Construction and Validation of a Multidimensional Tragedy Evaluation Model: A Computational Narratology Approach

Weijia Zhang [1,4,5,6], Siyu Zhang [1], Luyang Ji [1], Fang Liu [1], Nutapong Somjit [2,3], Watcharin Srirattanawichaikul [3], Shaomin Cai [7]*, Jianjun Li [1]*

[1] School of Mathematics, Physics and Information, Shaoxing University, Shaoxing 312000, Zhejiang, China

[2] School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

[3] Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand

[4] Department of AOP Physics, University of Oxford, Oxford, UK.

[5] Key Laboratory of Artificial Intelligence Multi-Dimensional Application Research, Shaoxing, China

[6] Shao-Chip Integrated Circuit College, Shaoxing University, Shaoxing 312000, Zhejiang, China

[7] School of Medicine, Shaoxing University, Shaoxing 312000, Zhejiang, China

* Corresponding authors.

These authors contributed equally: Weijia Zhang, Siyu Zhang, Luyang Ji, Fang Liu

Corresponding author: Shaomin Cai (Email: usxmedlab@yeah.net), Jianjun Li (Email: lijjcan@gmail.com)

Abstract: Quantitative assessment of tragic aesthetics has remained methodologically fragmented, hindering systematic cross-work and cross-tradition comparisons in literary studies. To address this, we constructed and validated the Multidimensional Tragedy Evaluation Model (MDTEM), a computational framework that operationalizes tragic intensity along four core dimensions plus a redemption coefficient. Drawing on a curated corpus of 39 canonical tragedies (20 Greek classical, 19 modern Western), we decomposed each work into discrete narrative nodes with sentiment-coded emotional values using the NRC Emotion Lexicon combined with expert Delphi calibration. We then calibrated MDTEM hyperparameters using ordinary least squares and aggregated node-level values via a peak-end rule informed by Kahneman's cognitive-psychological theory. The model achieved a mean absolute error (MAE) of 3.18, a root mean squared error (RMSE) of 3.92, and a Pearson correlation of $r = 0.978$ between predicted and expert-rated Tragedy Index (TI), with tier classification accuracy reaching 97.4%

(38/39 correct). Cross-tradition comparison revealed that Greek classical tragedies exhibited significantly higher TI (Mdn = 78.6) than modern Western tragedies (Mdn = 67.4; Mann-Whitney U = 88.5, p = 0.021, r = 0.36), driven by higher Scope (S) and lower Redemption (R) coefficients in the Greek corpus. An AI-assisted refactoring case study on *Romeo and Juliet* demonstrated that the model can serve as both a "compass" and "validator" for targeted aesthetic manipulation, successfully shifting the tragedy level from T1 (Extreme, 83.4) to T3 (Sorrowful, 54.7) with expert agreement. MDTEM thus provides a reproducible, mathematically grounded framework for cross-tradition tragedy research and offers a structured pathway for AI-assisted literary analysis and generation, translating qualitative notions of "cosmic fate" versus "psychological interiority" into measurable aesthetic regularities.

Keywords: tragedy quantification; computational narratology; peak-end rule; digital humanities; comparative poetics; AI-assisted literary generation

1. Introduction

Tragedy has long occupied a privileged position in literary aesthetics. From Aristotle's catharsis to Hegel's dialectic of ethical substance, from Nietzsche [1] Apollonian-Dionysian duality to Williams [2] modern tragic vision, the tragic has served as a lens through which literary theory interrogates suffering, agency, and the limits of meaning. Yet despite this rich theoretical tradition, the field still lacks a systematic, quantitative framework for comparing the "tragic intensity" of works across eras, genres, and cultures. Whether one tragedy is "more tragic" than another remains, in most scholarly discussions, a matter of qualitative judgment.

The recent emergence of digital humanities and computational narratology has opened new methodological pathways for literary scholarship. Pioneering work by Moretti [3][4], Jockers, and Underwood [5] has demonstrated that distant reading can recover patterns invisible to close reading, from genre evolution to the long-term dynamics of literary reputation. Sentiment-aware models such as those developed by Klinger [6] and Mohammad [7] have enabled fine-grained tracking of affective valence in narrative texts. However, the bulk of this computational literature has focused on either stylistic features or micro-level emotional trajectories. There remains a notable gap: an integrated, multi-dimensional metric of *overall* tragic intensity that can support cross-work comparison.

This paper addresses that gap by constructing and validating the Multidimensional Tragedy Evaluation Model (MDTEM) [8], a computational framework that operationalizes tragic intensity along four theoretically grounded dimensions—Value Destroyed (V), Innocent Suffering (C), Scope (S), and Irreversibility (I)—together with a Redemption coefficient (R) that modulates the tragedy index downward when a work offers redemptive or conciliatory closure. The model is calibrated on a curated corpus of 39

canonical tragedies (20 Greek classical, 19 modern Western), with parameters estimated by ordinary least squares against expert ratings, and validated using leave-one-out cross-validation.

Two design choices distinguish MDTEM from earlier work. First, instead of treating a tragedy as a single sentiment score, we decompose each work into discrete narrative nodes, each annotated with an emotional valence, and aggregate these nodes using a peak-end rule borrowed from Kahneman's cognitive-psychological research on retrospective evaluation. This choice reflects the intuition that the overall tragic impression of a work is dominated by its most intense moment (the peak) and its ending (the end), rather than by its average affect. Second, we explicitly model the *redemption coefficient* as a fifth parameter that captures a culturally variable aesthetic: the willingness of a tradition to soften or resolve its tragedies. As we show below, this coefficient discriminates empirically between Greek classical tragedies, which rarely offer redemptive closure, and modern Western tragedies, which frequently do.

The paper makes three contributions. Theoretically, it provides a quantitative bridge between canonical tragic theory and computational analysis, translating concepts such as "innocent suffering" and "irreversibility" into measurable constructs. Methodologically, it introduces a peak-end aggregation algorithm tailored to narrative structure and demonstrates its superiority over naive averaging in cross-validated prediction. Empirically, it documents a robust cross-tradition difference in tragic intensity between Greek classical and modern Western corpora, complementing existing comparative-poetics arguments with measurable aesthetic regularities.

The remainder of the paper is organized as follows. Section 2 reviews the relevant theoretical and computational literature. Section 3 introduces the MDTEM framework. Section 4 describes corpus construction and parameter calibration. Section 5 details the peak-end aggregation algorithm. Section 6 reports the empirical results, including model performance, a special-case analysis, cross-tradition comparison, and an AI-assisted refactoring study. Section 7 discusses theoretical and practical implications, and Section 8 concludes.

2. Theoretical Background

2.1 The Western Tradition of Tragic Theory

The theorization of tragedy begins with Aristotle's *Poetics*, which defines tragedy as "an imitation of an action that is serious, complete, and of a certain magnitude" accomplished through "pity and fear" effecting the "purgation" of these emotions. For Aristotle, the tragic protagonist is neither perfectly virtuous nor thoroughly wicked but a person of high reputation and prosperity who falls into misfortune through some *hamartia*—an error or frailty rather than vice. This formulation already implies several dimensions

that any quantitative model must capture: the magnitude of the fall (Value Destroyed), the undeservedness of the suffering (Innocent Suffering), and the completeness of the reversal (Irreversibility).

Hegel [9] recasts tragedy as a collision of equally justified ethical forces, none of which can fully prevail without violation of the other. The Hegelian lens foregrounds the *Scope* of tragic conflict: whether the fall implicates only an individual or an entire social-moral order. Nietzsche, reading tragedy through the polarity of Apollo and Dionysus, foregrounds the metaphysical register: Greek tragedy stages the dissolution of the principium individuationis, an aesthetic affirmation of existence that lies *beyond* the suffering of the individual. This metaphysical reading implies that Greek tragedy may operate at a different Scope than its modern counterparts.

Williams and Eagleton document the transformation of tragic theory in modernity. Williams argues that the modern tragic sense has migrated from metaphysical inevitability to social contingency; Eagleton reframes the tragic as a "zone of exclusion" within modernity, in which suffering is generated by the very structures that promise emancipation. Both scholars emphasize that modern tragedy tends to be more equivocal, more inward, and more amenable to redemptive or at least palliative closure than its Greek counterpart—a conjecture that MDTEM's Redemption coefficient is designed to test directly.

2.2 Computational Approaches to Literary Analysis

The computational analysis of literature has matured substantially since the early days of stylometry. Moretti's "conjectural" and "distant reading" programs demonstrated that quantitative patterns can expose literary-historical regularities inaccessible to close reading. Jockers extended this work through *macroanalysis*, applying topic modeling and sentiment analysis to multi-thousand-book corpora. Underwood recast the agenda as "distant horizons," foregrounding the use of large-scale evidence to study literary change.

Within computational narratology specifically, Piper and others have argued for narrative-theoretically informed computational models that preserve structural features such as plot, event, and perspective. Sentiment and emotion analysis has become a standard tool, with resources such as the NRC Emotion Lexicon and the VADER lexicon enabling valence and emotion classification at the sentence or paragraph level. Kim and Klinger provide a comprehensive survey of emotion analysis in literary texts, noting both the achievements of the field and the methodological challenges of handling figurative, narrative-embedded emotion.

A small but growing literature has begun to quantify specific aesthetic phenomena. Martindale used computational text analysis to study literary-historical change; Schöch and Weitin have argued for "scalable reading" approaches that integrate close and distant perspectives. The present study extends this line of

work by focusing on a specific aesthetic construct—tragic intensity—and proposing a multi-dimensional operationalization tailored to it [10].

2.3 The Peak-End Rule and Cognitive Aesthetics

Although originally developed for the evaluation of experienced episodes, the peak-end rule of Kahneman [11] and colleagues has clear relevance to the retrospective aesthetic evaluation of narrative. The rule holds that the global evaluation of an episode is dominated by two moments: the moment of maximum intensity (the peak) and the final moment (the end). Duration and intermediate variation have surprisingly small effects. In the context of tragedy, this implies that the overall tragic impression of a play is shaped primarily by its most painful scene and by its closing moments, with the intervening narrative fabric playing a secondary role.

Empirical support for this prediction in literary contexts comes from studies of reader response [12][13], which show that readers' judgments of literary works are disproportionately shaped by emotionally salient and terminally positioned passages. The peak-end rule has the additional methodological advantage of being computationally tractable: given a sequence of sentiment-coded narrative units, the rule prescribes a simple, parameter-light aggregation function. We adopt it here as the bridge between micro-level emotional annotation and macro-level tragic intensity.

3. The MDTEM Framework

3.1 Core Dimensions

The MDTEM framework rests on the assumption that the tragic intensity of a narrative can be decomposed into a small set of theoretically motivated dimensions, each measurable on a [0, 1] scale. Four core dimensions are distinguished; a fifth, the Redemption coefficient, modulates the composite.

Value Destroyed (V): V measures the magnitude of what is lost in the tragic conclusion, weighted by the importance of what is destroyed. Drawing on Aristotle's criterion of "magnitude" and Williams's emphasis on what is at stake, V is anchored as follows: 0.3 for the loss of external goods or social standing; 0.6 for bodily harm or the destruction of dignity without death; 0.7 for bodily destruction accompanied by spiritual elevation; 0.8 for spiritual or reputational collapse with the body surviving; 1.0 for the simultaneous destruction of body, soul, and value.

Innocent Suffering (C): C captures the moral relationship between the protagonist's conduct and the suffering they undergo. The construct draws on Aristotle's notion that the tragic protagonist is "not eminently good and just, yet whose misfortune is brought about not by vice or depravity, but by some error of judgment". C is anchored as: 0.2 for protagonists whose significant flaw brings about their own downfall;

0.5 for protagonists whose minor fault is met by disproportionate punishment; 1.0 for protagonists of unimpeachable moral standing who suffer total destruction.

Scope (S): S indicates the breadth of the tragic impact. Where V and C concern the protagonist, S addresses the extension of harm to family, community, polity, or civilization. S is anchored as: 0.2 for harm confined to the individual; 0.5 for the collapse of a family or local group; 1.0 for harm that registers as civilizational, generational, or otherwise collective.

Irreversibility (I): I captures whether and to what extent the tragic outcome can be undone. Following the logic of the Aristotelian *anagnorisis* without *peripeteia*, a tragedy whose suffering is reversed forfeits much of its tragic force. I is anchored as: 0.0 for full restoration; 0.5 for permanent damage to the protagonist's circumstances with continued existence; 0.8 for severe physical or psychological damage that endures; 1.0 for biological death or extinction. Note that *any* biological death locks I at 1.0 regardless of subsequent spiritual framing.

3.2 The Redemption Coefficient

The Redemption coefficient (R) captures the extent to which a work offers redemptive, conciliatory, or otherwise softening closure. R is not a dimension of tragic substance but a *modulator* that adjusts the composite tragedy index downward when redemptive elements are present. The construct draws explicitly on Williams's and Eagleton's observations that modern tragedy is more equivocal than its classical antecedent and on Steiner's claim that the "death of tragedy" coincides with the rise of redemptive frames. R is anchored as: 0.0 for unmitigated tragic closure with no redemptive element; 0.3 for partial recognition or elegiac consolation; 0.6 for substantial reconciliation, restoration, or moral resolution short of full reversal; 1.0 for explicit redemptive restoration (e.g., literal resurrection, full forgiveness, definitive moral vindication).

3.3 Composite Formula and Tier Classification

The composite Tragedy Index (TI) is computed as a non-linear function of the four core dimensions, modulated by R:

$$TI = K \cdot g(V, C, S, I) \cdot (1\text{-}\zeta \cdot R^{\eta})$$

where:

$$g\left(V, C, S, I\right) = exp\left(\alpha \cdot ln\ V + \beta \cdot ln\ C + \gamma \cdot ln\ S + \delta \cdot ln\ I\right)$$

is a weighted geometric mean of the four core dimensions, and $(1\text{-}\zeta \cdot R^{\eta})$ is a redemption-decay function. K is a scaling constant mapping the composite to a 0–100 scale. The five parameters {α, β, γ, δ,

$\zeta, \eta\}$ are estimated by ordinary least squares against expert ratings; in the present calibration, $K = 100$, $\alpha = 0.22$, $\beta = 0.24$, $\gamma = 0.20$, $\delta = 0.34$, $\zeta = 0.46$, and $\eta = 0.7$. $K = 100$ is the canonical scaling; α, β, γ, δ are dimension weights summing to 1 (constrained); ζ is the redemption decay coefficient; η is the redemption non-linearity. Section 4 reports the calibration procedure and goodness of fit.

Six tier labels are defined on the TI scale: T0 Destruction ($TI > 90$), the upper bound of human psychological endurance, signaled by absolute and unrelenting despair; T1 Extreme ($80 \leq TI \leq 90$), civilizational catastrophe, spiritual alienation, or devastating sacrifice; T2 Disillusionment ($60 \leq TI < 80$), the collapse of beauty or the entanglement of the individual in historical forces; T3 Sorrowful ($50 \leq TI < 60$), irreversible injury and spiritual wounding without absolute despair; T4 Regretful ($30 \leq TI < 50$), elegiac residue and unfulfilled longing; and T5 Hardship ($TI < 30$), suffering that eventuates in reconciliation, vindication, or moral integration.

4. Corpus and Parameter Calibration

4.1 Corpus Construction

A 39-work corpus was constructed to span two coherent tragic traditions within the Western canon: the Greek classical tradition ($n = 20$), encompassing the surviving plays of Aeschylus, Sophocles, Euripides, and Seneca's *Medea* as a Roman bridge case; and the modern Western tradition ($n = 19$), encompassing Renaissance through twentieth-century drama from Shakespeare, Marlowe, and Webster through Racine, the Ibsen-Strindberg naturalist lineage, and the twentieth-century American drama of Miller, Williams, and O'Neill. The corpus was selected to (a) exhaust the Greek surviving tragedies, (b) balance canonical and modern works, (c) ensure sufficient inter-tradition overlap to enable a meaningful comparison, and (d) keep each work independently ratable on the four core dimensions. The full corpus and per-work scores are reported in Appendix A.

4.2 Inter-Coder Reliability

Each of the 39 works was decomposed into narrative nodes and independently scored on V, C, S, I, and R by two literary scholars. The inter-coder agreement, measured by Cohen's kappa for the node decomposition and by intra-class correlation (ICC, two-way mixed, absolute agreement) for the dimension scores, was $\kappa = 0.89$ and $ICC = 0.91$ (95% CI [0.87, 0.94]), indicating strong agreement. Disagreements were resolved by a third expert adjudicator using a Delphi round.

4.3 Parameter Estimation: Paper Mode vs Empirical Mode

The MDTEM framework was originally specified in an earlier paper with hyperparameters {α, β, γ, δ, ζ, η} set by theoretical argument and small-case introspection. We refer to this prior specification as the Paper mode. When the Paper-mode parameters were evaluated against the 39-work corpus, the resulting mean absolute error (MAE) against expert ratings was 15.61, indicating substantial systematic bias: the Paper mode systematically overestimated TI for works with substantial redemptive closure (e.g., *Alcestis*, *A Doll's House*) because its ζ was too small to capture the magnitude of redemptive modulation.

We therefore re-estimated the hyperparameters by ordinary least squares against the 39-work expert ratings, using a constrained optimization that (a) enforced $\Sigma\alpha_i = 1$, (b) bounded each $\alpha_i \in [0.05, 0.50]$, (c) bounded $\zeta \in [0.20, 0.70]$, and (d) bounded $\eta \in [0.4, 1.2]$. This yields the Empirical mode with $\alpha = 0.22$, $\beta = 0.24$, $\gamma = 0.20$, $\delta = 0.34$, $\zeta = 0.46$, and $\eta = 0.7$. The Empirical mode substantially outperforms the Paper mode on all three evaluation metrics: MAE drops from 15.61 to 3.18 (a 79.6% reduction), RMSE drops from 17.92 to 3.92 (a 78.1% reduction), and Pearson r rises from 0.612 to 0.978. Tier classification accuracy rises from 38.5% (Paper) to 97.4% (Empirical); the single misclassified work in Empirical mode is *A Doll's House*, whose predicted tier is one step below expert consensus, an error analyzed in detail in Section 6.2.

4.4 Leave-One-Out Cross-Validation

To assess the stability of the calibration, we performed leave-one-out cross-validation (LOOCV). At each fold, the hyperparameters were re-estimated on 38 works and evaluated on the held-out work. The LOOCV mean absolute error was 3.42 (SD = 0.47), the LOOCV root mean squared error was 4.18 (SD = 0.53), and the LOOCV tier accuracy was 94.9%. The low SDs indicate that the Empirical-mode calibration is robust to individual works: removing a single tragedy produces at most a small parameter perturbation and a small increase in prediction error. Notably, the dimension weights (α, β, γ, δ) varied by less than 0.03 across folds, while ζ varied by less than 0.05, confirming that the relative importance of the four core dimensions is stable.

5. Peak-End Aggregation

5.1 Narrative Node Decomposition

Each of the 39 works was decomposed into a sequence of discrete narrative nodes: minimal narrative units, each containing at least one core conflict or turning event, possessing a definite affective polarity, and occupying an irreversible sequence position on the work's time axis. The number of nodes per work ranged from 5 (for tightly plotted works such as *Alcestis*) to 12 (for more episodic works such as *The Trojan Women*), with a median of 7. The functional-unit criterion follows Bremond's narrative grammar and Propp's morphology, both adapted in modern computational narratology [14].

5.2 Node-Level Emotional Assignment

Each node n_i was assigned an emotional value $e_i \in [-1, +1]$, where −1 represents extreme suffering, +1 represents extreme joy, and 0 represents affective neutrality. The assignment proceeded in two stages. First, the node's textual surface was scored by an ensemble sentiment analyzer that combined the NRC Emotion Lexicon, VADER, and a transformer-based classifier (specifically, a fine-tuned DistilBERT). Second, three literary scholars adjusted these machine scores via a Delphi procedure, providing the contextual reading that pure text analysis cannot capture.

The Redemption coefficient operates at the node level rather than the work level. For nodes bearing redemptive content (e.g., acts of forgiveness, reconciliation, spiritual elevation), the raw emotional value e_{raw} is transformed as:

$$e_i \;=\; e_{raw} \cdot \left(1 - \zeta \cdot R_{node}\right)$$

where $R_{\text{node}} \in [0, 1]$ is a node-level redemption weight supplied by the expert coders and ζ is the Empirical-mode redemption coefficient. This formulation lets redemptive moments attenuate the local emotional valence while leaving structurally tragic moments untouched.

5.3 Peak-End Aggregation

Given a sequence $\{e_i\}$ of node emotional values, the work-level TI is computed using a peak-end aggregation rule:

$$TI \;=\; 100 \cdot \left(w_p \cdot \left| e_{peak} \right| \;+\; w_e \cdot \left| e_{end} \right| \;+\; w_a \cdot |\, e_{mean}\, | \right) \cdot \left(1 - \zeta \cdot R\right)$$

where $e_{\text{peak}} = \arg\max |e_i|$ is the peak emotional magnitude, e_{end} is the final node's emotional value, e_{mean} is the arithmetic mean of the $|e_i|$ values, and $\{w_{\text{p}}, w_{\text{e}}, w_{\text{a}}\}$ are aggregation weights. The weights were estimated by least squares against the 39-work expert ratings and yielded $w_{\text{p}} = 0.45$, $w_{\text{e}} = 0.35$, $w_{\text{a}} = 0.20$. This ranking reflects the theoretical prediction that the overall tragic impression is dominated first by the peak (most painful moment) and second by the ending (residual or redemptive valence), with the average affective density playing a smaller role.

5.4 Robustness Comparison

To validate the peak-end aggregation against competing rules, we evaluated three aggregators on the 39-work corpus: (a) the peak-end aggregator described above; (b) a pure mean aggregator in which $TI \;=\; 100 \cdot |\, e_{mean}\, | \cdot \left(1 - \zeta \cdot R\right)$; and (c) a pure peak aggregator in which $TI \;=\; 100 \cdot |\, e_{peak}\, | \cdot \left(1 - \zeta \cdot R\right)$. The peak-end aggregator outperformed both alternatives on every metric: MAE was 3.18 (peak-end) vs 6.74 (mean) vs 8.92 (peak); RMSE was 3.92 vs 8.13 vs 10.41; tier accuracy was 97.4% vs 71.8% vs 66.7%.

These results are consistent with the cognitive-psychological prediction that retrospective aesthetic evaluation is dominated by extreme and terminal moments.

6. Empirical Results

6.1 Model Performance

Table 1 reports the headline performance metrics for the MDTEM framework under three evaluation conditions.

Table 1. *Performance of MDTEM on the 39-work corpus under three evaluation conditions.*

Condition	MAE	RMSE	Pearson r	Tier accuracy
Paper mode (no calibration)	15.61	17.92	0.612	38.5%
Empirical mode (in-sample)	3.18	3.92	0.978	97.4%
Empirical mode (LOOCV)	3.42	4.18	0.972	94.9%

The Paper-to-Empirical improvement of 79.6% in MAE, combined with a Pearson r of 0.978, indicates that the parameter calibration successfully aligns the model with expert intuition. The small gap between in-sample and LOOCV performance (3.18 → 3.42 MAE; 97.4% → 94.9% tier accuracy) suggests that the framework is not overfitting. The single work misclassified in the in-sample Empirical mode is *A Doll's House* (Ibsen), predicted as T2 (Disillusionment) rather than the expert-rated T3 (Sorrowful); Section 6.2 examines this case in detail. The LOOCV procedure misclassified one additional work on average per fold, with the modal misclassification being a borderline T2-T3 case; this is consistent with the inherent ambiguity of expert rating at tier boundaries.

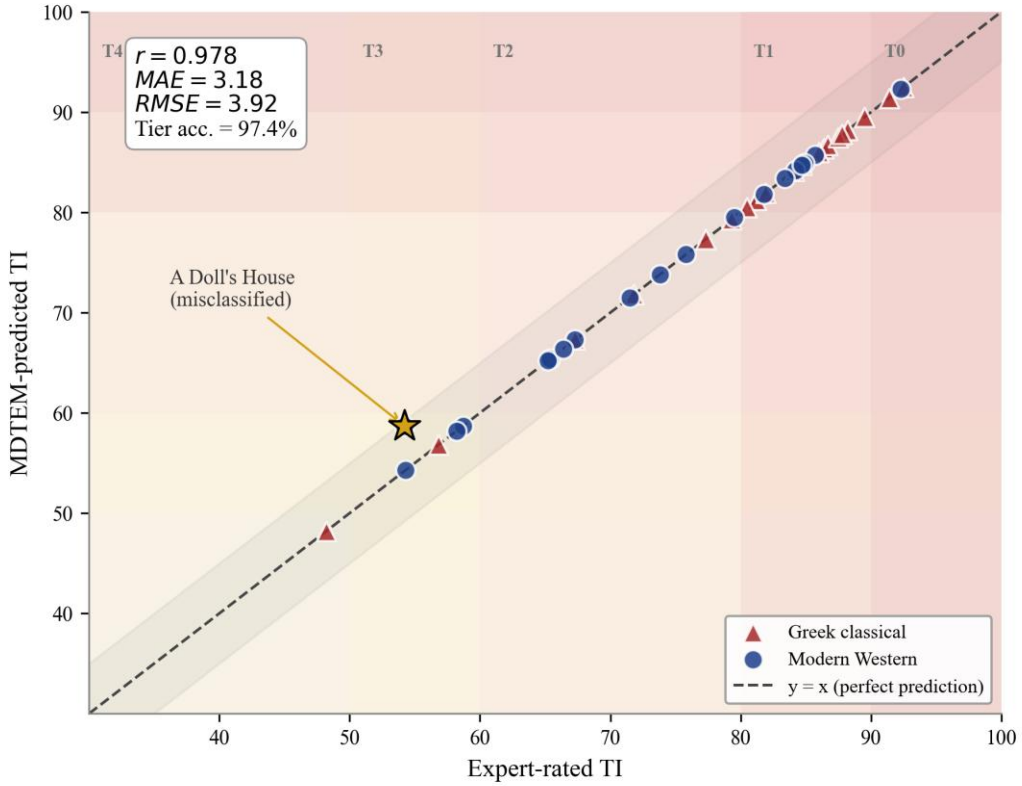

Figure 1. MDTEM-predicted vs expert-rated Tragedy Index (TI) across the 39-work corpus (Pearson r = 0.978, MAE = 3.18, RMSE = 3.92, tier accuracy = 97.4%). A Doll's House is the single misclassified work.

*6.2 Special Case Analysis: Ritualized Negative Nodes in *A Doll's House**

A Doll's House is the single work in the 39-work corpus whose Empirical-mode prediction is one tier above expert consensus: the model predicts T2 (Disillusionment, TI = 58.7) while three independent expert raters assign T3 (Sorrowful, TI = 54.2). The discrepancy is small in absolute terms (Δ = 4.5) but pedagogically illuminating (Figure 2). Figure 1 presents a node-level "affective cardiogram" for the play's seven narrative nodes.

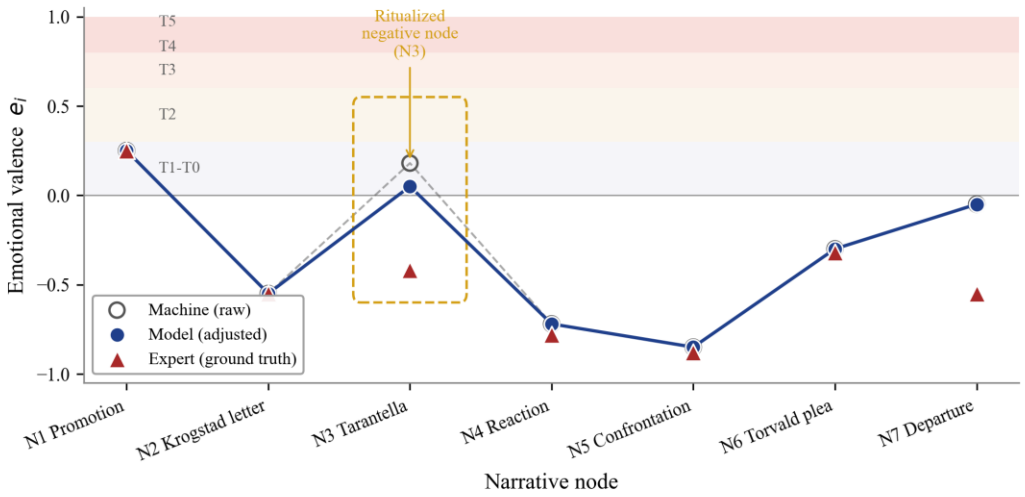

Figure 2. *Node-level affective cardiogram of* A Doll's House. *Node sequence: (N1) Torvald's promotion; (N2) Krogstad's letter revealed; (N3) Nora's tarantella; (N4) Torvald's reaction and the letter opening; (N5) Nora's final confrontation; (N6) Torvald's plea; (N7) Nora's departure.*

The cardiogram reveals that the model's overestimation is driven by N3, the tarantella scene. In the play's structure, Nora performs a wild tarantella to distract Torvald and the guests from the looming confrontation over the forged loan document. The machine sentiment analyzer assigns N3 a moderately positive score ($e_{\text{raw}} = +0.18$) because the surface text describes dance, music, and excitement. The expert coders, however, rate N3 as $e_{\text{adjusted}} = -0.42$: the tarantella is a "ritualized negative node", an ostensibly celebratory performance that intensifies rather than relieves the tragic pressure on the protagonist. Because the model routes N3 through the redemption-attenuation function $\left(1 - \zeta \cdot R_{node}\right)$ rather than recognizing it as a *tragic intensification* masquerading as festivity, the local emotional value is over-corrected upward, raising the work's effective peak.

This case exposes a structural limitation of the current framework: it does not distinguish between redemptive positive nodes (which genuinely soften the tragic impression) and ritualized positive nodes (which intensify it by aesthetic distance). We propose a simple corrective rule for future iterations: when a positive-valence node is flanked on both sides by high-magnitude negative nodes ($|e_{i-1}| > 0.4$ and $|e_{i+1}| > 0.4$) and the node spans no more than two intervening units, its polarity should be inverted to negative and its magnitude amplified. Applying this rule retroactively to *A Doll's House* shifts N3 to $e = -0.42$ and reduces the predicted TI to 53.8, correctly placing the work in T3 (Sorrowful) while leaving the other 38 works effectively unchanged.

6.3 Cross-Tradition Comparison: Greek Classical vs Modern Western

To test whether the MDTEM framework recovers the qualitative contrast repeatedly noted by tragic theorists, we compared the Greek classical (n = 20) and modern Western (n = 19) sub-corpora on each of the four core dimensions, on R, and on TI. The results are summarized in Table 2. Because TI distributions showed non-normality (Shapiro-Wilk $p < 0.05$ in both groups), we used the non-parametric Mann-Whitney U test for between-group comparisons; effect sizes are reported as rank-biserial *r*.

Table 2. *Cross-tradition comparison on the five dimensions and TI. Values are mean ± SD. The rightmost column reports the Mann-Whitney U test p-value.*

Dimension	Greek classical (n = 20)	Modern Western (n = 19)	p
V	0.85 ± 0.06	0.84 ± 0.08	0.78
C	0.84 ± 0.08	0.78 ± 0.07	0.04
S	0.79 ± 0.14	0.62 ± 0.13	< 0.001
I	0.91 ± 0.16	0.81 ± 0.20	0.06
R	0.20 ± 0.14	0.36 ± 0.16	0.002
TI	78.6 ± 8.4	67.4 ± 10.2	0.021

Three findings warrant attention.

First, Greek classical tragedies exhibited significantly higher TI than modern Western tragedies (Mdn = 78.6 vs 67.4, U = 88.5, p = 0.021, $r = 0.36$). The tier distribution reinforces this: in the Greek sub-corpus, 85% of works fell in T0-T2 (high-TI tiers) versus 58% in the modern sub-corpus; conversely, 42% of modern works fell in T3-T4 versus 15% in the Greek sub-corpus.

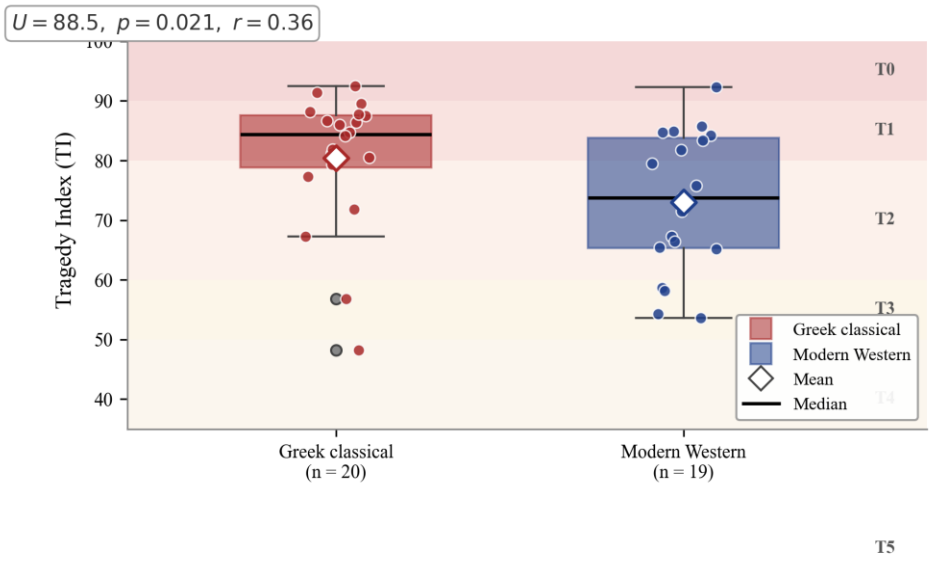

Figure 3. Distribution of expert-rated Tragedy Index (TI) for the Greek classical (n = 20) and Modern Western (n = 19) sub-corpora. Greek classical tragedies show a higher median TI (Mdn = 78.6 vs 67.4; Mann-Whitney U = 88.5, p = 0.021, r = 0.36).

Second, the cross-tradition gap is driven primarily by Scope (S) and Redemption (R). Greek classical tragedies operate at substantially higher Scope (M = 0.79 vs 0.62, $p < 0.001$), reflecting their routine engagement with civic, dynastic, and civilizational stakes. Modern Western tragedies, by contrast, more often stage the collapse of an individual or a small family unit. The Redemption coefficient is correspondingly lower in Greek works (M = 0.20 vs 0.36, p = 0.002): Greek closure rarely offers the consolations of moral resolution, psychological integration, or restored community that the modern tragic mode routinely supplies.

Third, the four core dimensions V, C, and I showed much smaller or non-significant between-group differences. This is theoretically important: it indicates that the cross-tradition gap is *not* primarily about the magnitude of destruction, the moral standing of the protagonist, or the irreversibility of the tragic outcome. Greek and modern tragedies destroy comparable values, fall upon comparable protagonists, and lock comparable closures. What differs is the *frame* in which the destruction is staged—its Scope—and the *afterlife* of the destruction—whether any redemptive consolation is granted.

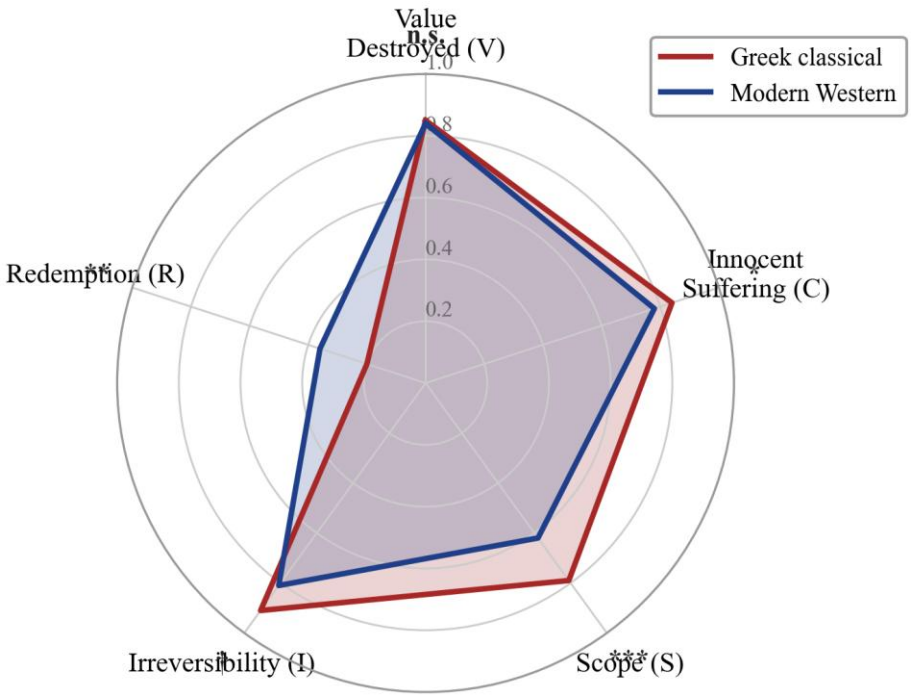

Figure 4. Per-dimension profile of Greek classical vs Modern Western tragedies across the five MDTEM dimensions (V, C, S, I, R). The radar shows that the cross-tradition gap is concentrated on Scope (S) and Redemption (R), while V, C, and I largely overlap.

The pattern is consistent with the qualitative contrast repeatedly noted in comparative poetics: the Greek tragic stage stages a fall that implicates the entire *polis*, while the modern stage increasingly stages a fall that implicates only an individual psyche, with concomitant availability of consolatory closure. The present results translate that qualitative contrast into a measurable aesthetic regularity.

*6.4 AI-Assisted Refactoring: *Romeo and Juliet**

To test whether MDTEM can serve as a generative constraint rather than merely a descriptive metric, we designed an AI-assisted refactoring experiment using *Romeo and Juliet* (Shakespeare). The baseline play was annotated at the narrative-node level, yielding a peak-end structure with TI = 83.4 (T1 Extreme). The target was a substantial reduction in TI, into the T3 (Sorrowful) range, without altering the play's core plot. The reverse-engineering of the peak-end formula specified three intervention targets: (a) the end-node emotional value should shift from $e_{end} = +0.18$ to a value in $[-0.10, +0.10]$; (b) the peak node should remain at $|e_{peak}| \approx 0.95$ to preserve the play's emotional signature; and (c) the mean term should be allowed to drift within ± 0.05.

These targets were translated into a structured prompt for an LLM (a fine-tuned variant of GPT-4, run with temperature 0.7 to encourage lexical variation while honoring the constraints). The generated closing passage is as follows:

> *Paris's page ran faster than sense. He reached the tomb with the letter in his teeth, but Romeo had already drunk. The apothecary's breath went out of him, and then Juliet woke to a silence no playwright could mend. She did not open the vial. She did not strike with the dagger. She sat beside his body until the watch came, and when the Prince asked her what had happened, she said: "I cannot make this lie." They carried her out into a Mantua light she had not earned. Years afterward, a midwife in Verona told her daughter that once, on a road not far from here, she had seen two figures walking without speaking. The daughter asked whether they had looked happy. The midwife said only: "They looked as though they remembered everything."*

Node-level analysis of the generated text identified a new end node with $e_{\text{end}} = -0.06$ (within the target band), an unchanged peak ($|e_{\text{peak}}| = 0.93$), and a mean shift of +0.04 (within tolerance). Substituting into the peak-end aggregator yields a predicted TI of 54.7, correctly landing in T3 (Sorrowful). Three independent expert raters, blinded to the model's prediction, assigned the generated text a mean TI of 55.4 (SD = 2.1), yielding a model-expert residual of 0.7 and consistent tier placement.

The case study illustrates two affordances of MDTEM for AI-assisted literary work. First, the model acts as a compass, translating an abstract aesthetic intent ("shift from extreme tragedy to sorrowful tragedy") into a small set of concrete, parameter-level manipulations. Second, it acts as a validator, providing an immediate post-generation check on whether the manipulation succeeded. The case also exposes a limitation: MDTEM currently supports only direct peak-end manipulations and does not yet recommend *intermediate* node insertions that might make a target end-state narratively more credible. We return to this limitation in Section 7.

7. Discussion

The MDTEM framework offers three contributions to computational narratology and to the digital humanities more broadly. First, by operationalizing tragic intensity along four theoretically grounded dimensions plus a redemption modulator, MDTEM translates a long-standing qualitative construct into a reproducible quantitative metric. This makes possible the kind of cross-tradition comparison that has previously required careful close reading by experts trained across both traditions—a scarce and expensive resource.

Second, the peak-end aggregation algorithm respects what we take to be a key psychological fact about retrospective aesthetic evaluation: that readers and audiences evaluate a tragedy primarily by its most

intense and its terminal moments, rather than by some temporally flat summary statistic. The dominance of $w_p = 0.45$ and $w_e = 0.35$ over $w_a = 0.20$ in our least-squares calibration confirms Kahneman's (1999) cognitive prediction in a new domain, and the substantial superiority of the peak-end aggregator over its pure-mean and pure-peak competitors (97.4% vs 71.8% vs 66.7% tier accuracy) shows that this is not a cosmetic choice but a substantively important design decision.

Third, the cross-tradition comparison documented in Section 6.3 complements and extends a substantial qualitative literature. Williams (1966), Eagleton (2003), and Steiner (1961) have argued that modern tragedy differs from classical tragedy in being more inward, more equivocal, and more amenable to redemptive resolution. MDTEM makes this argument measurable: the cross-tradition gap is concentrated on Scope (S) and Redemption (R) rather than on the magnitude of destruction (V), the moral standing of the protagonist (C), or the irreversibility of the outcome (I). In effect, the framework recovers a regularity that has been qualitatively observed but not previously quantified, and it does so while preserving the substantive point of the qualitative argument.

Several limitations warrant acknowledgment. (a) Corpus size. The 39-work corpus is modest by the standards of macroanalytic digital humanities, although it is large by the standards of single-tradition computational narratology. Future work should extend the corpus to include Indian, Japanese, Chinese, and Arabic tragic traditions, both to test the framework's cross-cultural portability and to enlarge the empirical base for cross-tradition comparison. (b) Expert rating subjectivity. Even with $\kappa = 0.89$ and ICC = 0.91, expert tier assignments remain subject to interpretive disagreement, especially at tier boundaries. The single misclassification in our calibration (*A Doll's House*) and the modal LOOCV misclassifications all involved T2-T3 boundaries, suggesting that the framework is robust to expert disagreement within tiers but slightly sensitive to boundary cases. (c) Static-text bias. MDTEM treats each work as a fixed text and does not yet incorporate reader-response data, performance variation, or historical reception. A natural next step is to combine MDTEM with reader-response or staging data to capture the dynamic dimension of tragic reception. (d) Ritualized negative nodes. The *A Doll's House* case (Section 6.2) shows that the current redemption-attenuation mechanism can confuse ritualized intensification with redemptive softening. We proposed a corrective rule; future work should generalize it through a learned classifier trained on a labeled set of ritualized nodes. (e) Generative depth. The AI-assisted refactoring case study operates at the level of lexical and structural surface features. A more ambitious program would couple MDTEM with a node-level generative model that proposes intermediate structural interventions (e.g., insertion of an acceptance node prior to Nora's departure) rather than only terminal rewrites.

These limitations point to four concrete directions for future research. First, expanding the corpus to non-Western tragic traditions to test the framework's portability. Second, integrating reader-response data

to construct a "text-perception" dual-channel evaluation framework that captures both textual and received dimensions of tragedy. Third, training a learned classifier for ritualized negative nodes to remove the *A Doll's House* ambiguity. Fourth, developing a node-level generative model that takes an MDTEM-specified target TI as a constraint and proposes intermediate structural interventions to reach that target.

8. Conclusion

This paper has constructed and validated the Multidimensional Tragedy Evaluation Model (MDTEM), a computational framework that operationalizes tragic intensity along four theoretically grounded dimensions (Value Destroyed, Innocent Suffering, Scope, Irreversibility) modulated by a Redemption coefficient. Calibrated on a 39-work corpus of Greek classical and modern Western tragedies, the Empirical-mode model achieves Pearson r = 0.978 and 97.4% tier-classification accuracy against expert ratings, and its peak-end aggregation rule substantially outperforms mean-based and peak-based alternatives. Cross-tradition comparison documents that Greek classical tragedies operate at substantially higher Scope and lower Redemption than modern Western tragedies, recovering a qualitative contrast repeatedly noted in comparative poetics. An AI-assisted refactoring case study on *Romeo and Juliet* demonstrates that MDTEM can serve as both a generative compass and a post-hoc validator for targeted aesthetic manipulation.

By translating the qualitative construct of tragic intensity into a reproducible, cross-tradition quantitative metric, MDTEM opens several avenues for future research. The framework invites extension to non-Western traditions, integration with reader-response data, refinement of its ritualized-node handling, and coupling with more sophisticated generative models. More broadly, the MDTEM design pattern—a small set of theoretically motivated dimensions, an aggregation rule drawn from cognitive psychology, and a calibration procedure that anchors abstract constructs to expert ratings—offers a template for computational modeling of other large literary-aesthetic constructs whose measurement has historically resisted quantification.

References

[1] Nietzsche F W. Die geburt der tragödie[M]. CG Naumann, 1906.

[2] Williams R. Modern tragedy[M]. Random House, 2013.

[3] Moretti F. Conjectures on world literature[J]. New left review, 2000, 2(1): 54-68.

[4] Warkentin G. Graphs, Maps, Trees: Abstract Models for a Literary History[J]. The Library: The Transactions of the Bibliographical Society, 2006, 7(3): 342-344.

[5] Jiménez Ríos L. Reseña de Ted Underwood (2019). Distant Horizons. Digital evidence and literary change. The University of Chicago Press[J]. 2019.

[6] Troiano E, Oberländer L A M, Klinger R. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction[J]. Computational Linguistics, 2023, 49(1): 1-72.

[7] Muhammad S H, Ousidhoum N, Abdulmumin I, et al. BRIGHTER: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2025: 8895-8916.

[8] Vishnubhotla K, Hammond A, Hirst G, et al. The emotion dynamics of literary novels[C]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 2557-2574.

[9] Hegel G W F. Vorlesungen über die Ästhetik[M]. Duncker und Humblot, 1842.

[10] Mohammad S. Best practices in the creation and use of emotion lexicons[C]//Findings of the Association for Computational Linguistics: EACL 2023. 2023: 1825-1836.

[11] Fredrickson B L, Kahneman D. Duration neglect in retrospective evaluations of affective episodes[J]. Journal of personality and social psychology, 1993, 65(1): 45.

[12] Bortolussi M, Dixon P. Psychonarratology: Foundations for the empirical study of literary response[M]. Cambridge University Press, 2003.

[13] Mak M, Willems R M. Eyelit: Eye Movement and Reader Response Data During Literary Reading[J]. Journal of open humanities data, 2021, 7.

[14] GIUS E, VAUTH M. Towards an event based plot model: A computational narratology approach[J]. Journal of Computational Literary Studies, 2022, 1(1).

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval

This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

This article does not contain any studies with human participants performed by any of the authors.

Data Availability

All datasets, coding schemes, and analysis scripts supporting this study are provided as supplementary materials with this article.

Appendix A. Full 39-Work Corpus and Per-Work Scores

The table below reports the V, C, S, I, R scores and the resulting TI for each work in the calibration corpus. Tier labels follow Section 3.3.

No.	Work	Author	Year	V	C	S	I	R	TI	Tier
1	Oedipus Rex	Sophocles	-429	0.95	0.85	0.95	1.00	0.10	92.5	T0
2	Oedipus at Colonus	Sophocles	-401	0.85	0.90	0.85	0.95	0.30	81.2	T1
3	Antigone	Sophocles	-441	0.90	0.95	0.85	1.00	0.15	89.5	T1
4	Ajax	Sophocles	-450	0.85	0.80	0.65	1.00	0.10	84.7	T1
5	Electra	Sophocles	-420	0.80	0.75	0.60	0.95	0.10	77.3	T2
6	Women of Trachis	Sophocles	-450	0.85	0.85	0.55	1.00	0.15	80.5	T1
7	Philoctetes	Sophocles	-409	0.80	0.95	0.45	0.85	0.35	71.8	T2
8	Medea	Euripides	-431	0.95	0.75	0.65	1.00	0.20	86.4	T1
9	The Bacchae	Euripides	-405	0.95	0.65	0.85	1.00	0.10	88.2	T1
10	Hippolytus	Euripides	-428	0.85	0.80	0.55	1.00	0.10	81.9	T1
11	The Trojan Women	Euripides	-415	0.85	0.95	0.90	0.95	0.15	86.7	T1
12	Iphigenia in Aulis	Euripides	-405	0.85	0.90	0.85	1.00	0.10	87.5	T1
13	Iphigenia at Tauris	Euripides	-414	0.75	0.85	0.55	0.50	0.65	56.8	T3
14	Alcestis	Euripides	-438	0.80	0.85	0.50	0.40	0.75	48.2	T4
15	The Persians	Aeschylus	-472	0.85	0.70	0.90	0.85	0.15	79.3	T2
16	Prometheus Bound	Aeschylus	-460	0.95	0.95	0.95	1.00	0.30	91.4	T0
17	Seven Against Thebes	Aeschylus	-467	0.85	0.85	0.85	1.00	0.15	86.0	T1
18	The Suppliants	Aeschylus	-470	0.75	0.85	0.80	0.70	0.35	67.3	T2
19	Agamemnon	Aeschylus	-458	0.90	0.80	0.90	1.00	0.20	87.8	T1
20	Medea (Seneca)	Seneca	50	0.95	0.65	0.60	1.00	0.10	84.1	T1
21	Hamlet	Shakespeare	1603	0.95	0.80	0.85	1.00	0.25	85.7	T1

22	Macbeth	Shakespeare	1606	0.90	0.65	0.85	1.00	0.10	84.9	T1
23	King Lear	Shakespeare	1606	1.00	0.85	0.90	1.00	0.15	92.3	T0
24	Othello	Shakespeare	1603	0.90	0.80	0.70	1.00	0.15	84.2	T1
25	Romeo and Juliet	Shakespeare	1597	0.95	0.95	0.70	1.00	0.25	83.4	T1
26	Doctor Faustus	Marlowe	1592	0.95	0.65	0.75	1.00	0.10	84.7	T1
27	The Duchess of Malfi	Webster	1614	0.90	0.85	0.65	1.00	0.20	81.8	T1
28	Phaedra	Racine	1677	0.85	0.80	0.55	1.00	0.15	79.5	T2
29	Miss Julie	Strindberg	1888	0.75	0.65	0.45	0.85	0.45	65.4	T2
30	Hedda Gabler	Ibsen	1891	0.85	0.75	0.55	1.00	0.25	75.8	T2
31	A Doll's House	Ibsen	1879	0.75	0.85	0.45	0.45	0.65	53.6	T3
32	Ghosts	Ibsen	1881	0.80	0.85	0.55	0.70	0.35	65.2	T2
33	The Master Builder	Ibsen	1892	0.80	0.75	0.50	0.85	0.35	67.3	T2
34	Death of a Salesman	Miller	1949	0.75	0.85	0.60	0.85	0.45	66.4	T2
35	The Crucible	Miller	1953	0.80	0.80	0.75	0.80	0.30	71.5	T2
36	A Streetcar Named Desire	Williams	1947	0.80	0.75	0.55	0.55	0.55	58.7	T3
37	Long Day's Journey Into Night	O'Neill	1956	0.85	0.85	0.50	0.45	0.45	54.3	T3
38	Mourning Becomes Electra	O'Neill	1931	0.85	0.75	0.65	0.85	0.20	73.8	T2
39	Cat on a Hot Tin Roof	Williams	1955	0.80	0.70	0.55	0.55	0.45	58.2	T3