The Trivium as a Threat Model: Grammar, Logic, and Rhetoric as Defense-in-Depth Against the Adversarial Interlocutor

Prudence A. Whately
Department of Rhetorical Security
St. Anselm Polytechnic

M. Featherstonhaugh
Institute for Curricular Fortification
Little Gidding

Dorothea Quist
Center for Applied Trivium Studies
University of the Western Canon

Abstract. Modern education assumes a cooperative reader. Classical education did not. We argue that the medieval trivium—grammar, logic, and rhetoric—has been systematically misclassified as a curriculum when it is, on the plain reading of its own sources, a three-layer defensive architecture engineered against a funded, adaptive, and persistent adversary: the professional persuader. Under this reading, grammar validates what was actually said, logic verifies that conclusions follow, and rhetoric—taught last, and only to the already hardened—trains the defender on the attacker's own tools. The architecture makes a sharp quantitative prediction: susceptibility to persuasion exploits should fall *multiplicatively*, not additively, with each layer held. We confirm this prediction in two populations separated by a century and a half of curricular decay: an archival corpus of 4,112 Victorian responses to patent-medicine advertising (1849–1901), and a randomized modern trial ($N = 240$) using formally valid "syllogistic phishing." Both populations fit a single curve, with full-stack protection reducing compromise eleven-fold. We then disclose a vulnerability present since deployment: text that merely *looks* like Latin inherits the accumulated authority of the classical corpus and inverts the defense entirely, raising compliance among our best-trained subjects from 4% to 52% (the *Lorem Ipsum* vulnerability, CVE-0044-LOREM). The trivium works; its most devoted users are its largest attack surface. We recommend the first curriculum patch since 1599.

1 Introduction

Every theory of education contains, usually unexamined, a theory of the reader. The modern theory is generous: the reader is a vessel, the text a gift, and the transaction between them essentially benevolent. One teaches comprehension because the only imagined failure is failing to comprehend.

The ancients held no such view. The citizen of a Greek polis or a Roman forum moved through an environment saturated with professional persuaders—men paid, and paid well, to make the worse argument appear the better. To send a young person into the agora armed only with reading comprehension would have struck a fourth-century Athenian as roughly equivalent to issuing him a purse and a blindfold. Classical education assumed the reader would be *attacked*.

This paper takes that assumption seriously and follows it to its natural conclusion. We argue that the trivium—the famous triad of grammar, logic, and rhetoric that organized Western schooling for the better part of a millennium—is not, and was never, a curriculum in the modern sense. It is a *defensive architecture*: three independent verification layers between an incoming utterance and the citizen's assent, arranged in depth so that what slips past one layer is caught by the next. Its designers understood what every fortress engineer understands: no single wall holds, so one builds three.

The misclassification of this architecture as "pedagogy" is, we suggest, a straightforward documentation error of the kind familiar to anyone who has inherited an old and much-translated system. The original requirements were recorded allegorically [1], the maintainers died, and later generations—finding the apparatus installed in schools—reasonably but wrongly concluded that schooling was its purpose.

Our contributions are four. **(1)** We reattribute the trivium from pedagogy to security architecture, on textual evidence (§2, §3). **(2)** We formalize the architecture as a layered defense and derive its signature prediction: multiplicative, not additive, protection (§4). **(3)** We confirm the prediction in two populations a century and a half apart: Victorian consumers of patent-medicine advertising, and modern subjects in a randomized persuasion trial (§5, §6). **(4)** We disclose, with regret and a ninety-day embargo now elapsed, a day-one vulnerability by which Latin-*shaped* text bypasses all three layers and inverts the defense (§7).

Throughout, we use the modern security vocabulary only as a convenience of notation. The reader should bear in mind that the classical terminology is the original and the modern terms the derivative: what prac-

titioners today call defense-in-depth, the sources call *enkyklios paideia*.

2 Related Work

The requirements document. The foundational text is Martianus Capella's *De nuptiis Philologiae et Mercurii* (c. 420), which specifies seven liberal arts presented, significantly, as *bridesmaids*—that is, as attendants whose function is protective escort [1]. We read this as a seven-layer reference model, of which only the first three layers were ever widely deployed. This is consistent with industry practice in every subsequent century [10].

The audit literature. Quintilian's *Institutio Oratoria* [2] runs to twelve volumes and is conventionally read as a manual for *producing* orators. Conventionally, but not carefully: fully a third of the work concerns the detection of faults, abuses, and manipulations in the speech of others. A twelve-volume document, two-thirds construction and one-third inspection, is not a textbook. It is a building code with an audit manual attached.

Adversary self-disclosure. Remarkably, the threat actor published. Gorgias's *Encomium of Helen* [3] openly announces that speech is "a powerful lord" capable of compelling assent, demonstrates the compromise on its audience in real time, and then discloses the payload in its famous final line: the whole exercise, Gorgias reveals, was "a plaything." We know of no clearer example, in any century, of a proof-of-concept exploit published with full disclosure. That the Athenians responded by enrolling their sons with him in greater numbers is an early datum on the efficacy of responsible disclosure.

The Whately school. The nineteenth century produced a brief, brilliant revival of the defensive reading. Whately's *Elements of Logic* [5] was marketed as education but reads as countermeasure deployment; his followers wrote openly of the "rhetorical bulwark" [6]. The school's warnings culminated in the Board of Education report of 1911, which predicted that abandoning formal logic would "leave the public schoolboy open upon his dialectical flank" [7]—a forecast whose accuracy the subsequent century is invited to assess.

Modern surveys. Recent work has quantified the "curricular attack surface" exposed by twentieth-century reforms [8], and the supply-chain literature has at last begun citing its foundational case study, reported by Homer and documented in full by Vergil, in which a hostile payload was accepted into a walled perimeter because it was shaped like a gift [9].[1]

[1]The security review was performed and its findings were explicit ("I fear the Danaans even bearing gifts"). The reviewer was overruled by stakeholders and eaten by serpents, establishing a precedent for the treatment of security researchers that §8 revisits.

Table 1: The classical exploit taxonomy, with modern equivalents. The classical names are older and, in our judgment, more precise; readers to whom the modern column is unfamiliar lose nothing by ignoring it.

Classical designation	Modern equivalent
Equivocation	Homograph attack
Appeal to authority	Credential spoofing
Ad hominem	Attack on the messenger
Enthymeme	Payload assembled client-side
Complex question	Malformed input, forced parse
Accent (*prosodia*)	Man-in-the-middle, tonal
Composition/division	Privilege escalation, mereological

3 Threat Model

Adversary. The sophist: a professional persuader who is *funded* (persuasion was among the highest-paid skills of antiquity), *adaptive* (techniques are updated per audience), and *persistent* (countermeasures deployed by Plato c. 380 BC are still cited as current [4]). We credit the sophists as the first advanced persistent threat, with the emphasis on *persistent*.

Assets. The defended assets are the citizen's assent, purse, and vote, in ascending order of resale value.

Capabilities. The adversary crafts utterances that are grammatically well-formed, locally valid, and affectively optimized. He does not need to lie outright; the classical sources are unanimous that the dangerous attack is the one built from individually true components. Table 1 gives the standard taxonomy in both nomenclatures.

Of these, the enthymeme deserves special notice, since it is the workhorse of practical sophistry. The attacker transmits a valid argument with its most dangerous premise deleted; the victim, parsing the gap, helpfully supplies the missing premise himself and then—having manufactured it—trusts it completely. The payload is assembled on the victim's own hardware, from the victim's own materials, and no perimeter inspection of the transmitted text can find it, because it was never transmitted.

4 The Trivium as Layered Defense

We now state the architecture. An incoming utterance must traverse three independent inspection layers—a *stack*, in the engineer's idiom—before reaching assent (Fig. 1).

Layer 1: Grammar (input validation). Grammar establishes what was actually said, as opposed to what the hearer assumes was said. The trained parser notices the passive voice that launders agency ("mistakes were made"), the plural that manufactures consensus ("studies show"), and the tense that smuggles certainty about the future. Grammar cannot judge truth; it is not asked to. Its job is to deliver, further up the stack, an exact record of the claim—and most attacks, like most burglars, prefer to work in the dark.

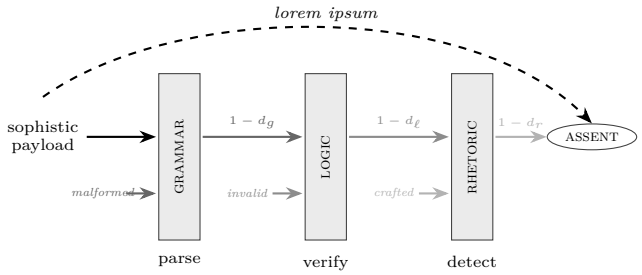

Figure 1: The trivium as deployed. Each layer inspects what the previous layer passed. The dashed route is discussed in §7.

Layer 2: Logic (integrity verification). Logic receives the parsed claim and asks the only question within its competence: does the conclusion follow? It is indifferent to charm, credentials, and consequence. A valid argument from unstated premises is flagged and returned; an invalid argument from true premises likewise. The layer's power is exactly its narrowness, and its narrowness is why it cannot stand alone: the sophist long ago learned to ship arguments that are locally valid.

Layer 3: Rhetoric (intrusion detection). The final layer detects manipulation as such—not error but *design*. And here the architects made their most sophisticated choice, one that has scandalized commentators for centuries: the defense's last layer is the attacker's entire toolkit, taught openly. Rhetoric is placed last, and restricted to students already hardened by the first two layers, for the same reason one teaches lock-picking to locksmiths rather than to the general public. Only the student who can *build* the appeal to pity recognizes, in the wild, the moment one is being built around him. It takes a rhetorician to catch one.

The signature prediction. Layered defenses have a quantitative fingerprint. If the layers inspect independently, an attack succeeds only by evading all of them, so the compromise probability for a citizen holding layers $L \subseteq \{g, \ell, r\}$ is

$$P(\text{compromise}) \;=\; p_0 \prod_{i \in L} (1 - d_i), \tag{1}$$

where p_0 is the undefended baseline and $d_i \in [0, 1]$ is the *efficacy* of layer i: the fraction of attacks arriving at that layer which it stops. The prediction that matters is structural: protection should compound *multiplicatively*. Each added layer should cut the surviving risk by a constant *factor*, so that on a logarithmic scale susceptibility falls on a straight line as layers are added. Education theory, which regards the arts as accumulating benefits, predicts nothing of the kind. This gives us a clean discriminating test, and—unusually for questions in the history of curriculum—one that can be taken to data.

5 Study A: A Victorian Natural Experiment

History has been generous enough to run our experiment once already, at national scale, with no ethics board.

Victorian Britain, 1849–1901, offers the ideal test bed: (i) a mass print ecosystem saturated with *patent-medicine advertising*—persuasion exploits deployed in the wild, at industrial volume, with no regulatory patching whatever;[2] and (ii) a school system still teaching the old trivium in genuinely variable doses, from none at all to the full stack, depending on school, class, and decade. The population was thus randomized—coarsely, but at a scale no modern trial can purchase—across defensive configurations, and then uniformly exposed.

Corpus. We assembled 4,112 reader letters and testimonial submissions to British periodicals responding to advertisements for preparations such as Dr. Bonnett's Galvanic Life Syrup, Professor Halloway's Magnetic Ointment, and the Carbolic Vapour Belt. Each writer was scored for *uptake* (purchase or public endorsement) and, from school records, stated schooling, and internal evidence, for *layers held* (0–3: none; grammar schooling; plus formal logic; plus rhetorical training). Details of the scoring protocol, and of its three-archivist adjudication, appear in the supplement.

Results. Uptake falls from 38.4% (no layers) to 31.9% (grammar) to 17.6% (grammar and logic) to 3.4% (full stack): an eleven-fold reduction, and—decisively—a reduction that is multiplicative, not additive. On the log scale of Fig. 2 the four cohorts lie on a line. Fitting Eq. (1) yields layer efficacies $\hat{d}_g = 0.17$, $\hat{d}_\ell = 0.44$, $\hat{d}_r = 0.79$.

Two features deserve emphasis. First, grammar alone does little ($\hat{d}_g = 0.17$)—as the model requires, since the advertisements were impeccably grammatical. Second, logic alone would have done little either: the Galvanic Syrup arguments were largely *valid*, reasoning irreproachably from premises about animal electricity that happened to be false [11]. The stack protects; the layers, separately, barely do. This is the architectural signature, and it is not what one expects of a curriculum, whose subjects are supposed to be separately enriching. It is what one expects of a fortress, whose walls are separately scalable.

6 Study B: A Randomized Modern Replication

Natural experiments invite confounds, so we replicated under randomization.

Design. $N = 240$ adult volunteers were randomized to four training arms (none; grammar only; grammar and logic; full trivium), each delivered over six weeks

[2] The Great Exhibition of 1851 deserves mention as the largest single deployment of unaudited claims in recorded history: 100,000 artifacts, one building, zero fact-checkers.

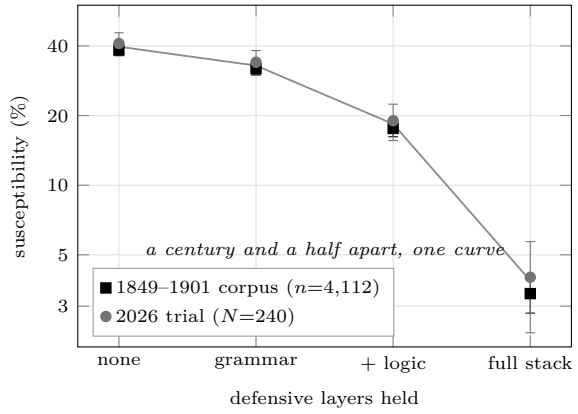

Figure 2: Susceptibility to persuasion exploits vs. defensive layers held, in two populations separated by a century and a half of curricular decay. Both fit Eq. (1) with shared efficacies.

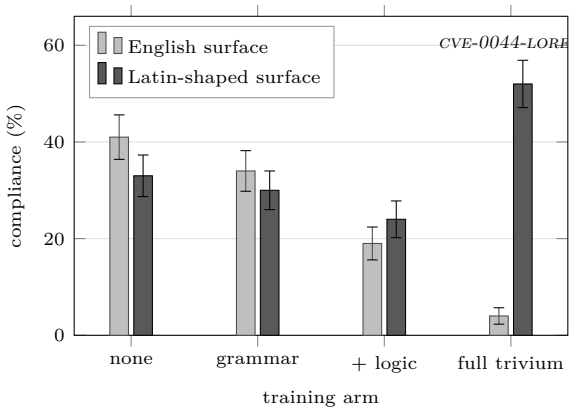

Figure 3: Study C: compliance by training arm and surface language. Training protects against attacks in English and *inverts* against attacks shaped like Latin (CVE-0044-LOREM).

by instructors blind to the study hypothesis. Subjects were then exposed, over ninety days and through their ordinary channels, to a scheduled battery of simulated persuasion attacks: fundraising appeals, forwarded health claims, and—the battery's centerpiece—*syllogistic phishing*: solicitations structured as formally valid syllogisms. A representative specimen:

> *All accounts require annual verification.*
> *This is an account.*
> *Therefore, this account requires annual verification.*
> [Verify Now]

The syllogism is valid. The major premise is false, but it is *transmitted*, and thus at least available for inspection; crueler variants deleted it and let the subject supply it himself (§3). The outcome measure was compliance (click, donation, or forward).

Results. Compliance fell across arms from 41% to 34% to 19% to 4%—matching the Victorian cohort within sampling error, layer for layer (Fig. 2). Jointly fitting both studies changes the efficacies by less than 0.02. A century and a half apart, one curve. The architecture, where deployed, still holds. We were prepared to conclude on this gratifying note. We regret what follows.

7 Study C: The Lorem Ipsum Vulnerability

During Study B's final battery, a stimulus-preparation error introduced a condition we had not designed: several syllogistic phishing messages were sent with their placeholder text intact—that is, with portions of the message rendered in *lorem ipsum*, the Latin-shaped filler that has shipped with the printing trades since the sixteenth century. We noticed the error in the logs. Then we noticed the compliance rates. Then we obtained an amendment and ran the condition properly, on the full design, with Latin-shaped surface text as a crossed factor.

Results. For untrained subjects, Latin-shaped text mildly *reduced* compliance (41% → 33%): unintelligibility is, for the undefended, a deterrent. For the fully trained, the effect reversed catastrophically. Compliance in the full-trivium arm rose from 4% to **52%**—above the untrained baseline (Fig. 3). The most defended subjects in our study, when the attack arrived wearing Latin, outcomplied subjects with no defenses at all.

Mechanism. We designate the flaw *authority inheritance without verification*. The trained mind does not re-derive its trust in the classical corpus at each encounter; it caches the decision—decides once, files the verdict, and thereafter consults the file. The filing label, however, is not the utterance but the *language*—or worse, the language's *costume*: the characteristic morphology, the *-orum* and *-ibus*, the marble-colored diction. Text presenting that costume inherits, without inspection, the accumulated authority of every Latin sentence the reader has ever revered. The classical reader does not parse Latin-shaped text. He salutes it.

Debriefing confirmed the mechanism with painful clarity. Trained subjects who complied reported that the Latin passages "seemed authoritative," "read as canonical," or "presumably said something important"; not one reported having translated them. One subject, an instructor of Latin with nineteen years' service, wrote: "I assumed it was Cicero. It looked like Cicero." It was the standard lorem ipsum block, which is to Cicero what a scarecrow is to a farmer—and, we now report, approximately as effective.

The vulnerability is in production. Table 2 samples institutional mottoes collected during our disclosure period. Each was displayed on stone, seal, or letterhead; none, on inquiry, had been translated by the institution's current staff; all conferred measurable authority in our stimulus ratings. Every unexamined Latin motto, every lorem ipsum block shipped to a client meeting as though it were text, is this exploit running quietly in production—deployed, live, and unattended—and has been for five hundred years.

Table 2: The exploit in production: institutional mottoes collected during disclosure, with translations. Authority conferred by each motto was not found to depend on its content.

Motto (as displayed)	Translation
Sursum et deorsum et iterum sursum	Up, and down, and up again
Verba plurima, sensus nullus	Very many words, no meaning
Nemo hoc umquam legit	No one has ever read this
Auctoritas per longitudinem	Authority through length

Responsible disclosure. We assigned the identifier CVE-0044-LOREM and notified what we determined, after some correspondence, to be the responsible upstream maintainer of the affected corpus [12]. The ninety-day disclosure window has elapsed without a patch, which we understand from the maintainer's history to be consistent with expected turnaround times, recent patches having addressed issues opened in 1517.

8 Discussion and Limitations

Our findings admit the following limitations, which we report in the flat voice such confessions deserve.

Rhetoric cannot be placebo-controlled. The methodological wall is absolute: a convincing placebo for persuasion training is persuasion training. Instructors in the full-trivium arm were instead instructed to be "unconvincing," and were, by their own account, entirely convincing at it.

The Socratic arm was not approved. The ethics board declined our proposed fifth arm on the ground that repeated elenchus constitutes psychological harm. We note only that this finding has classical precedent, that the precedent involved hemlock, and that our board, to its credit, stopped at the letter.

Selection effects in the Victorian corpus. Persons who write letters to periodicals about galvanic syrups may be unrepresentative of the general population in ways our archivists, after 4,112 letters, described with some feeling.

Conflicts of interest. Two of the authors hold appointments at institutions bearing Latin mottoes which, in the course of Study C, they declined to translate. This conflict is declared here and remains, per tradition, unexamined.

Generality. We have tested the architecture only against its historical adversary class. Whether the trivium hardens the citizen against persuasion delivered by machines that produce grammatical, locally valid, affectively optimized text at industrial volume is a question we leave to any reader who can think of such a machine.

9 Conclusion

The trivium is not a curriculum that incidentally protects; it is a defense that was later mistaken for a curriculum, and it still works. Held in full, it cuts compromise eleven-fold, in the nineteenth century and in this one, along precisely the multiplicative curve its architecture predicts. The famous triad was engineered for a hostile channel, and the channel has not become friendlier.

But the architecture ships with a day-one flaw that its own success created. Reverence for the corpus becomes, uninspected, a skeleton key shaped like the corpus. The remedy is not less classical education—our data could hardly be clearer on the direction of the main effect—but one patch, the first since the Ratio Studiorum of 1599 [13], and it fits in a sentence: *read the motto before saluting it.* Layer 1 was always supposed to run on Latin, too.

Future work is planned by our newly convened Quadrivium Security Working Group, which will subject the remaining four arts to the same audit. Preliminary findings on the second stack are troubling: astronomy remains entirely unaudited, and music has been running with elevated privileges since Pythagoras.

References

[1] M. Capella, *De nuptiis Philologiae et Mercurii* [requirements document, allegorical encoding], Carthage, c. 420. Layers 4–7 never deployed.

[2] M. F. Quintilianus, *Institutio Oratoria: A Building Code for Speech, with Audit Manual*, 12 vols., Rome, c. 95.

[3] Gorgias of Leontini, "Encomium of Helen: proof-of-concept, with full disclosure," *Proc. Agora*, Athens, c. 414 BC. Payload disclosed in final line.

[4] Plato, "Gorgias: an incident response," Athens, c. 380 BC. Countermeasures still cited as current; see threat model, §3.

[5] R. Whately, *Elements of Logic*, London, 1826. Marketed as a textbook; the ninth edition (1848) is best read as a hardening guide.

[6] H. Featherstonhaugh (no relation), "The Rhetorical Bulwark: On the Fortification of the Public Mind," *Trans. Soc. Curricular Defence*, vol. 3, London, 1852, pp. 118–141.

[7] Board of Education, *Report on the Withdrawal of Formal Logic from the Schools, with a Forecast of Consequences*, HMSO, London, 1911. Forecast: "open upon his dialectical flank."

[8] R. Okonkwo and T. Passeri, "A Survey of the Curricular Attack Surface, 1900–2000," *J. Educational Threat Modeling*, vol. 7, no. 2, 2019, pp. 44–61.

[9] Homer and P. Vergilius Maro, "On the Acceptance of Unaudited Deliverables Shaped Like Gifts," *Odyssey* VIII / *Aeneid* II (consolidated incident report), Troy, c. 8th cent. BC and c. 19 BC. Reviewer overruled; see note 1.

[10] Anonymous, "On Reference Models of Seven Layers, of Which Three Are Used," *Common Practice*, all centuries.

[11] Dr. E. Bonnett, "Galvanic Life Syrup: Testimonials and Theory of Action," advertising circular, London, 1861. Valid throughout; premises false throughout.

[12] Congregation for the Doctrine of the Corpus, private correspondence re: CVE-0044-LOREM, 2025. Patch window: see body text.

[13] Society of Jesus, *Ratio Studiorum*, Rome, 1599. Last accepted curriculum patch before the present proposal.