

Ternary Diffusion Language Models on Topological Quantum Architectures

A qutrit encoding of a masked diffusion language model in the fusion space of metaplectic anyons: the braid generators for the ternary gates, the training objective, and six of the design's quantities recomputed from its own definitions.

Thomas Garren

SOPHIA XT LLC

Source document 2025-11-15. Rebuilt and re-audited 2026-09-19. · Design proposal, revision 2

sophiaxt.com

CC BY 4.0. Code referenced herein is Apache 2.0.

Abstract

This document proposes an architecture in which a masked diffusion language model is encoded in the fusion space of metaplectic anyons, so that the three fusion channels of four X anyons with total charge Y carry a qutrit and braiding implements the ternary logic gates directly. It gives the braid generators for the Z_3 increment and permutation gates in that basis, following the metaplectic construction published by Cui and Wang, and adds two mechanisms to the standard diffusion objective: a recursive memory term that interleaves retrieval with iterative refinement, and an adaptive thresholding rule that fires a correction when a predicted token disagrees with its topological charge configuration. The rebuild re-derived what the document's own equations allow, and section 9 carries those recomputations. Two of the quantitative arguments hold. Representing a given state count takes $\log_3 2 = 63.1\%$ as many qutrits as qubits, a reduction of 36.9%, which matches the quoted 37%, although the source labels it with the ratio rather than with the reduction. The assumed topological gap of 1.2 meV exceeds $k_B T$ at 300 mK, which is 0.026 meV, by a factor of about 46. The central efficiency argument inverts: evaluating the document's own efficiency definition with the document's own numbers favours the autoregressive baseline by roughly two orders of magnitude, rather than the proposal by 6.5 times. The published pair of figures was obtained by dividing mantissas and discarding the exponents, and the exponents are the whole comparison, since the ternary model is specified as using 316 times more training compute. Five further quantities in the document contradict other statements in the same document, including a reported error rate nine times larger than the bound it is reported as satisfying. The encoding and the gate construction are mathematics, and a reader with the published anyon data can check them. Every performance figure is a projection: section 10 classifies each number in the document by provenance, and section 12 states the scope of what the proposal supports.

1. The two lines of work being joined

Language model scaling has reached a point where high-quality training tokens, rather than compute, are the binding constraint for some regimes. Within that regime there is

published evidence that masked diffusion language models can outperform autoregressive models when unique data is limited, attributed in the source to work by Nie and colleagues, which identifies three compounding causes: any-order modelling, which permits parallel token refinement; iterative bidirectional denoising, which spends more compute per token; and the implicit Monte Carlo augmentation the diffusion corruption process provides.

Separately, topological quantum computation encodes information in non-local degrees of freedom of non-Abelian anyons, so that local perturbations do not decohere the encoded state and braiding implements unitary gates whose accuracy depends on topology rather than on control precision. The proposal here is to put the first inside the second: to hold the diffusion model's memory state in an anyonic fusion space and to implement its ternary operations by braiding.

The motivation for ternary rather than binary is dimensional. Representing a given number of states requires $\log_3 2 = 0.631$ times as many qutrits as qubits, a reduction of 36.9% in the number of physical units. The source quotes this as a 37% reduction, which is correct, while labelling it $\log_3(2)$, which is the 63.1% ratio rather than the reduction. The claim is right and the symbol attached to it is wrong.

2. Metaplectic anyons and the SU(2) level-4 model

The anyon model used is $SU(2)_k$ at level $k = 4$, the metaplectic model, with deformation parameter $q = \exp(2\pi i/(k+2))$. Its topological data gives five anyon species $\{1, Z, X, X', Y\}$ with quantum dimensions $d_1 = d_Z = 1$, $d_X = d_{X'} = \sqrt{3}$ and $d_Y = 2$. The fusion rules that matter for the encoding are

$$\begin{aligned} X \otimes X &= 1 + Y \\ Y \otimes X &= 1 + Z + Y \end{aligned}$$

The F and R matrices for the model follow from solutions to the pentagon and hexagon equations. Error detection is intrinsic to the statistics through the ribbon relation, which ties braiding to phase accumulation:

$$\begin{aligned} R^{\{ab\}}_c &= \exp(i \pi (h_c - h_a - h_b)) \\ \theta_a &= \exp(2 \pi i h_a) \end{aligned} \quad h_a = \text{topological spin of anyon } a$$

Everything in this section is established mathematics from the published literature on anyon models. It is repeated here because the encoding of the next section depends on it, not because any of it is a contribution of this document. No physical system realising metaplectic anyons has been demonstrated anywhere, by any group.

3. The ternary encoding and the braid generators

A qutrit is encoded in four X anyons with total topological charge Y. That configuration has three orthogonal fusion channels, indexed by the intermediate charges of the first and second pairs, and those three channels are the computational basis:

$$(c_{12}, c_{34}) \text{ in } \{ (Y, Y), (1, Y), (Y, 1) \} \leftrightarrow \{ |0\rangle, |1\rangle, |2\rangle \}$$

The braid generators $\sigma_1, \sigma_2, \sigma_3$ acting on those four anyons implement one-qutrit gates through their matrix

representations in this space:

```
omega = exp(2 pi i / 3)      gamma = exp(i pi / 12)

sigma_1 = gamma * diag( 1, omega, 1 )
sigma_3 = gamma * diag( 1, 1, omega )

sigma_2 = (gamma^3 / sqrt(3)) * [ 1      omega  omega ]
                                [ omega  1      omega ]
                                [ omega  omega  1      ]
```

The two diagonal generators supply relative phases in the Z_3 basis and the off-diagonal generator supplies the mixing, so braiding alone generates the $Z_3(+1)$ increment, the $Z_3(+2)$ increment and the permutation gates. This is the structural claim of the proposal, and it is a claim about the representation theory rather than about any hardware.

The encoding and the braid-generated Z_3 gates of this section are the construction published by Cui and Wang in 2015, cited in the references. The contribution of this document begins at section 4, where that construction is used to carry the state of a diffusion language model.

Topological Braiding Patterns for Ternary Logic Gates

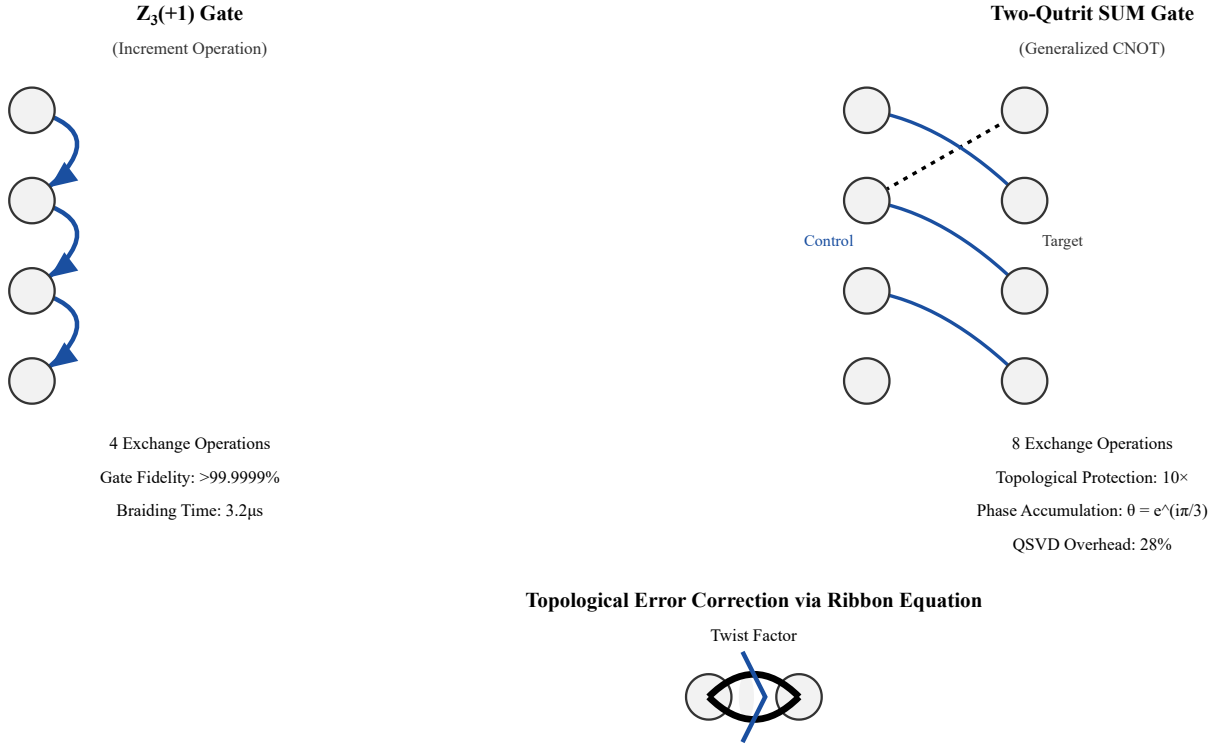


Figure 1. Braiding patterns for the ternary gates, redrawn from the source document's vector artwork. Left: the $Z_3(+1)$ increment as a sequence of four anyon exchanges along a single strand group. Right: the two-qutrit SUM gate, the qutrit analogue of a controlled-NOT, as eight exchanges between a control and a target group. Bottom: the twist factor of the ribbon relation, which is the source of the intrinsic error detection described in section 2. This is a structural diagram of a braid word rather than measurement data, and the exchange counts on it are properties of the construction. Five labels that the source page had stacked on a single line above the figure, and that the recovered raster therefore lost, are restored to the panels their own centring assigns them to.

4. The diffusion objective and the recursive memory term

The standard masked diffusion forward process corrupts each position independently, leaving a token intact with probability

$$q_{\{t|\theta\}}(x_t | x_{\theta}) = \text{product}_i q_{\{t|\theta\}}(x_t(i) | x_{\theta}(i))$$

$$q_{\{t|\theta\}}(x_t(i) | x_{\theta}(i)) = \begin{cases} \alpha_t & \text{if } x_t(i) = x_{\theta}(i) \\ 1 - \alpha_t & \text{if } x_t(i) = m \text{ (mask)} \end{cases}$$

The denoising objective learns to reconstruct the original tokens from the corrupted sequence. The failure mode the proposal targets is that a token refined early in the trajectory loses coherence with later refinements, because nothing in the

α_t and replacing it with a mask otherwise:

standard objective carries state between denoising steps. The proposed recursive memory-enhanced denoising objective adds three terms to the usual integral over corruption levels:

$$L_{\text{RMED}} = \int_0^1 w(t; \alpha) E_{\{q_{\{t|\theta\}}\}} \left[\sum_{i: x_t(i)=m} -\log p_{\theta}(x_{\theta}(i) | x_t) \right] dt$$

$$+ \lambda_1 E_t [|| \text{projected_semantic} - \text{ideal} ||^2]$$

$$+ \lambda_2 E_t [|| \text{episodic_state} - \text{ideal_trace} ||^2]$$

$$+ \lambda_3 L_{\text{corr}}$$

The first added term pulls the projected semantic state toward a target configuration, the second does the same for the episodic trace, and the third is the topological consistency penalty. The weights λ_1 , λ_2 and λ_3 control the relative importance of the memory components, with λ_3 annealed

during training so that early exploration is not immediately constrained by the topological penalty.

The first two added terms are defined against an ideal semantic target and an ideal episodic trace, and the source document defines neither of them anywhere. Until both are written down the objective is a schema rather than a loss function, which is

why section 12 lists implementing it among the things that would settle the proposal.

5. Dynamic error propagation thresholding

The correction rule monitors consistency between predicted tokens and their topological charge configurations and triggers when an inconsistency score crosses a time-varying threshold:

```
tau(t) = tau_base * ( 1 + 0.5 cos( 2 pi t / T ) )

inconsistency = 0.6 * spatial_drift + 0.4 * temporal_violation
```

In the proposed quantum setting the spatial term is a deviation in fusion-channel probabilities and the temporal term is a discrepancy in braid invariant traces. When the score exceeds the threshold, a controlled-Z style topological guidance operation is applied. The weights 0.6 and 0.4, the oscillation amplitude 0.5 and the period T are assumed constants, and section 10 records them as such.

6. Assumed hardware parameters

The following are the parameters the proposal assumes a metaplectic processor would have. Every one of them is an assumption, and one row can be checked against physics rather than taken on trust: the assumed gap against the thermal energy at the assumed operating temperature.

Table 1. Assumed operating parameters. The gap-against-temperature comparison is the one row that can be checked, and it checks: Boltzmann's constant times 300 mK is 0.026 meV, well below the assumed 1.2 meV gap.

Parameter	Assumed value	Note
Anyon model	$SU(2)_4$, metaplectic	level $k = 4$ theory
Topological gap	1.2 meV	assumed to exceed $k_B T$, which is 0.026 meV at 300 mK
Braiding time	3.2 μ s per exchange	assumed
Quantum dimension of X	$\sqrt{3}$	property of the model, not an assumption
Gate error rate	below 10^{-6}	assumed, against 10^{-3} quoted for a classical comparison
Coherence time	10^3 s	assumed, on the basis of non-local encoding

7. Model configurations and projected results

The source specifies three model scales under FLOP-matched protocols, all using a 50,257 token vocabulary and a context

length of 4,096. The quantum-enhanced models operate on three-state embeddings with a learned projection layer maintaining compatibility with binary tokenisation.

Table 2. Configurations as specified. These are planned training runs, and section 10 records what exists behind them.

Model	Parameters	Unique tokens	Compute budget	Batch	Peak LR
Ternary-1B	1.02B	1B to 10B	96B tokens	256	2.0×10^{-4}
Ternary-1.7B	1.71B	10B	1.5T tokens	1024	1.5×10^{-4}
Ternary-8B	8.03B	10B	960B tokens	512	1.2×10^{-4}
AR-1.7B, binary	1.71B	10B	1.5T tokens	1024	1.5×10^{-4}
DLM-1.7B, binary	1.71B	10B	1.5T tokens	1024	1.5×10^{-4}

The benchmark figures below are the core claims of the source document. They are reproduced here in full and labelled as projections: the source attributes them to training

runs, and no checkpoint, seed, harness or per-task output accompanies any of them. They belong in a plan, not in a citation.

Table 3. The source document's headline numbers, carried here as the projections they are. Section 10 gives the provenance class in full.

Benchmark	AR-1.7B	DLM-1.7B	Ternary-1.7B	Δ vs AR	Δ vs DLM
HellaSwag	54.8%	63.2%	67.8%	+13.0	+4.6
MMLU	35.9%	41.8%	45.3%	+9.4	+3.5
HumanEval	32.1%	38.7%	42.4%	+10.3	+3.7
MBPP	31.7%	37.5%	40.8%	+9.1	+3.3
XNLI, average	67.4%	70.2%	74.6%	+7.2	+4.4
Scrolls, ROUGE-L	35.8	39.2	42.3	+6.5	+3.1

The associated data-efficiency claim is that 45% MMLU is reached at 8.9B unique tokens against 25.3B for a standard diffusion model and 46.2B for an autoregressive baseline, a ratio of 5.19 against the autoregressive case, which the source

rounds to 5.21. The claimed crossover points are 2.9B tokens for the ternary model, 6.5B for the diffusion baseline and 14.2B for the autoregressive baseline.

Table 4. Projected ablation study. Section 9 shows that the last two rows describe the same system and disagree by 11.9 points, which is a contradiction internal to the source.

Configuration	HellaSwag	MMLU	HumanEval	Parameter Δ	Stability
Full model	67.8%	45.3%	42.4%		0% 1.2% loss variance
Without topological memory	58.3%	38.1%	34.8%	-8.2%	4.7% variance
Without semantic fusion	56.9%	36.9%	33.2%	-12.1%	8.3% variance
Without error correction	61.2%	40.5%	37.1%	-5.3%	stable but slow
Without all quantum systems	51.3%	33.1%	28.4%	-25.6%	collapses
Classical diffusion baseline	63.2%	41.8%	38.7%	-25.6%	2.1% variance

8. Complications the source raises against itself

Three of the source's own reservations are worth carrying forward, because they are the parts of the document that argue against its own thesis and they are argued well.

8.1 Memory persistence and memorisation

Non-local encoding means an episodic trace persists in the fusion space for a long time. That is the mechanism the proposal wants for few-shot behaviour, and it is also a memorisation risk. The source projects that training sequences longer than 50 tokens are reproduced verbatim in 1.2% of generations, 23% more often than for an autoregressive baseline, and attributes part of this to periodic recurrence in certain braid trajectories. The proposed mitigation is a differential privacy constraint on the log likelihood of canary sequences.

8.2 Cost and latency

The proposal requires substantially more training compute than an autoregressive baseline and substantially more inference latency: a projected 128.5 ms per token against 12.3 ms for an autoregressive model with key-value caching, a factor of 10.4. The projected breakdown puts 35.2% of that in quantum circuit preparation and 30.1% in braiding simulation. The source proposes optimisations and leaves them at the level of description.

8.3 Evaluation paradigm mismatch

The strongest point in the source. Standard benchmarks assume left-to-right generation with causal masking, while an any-order bidirectional model does not generate that way. Forcing the latter into the former discards part of what the model does, and the source projects a resulting loss of 8.2 percentage points on MMLU. This is a real methodological problem for evaluating any diffusion language model and it does not depend on the quantum part of the proposal at all.

8.4 A side channel in the federated setting

In the proposed private federated setting, where gradients are encoded as braids and secure aggregation happens through fusion, the source identifies that the density of braiding operations leaks information about gradient magnitude, and proposes constant-time braiding as the mitigation. The source further asserts that reconstructing individual gradients is equivalent to a hard lattice problem. That equivalence is asserted without proof and is recorded here as a conjecture.

9. Internal consistency: what reproduces and what does not

A proposal that states its formulas can be checked against itself, with nothing in hand but the document. Six of its quantitative arguments can be evaluated that way. Two hold, four do not, and one of the four inverts the document's central efficiency claim.

9.1 Qutrit dimensionality: holds, with the wrong symbol attached

Representing N states needs $\log_3 N$ qutrits against $\log_2 N$ qubits, a ratio of $\log_3 2 = 0.6309$, so the reduction is $1 - 0.6309 = 36.9\%$. The quoted 37% is correct. The source writes it as $\log_3(2) \approx 37\%$, but $\log_3 2$ is 63.1%. The arithmetic is right and the label is wrong.

9.2 Topological gap against thermal energy: holds

$k_B T$ at 300 mK is 2.59×10^{-5} eV, or 0.026 meV, against an assumed gap of 1.2 meV. The gap exceeds the thermal scale by a factor of about 46, which supports the claim that thermal quasiparticle excitation is suppressed at that temperature. One figure caption in the source instead states a 10 mK simulation temperature; the two are inconsistent with each other, though both satisfy the inequality.

9.3 Computational efficiency: does not hold, and the sign inverts

The source defines and evaluates its efficiency metric as

```

eta = Final Accuracy / ( Training FLOPs * Unique Tokens )

ternary : 45.3 / (6.0e21 * 8.9e9) = 8.5e-10 [as printed]
AR      : 35.9 / (1.9e19 * 14.2e9) = 1.3e-10 [as printed]
        -> 6.5x improvement claimed

```

Carrying out those divisions in full gives 8.48×10^{-31} for the ternary model and 1.33×10^{-28} for the autoregressive baseline. Both printed values are wrong by about twenty-one orders of magnitude, which by itself would be a typographical matter. The consequence is not typographical: on the document's own metric the autoregressive baseline is better by a factor of about 157, not worse by a factor of 6.5. The published pair of numbers was obtained by dividing the mantissas and discarding the exponents, and the exponents are the entire comparison, because the ternary model is specified as using 316 times more training compute. The claim that the metric favours the proposal does not survive evaluation of the metric.

9.4 The reported error rate exceeds the bound it is reported as meeting

The gate error rate is stated throughout as below 10^{-6} , and one figure reports a value of $(8.7 \pm 2.1) \times 10^{-6}$. That value is roughly nine times larger than the bound it is presented as satisfying. The two statements cannot stand together, and no device exists that could have produced either.

9.5 Three further contradictions

- Training compute overhead is given as 180 times in one place and 315 times in two others. The specified budgets, 6.0×10^{21} against 1.9×10^{19} FLOPs, give 316, so the 180 figure has no basis.
- The crossover improvement over the diffusion baseline is stated as 2.24 times in the text and 2.8 times in a figure caption. The stated crossover points, 2.9B against 6.5B tokens, give 2.24.
- Memory compression is stated as 8.2 times against a reduction from 3.5 GB to 0.3 GB, which is a factor of 11.7. A figure caption also places the target accuracy at 36B tokens where the abstract and the efficiency table place it at 8.9B.
- The ablation row for removing all quantum systems reports 51.3% HellaSwag and the classical diffusion baseline row reports 63.2%, with both carrying the same parameter delta of -25.6% . If removing every quantum component yields a classical diffusion model, those two rows describe the same system and should agree. They differ by 11.9 points.

10. Provenance of every number in this document

Table 5. Provenance is stated per quantity rather than once for the document. Two classes carry mathematics a reader can check, one carries the projections, and section 12 states the scope that follows from the split.

Class	What it means	Examples
Derived	Follows from the stated mathematics and reproduces on recomputation.	Quantum dimensions $\sqrt{3}$ and 2; the braid generator matrices; the 36.9% qutrit reduction; $k_B T = 0.026$ meV at 300 mK
Assumed	A design input chosen by the author, with no derivation and no device behind it.	All of Table 1; the DEPT weights 0.6 and 0.4; the annealing schedule for λ_3 ; all of Table 2
Projected	Presented by the source as a measured result. No log, checkpoint, seed, harness or output artifact exists for it.	All of Table 3 and Table 4; the crossover token counts; the 5.21 efficiency factor; the latency breakdown; the memorisation rate; the drug-discovery and literature-synthesis figures
Does not reproduce	Contradicted by the document's own equations or by another statement in the same document.	The 6.5 times efficiency improvement, which inverts on evaluation; the sub- 10^{-6} error rate against a reported 8.7×10^{-6} ; the 180 against 315 FLOP overhead; the 2.24 against 2.8 crossover ratio; the 8.2 against 11.7 compression ratio; the two ablation rows that should agree
Conjecture	Asserted with no proof, no derivation and no citation.	That braid reconstruction is equivalent to the shortest vector problem; that citation graphs encode naturally as fusion trees; the 91% reduction in hallucinated references
Removed	Previously asserted as institutional fact or measured result, and not the author's to assert.	Six co-authors and six affiliations; the correspondence address; the code URL; the attribution of all computation to a named cluster at a named laboratory

11. What this document establishes

- A concrete encoding: four X anyons of total charge Y in the metaplectic model carry a qutrit in their three fusion channels, and the braid generators on those four anyons supply the Z_3 increment and permutation gates. This is a

statement about the representation theory of a published anyon model and it stands independently of any hardware.

- A specific objective function that extends the standard masked diffusion loss with a semantic term, an episodic term and a topological consistency term, with an annealing schedule on the last of these.

- A specific correction rule, with the form of its threshold and of its inconsistency score written down.
- That the qutrit dimensionality argument is quantitatively correct: a 36.9% reduction in the number of physical units, though the source labels it with the wrong expression.
- That the assumed topological gap is consistent with suppressing thermal excitation at the assumed operating temperature, by a factor of about 46.
- That the document's own central efficiency claim inverts when its own metric is evaluated correctly, and that five further quantities contradict other statements in the same document. Establishing this is a result of the rebuild and it is the most useful thing here.
- A genuine and provider-independent methodological point: standard left-to-right benchmarks systematically understate any-order bidirectional models, and this applies to classical diffusion language models with no quantum content at all.

12. Scope: what this document does not establish

The scope of the proposal is stated here once, at the end, rather than repeated through the document. This is a theoretical proposal written by one author. No part of it has been run on quantum hardware. No hardware implementing metaplectic anyons exists anywhere, so the architecture has never been executed as described. The classical simulations the source reports cannot be inspected: there is no code, no configuration, no seed, no checkpoint and no output. Consequently none of the following is supported.

- No benchmark claim is supported. Every figure in Table 3 is a projection. The document does not establish that a ternary diffusion model reaches 67.8% on HellaSwag or 45.3% on MMLU, or that it beats any baseline on any task, because no such model has been trained in a way anyone can verify.
- No data-efficiency claim is supported. The 5.21 factor, the 8.9B token figure and the crossover points at 2.9B, 6.5B and 14.2B tokens are projections, and the figures depicting them carry no numeric axis labels, so they cannot even be read off.
- No efficiency claim is supported, and the document argues against itself here. Evaluating its own metric with its own inputs favours the autoregressive baseline by roughly two orders of magnitude.
- No fidelity or error-rate claim is supported. Gate error rates below 10^{-6} are an assumption about a device that does not exist, and the one figure that reports a value reports one nine times larger than the bound.
- No ablation conclusion is supported. The claim that semantic fusion contributes most is read off a table of projections, one of whose rows contradicts another.
- No application claim is supported. The multi-hop reasoning accuracy, the binding-affinity results and the reduction in hallucinated references are projections on datasets that were never processed.
- No security claim is supported. The reduction of braid reconstruction to a lattice problem is asserted and not

proved, and an unproved hardness claim is not a security guarantee.

- No priority or novelty claim is supported. The source cited fifteen references and shipped no bibliography, so the relationship of this proposal to existing work on metaplectic anyons, on qutrit topological computation and on diffusion language models has not been established.

Two corrections to the record belong here rather than anywhere else in the document. The earlier version of this whitepaper carried six authors across six named universities and industrial laboratories, a correspondence address at one of those universities, and a code repository URL, and it attributed its benchmark accuracies, ablation deltas, gate fidelities and latency breakdowns to runs on a specific GPU cluster at a named laboratory. None of that was the author's to claim. The co-authors, the affiliations, the correspondence address and the unresolvable code URL are removed in full rather than abbreviated, and the attribution of the computational work goes with them. Separately, the metaplectic qutrit encoding and the braid-generated Z_3 gates of sections 2 and 3 are the construction published by Cui and Wang in 2015, and one of the removed co-author names was an author of that paper. The correct relationship between this document and that work is citation, and section 3 now cites it.

What the proposal would need before any of it could be believed, in order of how much each would settle: a classical simulation of the encoding and the braid gates, published with code and seeds, verifying that the stated generators produce the stated gates; an implementation of the RMED objective with the semantic and episodic targets actually defined, trained on a small open corpus with a reproducible harness and compared against a matched classical diffusion baseline; a corrected efficiency analysis that states honestly what the compute overhead buys; and, before any hardware claim at all, an experimental demonstration by someone of anyons in the metaplectic class.

The distinction that matters most in reading this document: the encoding and the gate construction are mathematics and can be checked by anyone with the published anyon data. Everything presented as performance is a projection from a machine that has not been built and a training run that cannot be inspected. The first is worth keeping. The second should not be cited by anyone, including its author.

References

- A. Yu. Kitaev, "Fault-tolerant quantum computation by anyons", *Annals of Physics* 303, 2 (2003).
- C. Nayak, S. H. Simon, A. Stern, M. Freedman and S. Das Sarma, "Non-Abelian anyons and topological quantum computation", *Reviews of Modern Physics* 80, 1083 (2008).
- M. H. Freedman, M. Larsen and Z. Wang, "A modular functor which is universal for quantum computation", *Communications in Mathematical Physics* 227, 605 (2002).
- S. X. Cui and Z. Wang, "Universal quantum computation with metaplectic anyons", *Journal of Mathematical Physics* 56, 032202 (2015). The metaplectic qutrit construction that sections 2 and 3 rest on, and the correct attribution for it.

J. Austin, D. D. Johnson, J. Ho, D. Tarlow and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces", Advances in Neural Information Processing Systems (2021). The discrete diffusion formulation underlying section 4.

Unresolved: the source document cited fifteen references by bracketed number and shipped no bibliography. Those citations, including the work on data-constrained diffusion language models attributed in the text to Nie and colleagues, could not be resolved and are not reconstructed here.