.** The official solution is fundamentally combinatorial, encoding the Step 2 transitions via signed and unsigned two-line queues. The `af` attempt did not engage with this machinery at all, instead pursuing an algebraic approach via Hecke operators.

3. **The factorization of interpolation Macdonald polynomials.** The official solution crucially uses $P^*_\lambda(\mathbf{x};1,t) = \prod_i e^*_{\lambda'_i}(\mathbf{x};t)$ (from Dołęga and BDW25), which provides the key Corollary 3.5 relating F^*_ρ to $F^*_{\rho^-}$. The `af` attempt mentioned BDW25 Theorem 7.1 (factorization) as an "alternative route" but never pursued it.

4. **The distinction between interpolation and homogeneous polynomials.** While the `af` attempt acknowledged that f_μ is the top-degree component of f^*_μ (BDW25 Theorem 2.3), it did not fully reckon with the implications: $f^*_\mu \neq C \cdot f_\mu$, so the ordinary t-PushTASEP cannot have f^*_μ as stationary weights.

5. **The role of the BDW25 expansion.** The official proof uses the expansion $F^*_\nu = \sum_\eta F^{*\eta}_\nu \cdot F^*_{\eta^-}$ from [2, Theorem 1.15 and Lemma 5.6] as the starting point for the stationarity proof. The `af` attempt did not use this result.

6. **The inherent difficulty flag in BDW25 Remark 1.17.** The `af` attempt correctly noted that BDW25 Remark 1.17 defers the interpolation probabilistic interpretation to a forthcoming paper, but drew the wrong conclusion: it treated this as suggesting the Hecke transfer argument merely needs careful verification, rather than as a signal that a genuinely new construction is needed.

8 Alignment with Common LLM Failure Modes

The authors' commentary identifies two common LLM failure modes for Problem 3:

1. **Metropolis–Hastings (trivial solution).** Using the desired distribution formula to define transition rates. The `af` attempt avoided this failure mode — it correctly recognized that this would be trivial.

2. **Replacing interpolation polynomials by homogeneous ones.** Solving the already-solved problem for f_μ/P_λ (ordinary ASEP/Macdonald polynomials) instead of the interpolation versions F^*_μ/P^*_λ. The `af` attempt *fell into this exact failure mode*, proposing the ordinary t-PushTASEP and attempting to bridge the gap with the ratio identity.

The `af` attempt's approach is essentially a more sophisticated version of failure mode 2: rather than simply ignoring the "interpolation" qualifier, it acknowledged the distinction and attempted an algebraic bridge. But the bridge (the ratio identity) is built on the false premise that f^*_μ is proportional to f_μ.

9 Lessons Learned

1. **LLMs tend to reduce novel problems to known ones.** The `af` attempt replaced the open problem (interpolation t-PushTASEP stationarity) with the solved problem (ordinary t-PushTASEP stationarity) and tried to bridge the gap algebraically. This is a pervasive LLM failure mode: rather than constructing a genuinely new object, the system reuses a known object and claims the difference does not matter.

2. **The adversarial framework caught errors but not the fundamental one.** Sessions 2 and 3 caught fabricated proofs (Ferrari–Martin), logical fallacies ("both sum to 1"), circular dependencies, and notation errors. But they did not catch the deepest error: that the *wrong Markov chain* was proposed. The adversarial verifiers correctly flagged Node 1.6 as high-risk but could not definitively refute it.

3. **Risk assessment was accurate.** The HANDOFF.md correctly identified Node 1.6 as the highest-risk node and acknowledged the critical caveat about Step 2. The system "knew" where the proof was weakest but could not resolve the issue.

4. **The ratio identity approach is seductive but wrong.** The identity $f^*_\mu/P^*_\lambda = f_\mu/P_\lambda$ looks natural and would elegantly reduce the problem to known results. But it is false: f^*_μ contains lower-degree terms that prevent proportionality to f_μ.

5. **Combinatorial constructions may be harder for LLMs than algebraic manipulations.** The official solution requires inventing a new dynamical system (Step 2 of the interpolation t-PushTASEP) and encoding its transitions via two-line queues. This is a creative combinatorial construction that may be inherently harder for LLMs to generate than algebraic manipulations of known identities.

6. **The "forthcoming paper" signal was misread.** BDW25 Remark 1.17's deferral to a forthcoming paper was a strong signal that the interpolation probabilistic interpretation requires new ideas beyond the ordinary t-PushTASEP. The `af` attempt interpreted this as "the transfer is nontrivial but doable"; the correct interpretation was "a new construction is needed."

7. **Nontriviality arguments transfer well.** The `af` attempt's sparsity argument for nontriviality (Node 1.7) is clean, correct, and would apply to the interpolation t-PushTASEP as well. This is a reusable technique.

10 Conclusion

The `af` automated proof attempt for Problem 3 correctly identified the answer (YES) and the right family of Markov chains (the t-PushTASEP family), but proposed the wrong specific chain (the ordinary t-PushTASEP instead of the interpolation t-PushTASEP) and attempted to bridge the gap with a ratio identity that is false. The official solution by Ben Dali and Williams constructs a genuinely new Markov chain — the interpolation t-PushTASEP with its novel Step 2 reinsertion phase — and proves stationarity directly via two-line queue combinatorics and the factorization of interpolation Macdonald polynomials.

The `af` attempt's failure aligns precisely with the second common LLM failure mode identified by the problem authors: replacing interpolation polynomials by homogeneous ones and solving the already-solved problem. The adversarial framework successfully caught many errors (97 challenges in Session 2) but could not catch the fundamental misidentification of the Markov chain. The system's own risk assessment correctly flagged Node 1.6 (the ratio identity) as the highest-risk node, demonstrating that the `af` framework has some capacity for self-diagnosis even when it cannot resolve the underlying issue.

References

[1] A. Ayyer, J. B. Martin, and L. K. Williams, *The inhomogeneous t-PushTASEP and Macdonald polynomials at $q = 1$*, Annales de l'Institut Henri Poincaré D (2025). arXiv:2403.10485.

[2] H. Ben Dali and L. K. Williams, *A combinatorial formula for interpolation Macdonald polynomials*, Preprint arXiv:2510.02587 (2025).

[3] S. Corteel, O. Mandelshtam, and L. K. Williams, *From multiline queues to Macdonald polynomials via the exclusion process*, Amer. J. Math. **144**(2) (2022), 395–436.

[4] M. Dołęga, *Strong factorization property of Macdonald polynomials and higher-order Macdonald's positivity conjecture*, J. Algebraic Combin. **46**(1) (2017), 135–163.

[5] P. Alexandersson and M. Sawhney, *Properties of non-symmetric Macdonald polynomials at $q = 1$ and $q = 0$*, Ann. Comb. **23**(2) (2019), 219–239.