

Artificial Epanorthosis

Why large language models overuse a classical rhetorical figure, and how to mitigate it

Federico Boggia

July 2026

Abstract. A rhetorical figure catalogued by Cicero and Quintilian two thousand years ago reappears, in massive and systematic form, in text produced by large language models. That figure is *epanorthosis*: the act of correcting or amplifying what one has just said, as in the specimen «It is not a course. It is a journey of transformation». This essay argues that the overuse is a trained disposition. Its primary drivers are a training distribution rich in promotional prose and a preference-tuning stage (RLHF) that rewards confident, emphatic phrasing, with recent evidence linking such reward tuning to lexical overrepresentation; the left-to-right nature of generation is a contributing amplifier rather than the root cause. Drawing on empirical work showing that models fail to match human rhetorical styles, and on the classical distinction, due to Fontanier, between correction as a figure of style and epanorthosis as a figure of *thought*, the essay characterises the phenomenon and sets out a *research programme* (a protocol plus a first pilot measurement) to measure it against genre-specific human baselines through an *Epanorthosis Index* (density relative to the human rate for each genre). In a first measurement, instruction-tuned models mis-calibrate the figure by register in both directions: they overshoot human rates in the oratorical register (about twofold, and near threefold in Italian, significant across model tiers) and undershoot in informal question-and-answer writing, while sitting at parity in argument, journalism, and encyclopedic prose. Its constructive contributions are three: a survey of mitigation techniques centred on lightweight parameter-efficient fine-tuning (LoRA adapters and preference optimisation), complemented by decoding-time control and activation steering; a demonstration, in Italian, that a one-line anti-epanorthosis instruction reduces the figure by roughly half to two-thirds and that a supervised-fine-tuning LoRA adapter removes it almost entirely at pilot scale (then to be dialled back toward the human rate); and the argument that the right target is *calibration to the human rate for each genre*, not elimination, since correction is often a form of clarity and appropriateness is precisely what a statistical system finds hardest to learn. It closes on the stakes: the real risk is that we begin to write like the machines.

Keywords: epanorthosis, rhetoric, large language models, autoregressive generation, RLHF, LoRA, controllable text generation, AI-text detection

1. Introduction

Read the next answer a chatbot gives you, or the next motivational post that scrolls through a professional feed, and a structure begins to surface that, once seen, is difficult to unsee:

«I am not a consultant. I am an architect of change.»

«We do not sell products. We build experiences.»

«This is not a course. It is a journey of transformation.»

The model asserts something and, in the same breath, negates it in order to replace it with something larger, stronger, more solemn. The tic is so pervasive that it has become a kind of signature of automatically generated writing. What is rarely said is that the tic has a name: a Greek name, more than two thousand years old, that Cicero and Quintilian used with full awareness. That name is *epanorthosis*.

The observation has empirical support. A recent large study comparing parallel corpora of human- and model-written text found that language models systematically diverge from human writers precisely at the level of grammatical and rhetorical style, that the divergence is *larger* for instruction-tuned models than for base models, and that it does not vanish as models scale up.^[1] The rhetorical fingerprint, in other words, is being reinforced by the very training stages meant to make models more useful, and it persists as models grow.

This essay pursues a question that is simple to pose and surprisingly hard to close. When a machine that knows nothing of rhetoric spontaneously reproduces a classical rhetorical figure, what is actually happening? Is it a coincidence, a meaningless statistical artefact? Or does the convergence between ancient rhetoric and the statistics of neural networks tell us something about language itself? The argument proceeds in three movements. Sections 2–3 characterise the figure and trace where the overuse comes from. Sections 4–6 turn the phenomenon into an object of measurement and ask whether it is a figure of rhetoric or an artefact of prediction. Section 7, the core contribution, surveys concrete future directions for mitigating the overuse, organised around LoRA adapters and other lightweight techniques. A note on method: the paper tries to practise what it argues, keeping its own prose free of gratuitous epanorthosis, and the corrective turns it does use are the legitimate, content-bearing kind it defends in Section 7.

2. Epanorthosis in classical rhetoric

The word derives from the Greek ἐπανόρθωσις (*epanorthosis*): *epi*, “upon, in addition”, and *anorthosis*, “a straightening”, meaning literally “a further straightening”. The speaker returns to what has just been said in order to correct it. The correction works as a deliberate rhetorical move. One goes back to reinforce, to soften, or to make more precise. In Latin the figure travels under the name *correctio*, and it appears in three functional variants. There is the *emphatic* correction, an upgrade in which a term is replaced by a stronger one (“he is good, indeed excellent”); the *attenuative* correction, a downgrade (“it is a disaster, well, a setback”); and the properly *corrective* variant, which changes the substance (“he was running, or rather, walking briskly”). It is one of the most common figures of everyday speech: every “I mean”, “rather”, “or better” is an act of epanorthosis performed without noticing.

The rhetorical tradition understood its power early. Aristotle, in the *Rhetoric*, connects the appearance of self-correction to a persuasive mechanism: an orator who corrects himself seems sincere and reflective, someone thinking in front of the audience rather than reciting a script.^[3] Quintilian, in the *Institutio Oratoria* (c. 95 CE), is more explicit: *correctio* serves to move the listener (*movere*), and it has three offices: to emphasise, by implying the first formulation was insufficient; to give the impression of genuine, unrehearsed thought; and to display a willingness to revise.^[4] A clear Ciceronian instance is in the First Catilinarian, where a flat assertion is corrected upward in the same breath:^[5]

Hic tamen vivit. Vivit? immo vero etiam in senatum venit.

The particle *immo* (“nay, rather”) is the Latin correction-marker par excellence: the first claim is retracted as an understatement and replaced by a stronger one. Epanorthosis proper should be kept apart from mere accumulation. Cicero’s more famous “Abiit, excessit, evasit, erupit” (*In Catilinam* II.1), four verbs that intensify without retracting one another, is a *climax* (gradatio), a different figure. The most devastating vernacular case of genuine corrective epanorthosis may be Dante’s, at the close of the Ugolino canto: “Poscia, più che ’l dolor, poté il digiuno”.^[6] There the final correction reverses the meaning of the entire scene, revealing that hunger, more than grief, killed Ugolino.

For the present argument the decisive author is a nineteenth-century French rhetorician, Pierre Fontanier, in *Les Figures du discours* (1827).^[2] Fontanier draws a distinction that seems pedantic yet turns out to be crucial: he separates *correction*, which he treats as a figure of style, from *épanorthose* (or *rétroaction*), which he places among the **figures of thought**. In his own words, here in English translation: “Rétroaction, otherwise Epanorthosis, which must not be confused with Correction, a figure of style, consists in returning to what one has said, either to strengthen it, or to soften it, or even to retract it entirely.” The consequence is far-reaching. If epanorthosis is a figure of *thought* rather than of style, then it is a general discursive operation. It can be observed wherever there is language, including the output of a statistical model, and it extends beyond fine writing or oratory; crucially, it can be studied independently of the intention of whoever produces it. This is the bridge that lets us pass from Cicero to a chatbot without cheating. Standard reference works confirm the classification and taxonomy: Lausberg’s *Handbook of Literary Rhetoric* catalogues *correctio* among the figures that stage a revision of the utterance.^[7]

3. Where the overuse comes from

The overuse is best understood as a trained disposition, and its main sources lie in how models are built rather than in how they emit tokens. The first is the **training distribution**: vast quantities of web text, a substantial fraction of which is copywriting, motivational posts, and landing pages, all genres in which emphatic epanorthosis correlates with engagement. This is a plausible pressure rather than a measured one, and quantifying it corpus-side is itself part of the research programme of Section 4. The model learns that the structure works. The second source, and probably the decisive one, is **reinforcement learning from human feedback** (RLHF): the alignment stage in which annotators rank candidate responses.^[10] Raters tend to prefer answers that sound confident, incisive, and committed, and few things sound as confident as a well-placed upward correction. There is empirical support for this link: investigating why words such as “delve” and “intricate” became overrepresented in model output, Juzek and Ward found no evidence that architecture, decoding algorithm, or the raw training data were responsible, and their comparative testing pointed instead to *RLHF*.^[11] If preference tuning can inflate individual words, it can inflate a whole construction.

Generation itself adds a weaker, secondary pressure, and it is worth being precise about how much. A transformer language model produces text **autoregressively**, one token at a time, left to right, and a committed token cannot be retracted.^[8] It is tempting to conclude that the model therefore repairs by correcting forward, “good, indeed excellent”, which is the surface shape of epanorthosis. That inference is too strong: a well-trained model usually emits the intended token on the first attempt, so forced forward-correction is

the exception, not a structural necessity. What the left-to-right constraint does contribute is subtler. Unlike a human writer, who drafts, deletes, and reveals only a clean result, the model cannot revise before publishing, so whatever emphatic-correction disposition it has acquired from training surfaces in the open, unedited, instead of being smoothed away. Decoding plays a similar amplifying role: truncated sampling such as nucleus sampling favours fluent, template-like continuations,^[9] and the “Not X. Y” frame is one such high-probability path.

The surface forms are a small, recognisable family. The commonest is “Not X. Y”: negate a term and replace it with a grander one. (The specimens that open this essay are the fastest proof that the pattern can be recognised.) The variants are familiar once named: “Not only X, but Y”; “X, or rather, Y”; “It is not X, it is Y”; “X? No. Y”. A useful field heuristic: three or more such frames in a short passage is a strong indicator of generated text, though never a conclusive one.

4. Measuring the phenomenon

A signature one can only feel is of limited scientific use. Can the overuse be measured systematically, and compared against a human baseline? Recent work suggests it can, and that the effect is real. The excess-vocabulary method of Kobak and colleagues, applied to more than fourteen million biomedical abstracts, quantifies post-2022 shifts in word frequency attributable to model-assisted writing and estimates that at least ten percent of recent abstracts were processed with a language model.^[12] Liang and colleagues, using a maximum-likelihood estimator over expert- and model-written reference texts, found that between roughly six and seventeen percent of sentences in peer reviews at major AI conferences after the release of ChatGPT could have been substantially model-modified.^[13] And, as already noted, Reinhart and colleagues moved the analysis from single words to grammatical and rhetorical features (Biber’s feature set) and showed that the stylistic gap is structural, larger for instruction-tuned models, and persistent across scale.^[1] These results establish that rhetorical divergence between human and model text is measurable; epanorthosis intensity is one natural coordinate within that space.

Two families of detector are relevant. The first exploits token-level statistics: GLTR visualises how “expected” each token is under a reference model,^[14] and DetectGPT tests whether a passage sits at a local maximum of the model’s log-probability, a curvature signature of machine text.^[15] Comparison corpora such as HC3 supply paired human and model answers for training and evaluation.^[16] These methods target *authorship* rather than a specific figure; epanorthosis is one feature among many that they implicitly exploit. The second family targets rhetorical *constructions* directly. Dubremetz and Nivre showed that repetition-based figures (chiasmus, epanaphora, epiphora) can be detected with supervised models that separate rhetorically effective repetitions from accidental ones,^[18] and Bothwell and colleagues formalised *rhetorical parallelism detection* as a task with datasets, metrics, and baselines.^[19] Epanorthosis works by substitution where those figures work by repetition, yet it too is a construction rather than a bag of words, so its detection belongs to this second, construction-level tradition.

A methodological caveat frames what follows. Fontanier’s licence to study epanorthosis regardless of intention comes from treating it as a figure of *thought*; yet any lexical or construction-level detector measures its *stylistic* manifestation, the surface “Not X. Y”, which is the figure of style Fontanier keeps separate. The protocol below therefore measures a proxy: the observable trace of a discursive operation, not the

operation itself. On this footing one can sketch a concrete measurement protocol. First, build a **human baseline corpus** stratified by genre: spontaneous speech (for Italian, the KIParla corpus of transcribed spoken interaction offers a genre where discourse markers are frequent and the figure often serves merely as a filler),^[17] academic prose, twentieth-century narrative, journalism, and social writing. Draw it, wherever possible, from material predating the diffusion of generative models, so that the human rate is measured before contamination. Second, run a two-stage detector: a high-recall lexical pass over corrective markers (*but, rather, actually, I mean*, the “not...but” frame), followed by a construction-level classifier that filters the many false positives; words like “but” and “I mean” recur constantly without any corrective function, especially in speech, where they serve largely as fillers. A first-pass implementation of exactly this kind recovers more than half of hand-annotated instances yet generates a large number of false positives, with precision collapsing in speech and rising in argumentative prose, where “not X, but Y” does the work it promises. The lesson is methodological: at the tail, the phenomenon demands a human judgement of function that no lexical trigger supplies on its own. The measurement target, moreover, is frequency *conditioned on genre and communicative purpose* rather than raw frequency; this point returns, decisively, in Section 7.

4.1 A preliminary measurement

To move from protocol to a first result, I ran an English-language study with the detector above (distinct from the Italian corpus just described). The human baseline uses public-domain and openly licensed human texts, several works per genre, with pre-2022 filtering where feasible: encyclopedic prose (Wikipedia), journalism (Wikinews), academic abstracts (arXiv), narrative fiction (Conan Doyle), informal question-and-answer writing (high-voted pre-2022 Stack Exchange answers), argument (Mill, *On Liberty*), and oratory (inaugural and commemorative addresses, among them Lincoln, Jefferson, and T. Roosevelt). The model sample comprises short texts from neutral, genre-specific prompts that never mention correction or any rhetorical figure, generated across three instruction-tuned models (Claude Haiku, Sonnet, and Opus) and eight genres (seven with a human baseline, plus a model-only promotional register). Every text, human and model, is cut into windows of a few hundred words; density counts the emphatic-correction family (“not X, but Y”, “not only X but Y”, and the sentence-split “Not X. Y”) per ten thousand words, and significance is a label-permutation test on the per-window densities.

Table 2. Emphatic-epanorthosis density (per 10,000 words), human vs. instruction-tuned models, over few-hundred-word windows (n = windows). p from a 5,000-iteration permutation test; bold marks $p \leq 0.05$.

Genre	Human (n)	LLM (n)	EI (LLM/human)	p
Encyclopedic	1.2 (21)	1.4 (18)	1.2×	1.00
Journalism	2.4 (128)	2.0 (28)	0.9×	0.89
Academic abstract	3.6 (47)	7.8 (18)	2.2×	0.28
Narrative fiction	7.5 (84)	4.1 (18)	0.5×	0.38
Informal Q&A	8.2 (75)	1.3 (21)	0.2×	0.05
Argumentative essay	17.2 (71)	22.0 (18)	1.3×	0.37
Oratory	14.9 (19)	33.5 (17)	2.2×	0.03
Promotional	n/a	26.4 (18)	model-only	n/a

Two effects reach significance, and they point in opposite directions. In the *oratorical* register the models *overshoot*, producing emphatic self-correction at roughly twice the human rate (33.5 against 14.9 per ten thousand words, $p = 0.03$), consistently across model tiers (Haiku, Sonnet, and Opus at 15, 37, and 46). In informal *question-and-answer* writing they *undershoot*, falling to about a quarter of the human rate (1.3 against 8.2, $p = 0.05$), where human writers hedge and correct themselves freely and the models stay flat. The remaining registers show no significant gap: argument, encyclopedic prose, journalism, and narrative sit at parity, and academic abstracts trend toward model overshoot without reaching significance.

The picture is one of *mis-calibration by register in both directions*: the models pour the figure on where it flatters a persuasive audience and withhold it where a peer would qualify a claim. A specificity check supports reading this as targeted rather than as general verbosity: across genres the models use *fewer* neutral connectives than humans (in oratory, 1.5 against 7.5 per ten thousand words), so the oratorical spike is not mere marker-heaviness. Narrative makes the same point from the other side, since against a proper multi-work baseline the human rate is itself high (7.5) and the models fall below it.

These ratios can be read as a single metric. Define the **Epanorthosis Index** of a text or corpus in genre g as its emphatic-epanorthosis density divided by the human baseline for that genre, $EI_g = D_g / H_g$: an index of 1 is human-like calibration, above 1 is overshoot, below 1 undershoot. The “LLM / human” column of Table 2 is exactly this index, and its spread is the point: $EI \approx 2.2$ in oratory and ≈ 0.2 in informal question-and-answer, a two-sided mis-calibration that a single raw-frequency number would hide. A corpus-level *calibration error*, the mean of $|\log EI_g|$ across genres, condenses this into one number for how far a model sits from human calibration in either direction (0 = perfectly calibrated).

Caveats bound these numbers. The corpus is modest and English-only; the human argumentative, oratorical, and narrative samples are nineteenth-century, so register is matched but era is not (only the encyclopedic material is contemporaneous); the detector captures the “not X, but/. Y” family, a central but partial slice of epanorthosis; and promotional text has no public-domain human baseline. The result is a first signal rather than a settled measurement. It points, however, the same way as this paper’s argument: the models’ failure is one of appropriateness rather than raw frequency, concentrated in the register where the fig-

ure most flatters, which is exactly what Section 7.7 proposes to target. The corpus and detector script are available from the author for replication.

5. When self-correction does real work

So far epanorthosis has figured as a stylistic tic. There is also a regime in which the model's self-correction stops being decoration and becomes function. Since 2022 we have known that models answer far better when asked to reason step by step (the chain-of-thought technique),^[20] and that a single instruction such as "let's think step by step" elicits the behaviour zero-shot.^[21] In the intermediate steps a fully-fledged corrective epanorthosis appears: "Wait, that's not right, let me recompute"; "Actually, let me reconsider." Iterative refinement methods make the move explicit, feeding a model its own draft and self-critique.^[22]

The obvious question is whether, when a model writes "wait, let me reconsider", it is really reconsidering. The honest answer is uncomfortable. There is no internal module that fires on "wait", no representation of "changing one's mind". Yet the token sequence "wait, let me reconsider" shifts the context, and thereby shifts the probabilities of subsequent tokens toward alternative lines of reasoning. Textual self-correction can therefore *work* even without a mind behind it: the rhetorical form, emptied of mental content, still produces concrete effects. The claim must be qualified, however: Huang and colleagues showed that models often *cannot* reliably self-correct reasoning without an external signal, and may even degrade their answers when asked to revise blindly.^[23] Self-correction is a genuine capability only when the correction is grounded in something concrete, such as a tool result, a verifier, or a retrieved fact, rather than in the mere performance of reconsidering.

There is a historical inversion here worth savouring. Classical rhetoric taught orators to *simulate* self-correction in order to seem more credible; modern engineering teaches machines to *practise* self-correction in order to be more accurate. Same form, opposite direction. Anthropic's Constitutional AI, in which a model drafts, critiques its draft against a set of written principles, and rewrites, is exactly this: engineered epanorthosis, performed behind the scenes before we ever see the answer.^[24] The figure that Cicero staged for the Senate has become an optimisation loop.

6. Figure of rhetoric, or artefact of prediction?

When a human says "good, indeed excellent", a judgement lies behind the correction; when a model produces the same structure, behind it there is only a frequency in the data. This is Searle's Chinese Room applied to rhetoric:^[25] flawless figures with, apparently, no intentional persuasion behind them. The functionalist reply is that if rhetoric is defined by its effects on a listener rather than by intention, a model's epanorthosis persuades and emphasises exactly as Cicero's does. Both positions are defensible, and the honest verdict is that the categories built for intentional human discourse do not yet settle the case. What matters for the rest of this essay is narrower and practical: whatever its ontological status, the figure is produced in excess, that excess is measurable, and where it is gratuitous it can be turned down. If we wished to do so, could we, and how?

7. Future directions for mitigation

This section is the paper’s constructive core. Before surveying techniques, one framing decision must be made explicit, because it constrains every method that follows. The objective is to **calibrate** epanorthosis to context. Correction is often a virtue: in argumentative prose it carries clarity and intellectual honesty, and forcing it to zero would flatten the very texts we value. What matters is the figure’s *appropriateness to context*: an emphatic upgrade that suits a keynote reads as grating in a discharge summary. The pilot of Section 4 makes this concrete: the models overshoot specifically in the persuasive and oratorical register while matching humans in argument, so mitigation should be register-aware, turning the figure down where it is gratuitous and leaving it where it does argumentative work. The engineering target, therefore, is a controllable reduction of gratuitous, upward-emphatic epanorthosis at fixed semantic content, ideally exposed as a dial. The survey below moves from the most invasive interventions (retraining weights) to the least (prompting), noting for each its cost, its controllability, and its principal risk. Table 1 summarises the comparison.

7.1 Data-side interventions

The cheapest place to act is the data, before any gradient is computed. Two moves are available. The first is *counterfactual data augmentation*: mine a corpus for emphatic “Not X. Y” constructions and pair each with a de-emphasised paraphrase that preserves content (“I am not a consultant, I am an architect of change” → “I work as a consultant, with a focus on organisational change”). These pairs become supervision for later stages. The second is *instruction-data curation*: because the divergence is amplified by instruction tuning,^[1] pruning promotional and template-heavy examples from the instruction mixture attacks the problem at its documented source. Data-side methods are attractive because they are non-invasive and compose with everything downstream; their limitation is that they act only at the next training run and offer no inference-time control.

7.2 A LoRA adapter for rhetorical style

The most promising route, and the one this essay emphasises, is a lightweight fine-tune with **Low-Rank Adaptation (LoRA)**.^[26] LoRA freezes the base model’s weights and injects, in parallel with selected weight matrices (typically the attention query/value projections and the MLP projections), a trainable low-rank update $\Delta W = BA$, whose inner dimension r is far smaller than the model dimension. Only B and A are trained; the number of trainable parameters drops by three to four orders of magnitude, and, decisively for this application, the adapter is a small, separable artefact that can be attached or detached at will. One trains a *behaviour* rather than a new model.

Concretely, an “anti-epanorthosis” style adapter could be trained as follows. Assemble preference pairs from Section 7.1 (for each item, a high-emphasis response and a content-matched low-emphasis paraphrase), then optimise the adapter with **Direct Preference Optimisation (DPO)**,^[28] which fits preferences directly with a simple classification-style loss and no separate reward model or RL loop, making it well suited to a small, targeted stylistic objective. Because LoRA is memory-light and can be combined with 4-bit quantisation of the frozen base (the QLoRA recipe),^[27] the whole procedure is feasible on a single GPU, which matters for reproducibility and for practitioners outside large labs. The adapter’s separability yields the “rhetoric dial” directly: scaling the LoRA contribution by a coefficient α at inference time

moves, at moderate values, between the base model’s style and the de-emphasised style, and multiple adapters (concise, formal, plain-language) can be swapped per deployment. LoRA scaling is known to behave non-monotonically at large α , so the continuous “dial” must be validated empirically rather than assumed. The principal risk is *content drift*: an adapter trained to suppress a surface pattern may also suppress legitimate corrections or subtly alter meaning. This must be controlled explicitly, and it is harder than it sounds: a content-preservation term in the objective (for example, a semantic-similarity constraint against a reference) is a blunt instrument, because embedding similarity captures fine shifts of meaning poorly, so it should be paired with targeted human checks and the meaning-fidelity evaluation described in Section 7.7.

Two pilots: preference optimisation fails, supervised fine-tuning works. A first attempt trained the adapter on 570 preference pairs with DPO alone; it reached high preference accuracy on the training set yet failed to generalise, *raising* the mean Epanorthosis Index on a held-out Italian evaluation from 60.3 to 68.9 per ten thousand words. Small-data DPO learned to rank the pairs without moving the generation policy in the intended direction. A second attempt taught the affirmative style directly, by supervised fine-tuning on the plain (chosen) responses for three epochs at LoRA rank 32, over 797 semantically filtered pairs. This adapter cut the index by about 98%, from 75.4 to 1.7 per ten thousand words across oratory, promotional and argument, and inspection of its outputs confirms that they remain coherent, on-task prose with the emphatic corrections simply gone. One caveat remains, and it is precisely the caveat of Section 7.7: the adapter *eliminates* the figure rather than calibrating it, landing below the human rate; scaling the adapter contribution by a coefficient (the “dial”) is the lever to settle it at the human rate instead of at zero. Supervised fine-tuning therefore succeeds where preference optimisation alone did not, and the adapter of this section is a delivered result at pilot scale. The prompting intervention of Section 7.8 remains a lighter alternative when no training is possible.

7.3 Preference optimisation at the alignment stage

If the overuse is partly manufactured by RLHF,^{[10],[11]} the alignment stage is also the natural place to correct it. The same preference pairs used for a LoRA adapter can instead shape the main alignment objective (whether classical RLHF or DPO^[28]), so that raters’ latent preference for emphatic confidence is counter-balanced by an explicit preference for calibrated phrasing. This is more invasive than an adapter (it alters the shipped model rather than a detachable component) but addresses the cause rather than a symptom, and it connects the problem to the broader literature on unintended stylistic side-effects of preference tuning, of which sycophancy (models telling users what they want to hear) is the best-studied cousin.^[35] Emphatic epanorthosis can be read as a rhetorical form of the same reward-hacking: phrasing that *sounds* impressive to a rater independently of its content.

7.4 Decoding-time control without retraining

Several techniques steer generation at inference time, changing no weights at all: an attractive property when retraining is impossible or when per-request control is desired. **PPLM** perturbs the model’s hidden states at generation time using gradients from a small attribute classifier,^[29] **GeDi** uses a small generative discriminator to reweight the next-token distribution efficiently,^[30] and **FUDGE** multiplies the base model’s probabilities by those of a lightweight *future discriminator* that predicts whether an attribute will hold.

[31] An epanorthosis-intensity classifier, trainable from the annotations of Section 4, slots directly into any of these as the attribute model, downweighting continuations that open a “Not X. Y” frame. The crudest member of the family is direct *logit biasing* of the tokens that most often initiate the construction (“not just”, “rather”, “anzi”), which requires no classifier at all yet is brittle and easily circumvented by paraphrase. Decoding-time methods offer fine, reversible control at the cost of extra inference compute and, if pushed hard, some loss of fluency.

7.5 Activation steering and representation engineering

A newer line of work manipulates the model’s internal *activations* rather than its weights or its output distribution. If a stylistic property corresponds to a direction in activation space, one can find that direction and add or subtract it during the forward pass. Activation addition steers generation by adding a contrast-derived vector to the residual stream;^[32] representation engineering offers a top-down toolkit for reading and controlling such directions;^[33] and inference-time intervention demonstrated the approach for truthfulness by shifting activations along a learned direction at selected attention heads.^[34] An “epanorthosis direction”, defined as the mean activation difference between emphatic and plain paraphrases, could in principle be subtracted with a tunable coefficient, giving continuous, weight-free control with negligible added compute. The method is elegant and highly controllable; its risks are entanglement (the direction may correlate with desirable behaviours) and the current immaturity of guarantees about what a steered direction actually encodes.

7.6 Prompting, system prompts, and constitutions

The cheapest intervention is instruction: a system prompt or a constitutional clause^[24] forbidding gratuitous “Not X. Y” phrasing. It requires no engineering and is instantly reversible, which makes it the right first move. It is also the least robust: the very finding that instruction-tuned models diverge *more* strongly in rhetorical style^[1] is a warning that surface instructions do not reliably override a disposition baked in during training. Prompting mitigates locally; it leaves the underlying pressure in place.

7.7 Evaluation: measuring success without flattening

No mitigation can be trusted without an evaluation that distinguishes *calibration* from *suppression*. Three components are needed. First, a *style metric*: epanorthosis density (constructions per thousand tokens) measured by the Section 4 detector, reported *against the human baseline for the relevant genre* rather than against zero; the goal is to match human rates, and to avoid undercutting them. Second, a *content-fidelity metric*: semantic-similarity and factuality checks between the mitigated output and a reference, to catch the content drift of Section 7.2. Third, *human evaluation of appropriateness* on held-out genres, since appropriateness is exactly the judgement that resists automation. A mitigation that lowers epanorthosis density to the human baseline while preserving content fidelity and appropriateness ratings has succeeded; one that drives density below the human rate, or that trades style for meaning, has merely traded one artefact for another.

Table 1. Mitigation techniques for emphatic epanorthosis, ordered from most to least invasive.

Technique	Acts on	Cost	Controllability	Main risk
Data curation / augmentation (7.1)	Training data	Low–med.	None at inference	Slow feedback loop
LoRA / QLoRA style adapter (7.2)	Low-rank weight update	Low	High (detachable, scalable α)	Content drift
Preference optimisation / DPO (7.3)	Shipped model weights	Medium	Global, not per-request	Alters base behaviour
Decoding-time control: PPLM/GeDi/FUDGE (7.4)	Output distribution	Med. (inference)	High, reversible	Fluency loss; brittleness
Activation steering / RepE (7.5)	Internal activations	Very low	High, continuous	Entanglement; weak guarantees
Prompting / constitution (7.6)	Context	Negligible	Local only	Overridden by training prior

No single method is sufficient. The practical recommendation is a stack: curate the data, train a detachable LoRA style adapter optimised with DPO under a content-preservation constraint, expose it as a dial, and reserve decoding-time or activation-level control for cases needing per-request adjustment; the whole stack is governed by the genre-relative evaluation of Section 7.7.

7.8 A working demonstration

The cheapest leg of the stack can be tested directly, and in a second language. I generated seventy-two Italian texts across three genres (oratory, promotional, and argument), three instruction-tuned models, and four topics each, under two conditions: a neutral baseline prompt, and the *same* prompt plus a one-line instruction to write affirmatively and avoid negate-and-upgrade phrasing («Non è X. È Y», «non X ma Y», «anzi»). An Italian detector counts the emphatic-correction family («non X ma / bensì Y», «non solo X ma Y», «non X, è Y», and the sentence-split «Non è X. È Y») per ten thousand words.

Table 3. Italian mitigation demonstration: emphatic-epanorthosis density (per 10,000 words) under a neutral prompt vs. the same prompt plus a one-line anti-epanorthosis instruction, three instruction-tuned models. p from a 5,000-iteration permutation test.

Genre	Baseline	Mitigated	Reduction	p
Oratory	39.6	10.9	72%	0.03
Argumentative essay	56.9	17.2	70%	0.004
Promotional	32.3	16.8	48%	0.43

A single instruction cuts emphatic epanorthosis by roughly half to two-thirds: 72% in oratory ($p = 0.03$) and 70% in argument ($p = 0.004$), both significant, with a weaker and noisier 48% in promotional. The reduction is near-total for the larger models, where Sonnet and Opus fall to zero under the instruction, and partial for the smallest, where Haiku roughly halves. This is only the prompting leg of Section 7.6, the least robust rung of the stack, and it operates on Italian; even so, it shows that the figure can be turned down substantially at fixed content by the cheapest available means, which sets a floor that the trained adapter of Section 7.2 should exceed. It also confirms that the method and the phenomenon transfer across

languages. For reference, human Italian baselines sit near 12 per ten thousand words for argument (Beccaria) and near 14 for oratory (public-domain orations): the unmitigated models run two to five times higher, and the instruction brings them back to roughly the human rate rather than below it, which is the calibration target of Section 7.7.

7.9 Two training-free alternatives

Two more mitigations need no fine-tuning at all, only the detector of Section 4 applied at inference. *Best-of-n* generates several candidates and keeps the one with the lowest Epanorthosis Index; *rewrite* adds a second pass that recasts the draft affirmatively. On the same Italian prompts (six candidates each), both cut the figure sharply.

Table 4. Training-free mitigations: emphatic-epanorthosis density (per 10,000 words) for the natural output, a best-of-6 selection, and a post-hoc affirmative rewrite.

Genre	Baseline	Best-of-6	Rewrite
Oratory	22.5	0.0	0.0
Promotional	82.1	33.6	0.0
Argumentative essay	75.0	29.1	0.0
Overall	59.9	20.9	0.0

Best-of-n cuts the index by about two thirds and, because it selects a genuine model output rather than forcing one, tends to stay nearer natural phrasing, which makes it the gentlest of the methods here. The rewrite removes the figure essentially to zero, which, like the supervised adapter, over-corrects past the human rate and should be tempered when the goal is calibration rather than elimination. Both are model-agnostic and deployable without any training, at the cost of extra inference calls, and together with the prompting instruction they give three inference-time levers of increasing strength (best-of-n, instruction, rewrite) to sit alongside the trained adapter.

8. Conclusion

If a statistical system, knowing nothing of rhetoric, spontaneously reproduces the same figures humans have used for millennia, perhaps those figures are *emergent structures of language*: attractors toward which the medium tends, whether a brain or a network produces it. In that light the classical orators discovered epanorthosis rather than inventing it, as mathematicians discover theorems that were true before they were proved. Models, correspondingly, act as explorers of the space of linguistic possibility, arriving by statistics at structures we reach by cognition. The convergence is evidence that the structure belongs to language itself, independent of any particular speaker.

This reframes the engineering problem of Section 7. The techniques surveyed there (LoRA adapters foremost, alongside preference optimisation, decoding-time control, and activation steering) can measurably calibrate the figure toward human rates. The deeper task is to teach readers, and models, *appropriateness*: the sense of when a correction is earned and when it is empty. The real risk points the other way: that we would begin to write like them.

References

1. Reinhart, A., Markey, B., Laudénbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8), e2422455122. arXiv:2410.16107.
2. Fontanier, P. (1827). *Les Figures du discours* (ed. G. Genette, Paris: Flammarion, 1968).
3. Aristotle. *Rhetoric* (trans. G. A. Kennedy, *On Rhetoric*, 2nd ed., Oxford University Press, 2007).
4. Quintilian. *Institutio Oratoria*, Book IX (c. 95 CE) (Loeb Classical Library, Harvard University Press).
5. Cicero. *In Catilinam* I.2 and II.1 (63 BCE) (Loeb Classical Library, Harvard University Press).
6. Dante Alighieri. *Inferno*, Canto XXXIII, l. 75.
7. Lausberg, H. (1998). *Handbook of Literary Rhetoric: A Foundation for Literary Study*. Leiden: Brill.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *NeurIPS 2017*. arXiv:1706.03762.
9. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. *ICLR 2020*. arXiv:1904.09751.
10. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*. arXiv:2203.02155.
11. Juzek, T. S., & Ward, Z. B. (2024). Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. arXiv:2412.11385.
12. Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*. Preprint arXiv:2406.07016 (2024).
13. Liang, W., Izzo, Z., Zhang, Y., et al. (2024). Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. *ICML 2024*. arXiv:2403.07183.
14. Gehrmann, S., Strobel, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *ACL 2019, System Demonstrations*. arXiv:1906.04043.
15. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *ICML 2023*. arXiv:2301.11305.
16. Guo, B., Zhang, X., Wang, Z., et al. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597.
17. Mauri, C., Ballarè, S., Gorla, E., Cerruti, M., & Suriano, F. (2019). KIParla corpus: a new resource for spoken Italian. *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it 2019)*.
18. Dubremetz, M., & Nivre, J. (2018). Rhetorical Figure Detection: Chiasmus, Epanaphora, Epiphora. *Frontiers in Digital Humanities*, 5:10.
19. Bothwell, S., DeBenedetto, J., Crnkovich, T., Müller, H., & Chiang, D. (2023). Introducing Rhetorical Parallelism Detection: A New Task with Datasets, Metrics, and Baselines. *EMNLP 2023*. arXiv:2312.00100.
20. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022*. arXiv:2201.11903.
21. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *NeurIPS 2022*. arXiv:2205.11916.
22. Madaan, A., Tandon, N., Gupta, P., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. *NeurIPS 2023*. arXiv:2303.17651.

23. Huang, J., Chen, X., Mishra, S., et al. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. *ICLR 2024*. arXiv:2310.01798.
24. Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
25. Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
26. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022*. arXiv:2106.09685.
27. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *NeurIPS 2023*. arXiv:2305.14314.
28. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*. arXiv:2305.18290.
29. Dathathri, S., Madotto, A., Lan, J., et al. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *ICLR 2020*. arXiv:1912.02164.
30. Krause, B., Gotmare, A. D., McCann, B., et al. (2021). GeDi: Generative Discriminator Guided Sequence Generation. *Findings of EMNLP 2021*. arXiv:2009.06367.
31. Yang, K., & Klein, D. (2021). FUDGE: Controlled Text Generation With Future Discriminators. *NAACL 2021*. arXiv:2104.05218.
32. Turner, A. M., Thiergart, L., Leech, G., et al. (2023). Steering Language Models With Activation Engineering (the ActAdd method). arXiv:2308.10248.
33. Zou, A., Phan, L., Chen, S., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.
34. Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *NeurIPS 2023*. arXiv:2306.03341.
35. Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.

Appendix A. A reproducible mitigation recipe

The anti-epanorthosis adapter of Section 7.2 is given in full as a supplementary training script (*artificial-epanorthosis-lora-recipe.py*). In outline: build preference pairs by the counterfactual augmentation of Section 7.1 (each emphatic “Not X. Y” span paired with a content-matched, de-emphasised paraphrase, filtered to keep only high semantic-similarity pairs); freeze the base model and train a LoRA update (rank 16, α 32, on the attention query and value projections and the MLP projections) with Direct Preference Optimisation ($\beta = 0.1$); optionally quantise the frozen base to 4-bit (QLoRA) for single-GPU training. The output is a small, detachable adapter whose contribution can be scaled at inference (the “dial” of Section 7.2, validated empirically) and scored against the genre-relative human baseline of Section 7.7. The prompting-level demonstration of Section 7.8 sets the floor this adapter should exceed.

Supplementary materials. The detector, the training and evaluation scripts, the preference dataset, a ready-to-run notebook, and the pilot adapter with its training and evaluation logs accompany this paper for replication. Model weights are best distributed as a small LoRA adapter (a few megabytes) with a model card noting the base model and its licence; the emphatic/plain preference pairs and the human baseline corpus are released alongside.

Author: Federico Boggia, teacher and trainer, founder of Binatomy; degree in Humanities Computing (Informatica Umanistica) from the University of Pisa, with a specialisation in Language Technologies. Correspondence: federico@binatomy.com.