Probing Stylistic Appropriation using Large Language Models: An Evaluation Framework for Copyright Infringement under EU Law

Noah Scharrenberg[1,2] and Chang Sun[1*]

[1*]Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands.
[2]Contractuo, Eindhoven, The Netherlands.

*Corresponding author(s). E-mail(s): chang.sun@maastrichtuniversity.nl;
Contributing authors: noah.scharrenberg@maastrichtuniversity.nl;

Abstract

Large language models (LLM) trained on web-scale corpora generate output that may infringe copyright, yet existing technical safeguards focus narrowly on verbatim memorisation. EU copyright doctrine applies a broader standards: substantial similarity, which extends to stylistic choices, narrative structure, and creative elaboration. This mismatch between what current methods detect and what the law protects leaves a significant compliance gap. We introduce PSALM, an LLM-as-a-judge framework that operationalises EU copyright doctrine through ten evaluators assessing computational overlap, stylistic dimensions (writing style, narrative voice), content dimensions (character, plot, scene, world building), and statutory exceptions (parody, pastiche, quotation, scènes à faire). Applying PSALM to Llama 3.2 models fine-tuned on translated historical Dutch literary works, we find that: 1) instruction-tuned models exhibit non-trivial baseline stylistic similarity prior to corpus exposure; 2) fine-tuning induces systematic stylistic appropriation across all infringement-relevant dimensions, extending beyond verbatim memorisation to abstract narrative patterns; 3) Negative Preference Optimisation unlearning substantially reduces similarity but leaves detectable residual stylistic patterns. These findings indicate that safeguards targeting literal copying alone are insufficient to mitigate broader copyright risks. PSALM provides infrastructure for auditable, legally informed compliance evaluation, though the relationship between automated similarity scores and infringement determinations requires validation by legal experts. This

work bridges qualitative legal standards and quantitative technical measurement, exposing fundamental tensions between generative AI and EU intellectual property law.

Keywords: EU Copyright, LLM-as-a-Judge, Unlearning, LLM Evaluation, PSALM

1 Introduction

The rapid advancement of Large Language Models (LLMs) has sparked a fundamental tension between technological capability and legal compliance. Recent demonstrations have shown that current LLMs can reproduce substantial portions of copyrighted works with minimal prompting (Cooper et al. 2025), raising urgent questions about the suitability of existing copyright frameworks. However, current technical approaches to LLM copyright compliance focus predominantly on verbatim memorisation, literal copying of training data (Carlini et al. 2021; Cooper et al. 2025; Ippolito et al. 2023), with evaluation benchmarks focusing on exact-match and high-overlap metrics (Maini et al. 2024; Shi et al. 2025; Wei et al. 2024), while overlooking more subtle forms of infringement that may be equally consequential under copyright law (Chen et al. 2024).

European copyright law, particularly codified in the Copyright in the Digital Single Market (CDSM) Directive and the Artificial Intelligence (AI) Act, establishes copyright protection based on originality as "the author's own intellectual creation" (European Parliament and Council 2019, 2024; Court of Justice of the European Union 2009, 2019a). Infringement determinations extend beyond literal copying to encompass substantial similarity in expressive elements, including narrative structure, character development, stylistic choices, and the selection and arrangement of creative elements (Lucchi and Hunter 2025). Yet existing technical safeguards, including deduplication filters, similarity thresholds, and machine unlearning methods, are calibrated primarily to detect and suppress verbatim reproduction, leaving a critical gap in compliance verification for stylistic and structural appropriation.

This gap reflects a deeper epistemic misalignment between the legal and technical evaluation paradigms. Copyright law operates through qualitative, context-dependent standards adjudicated retrospectively by courts, weighing factors such as the overall impression, expressive overlap, and market substitution effects. In contrast, LLM development relies on quantitative, automated metrics (such as ROUGE-L, cosine similarity, and edit distance) that can be embedded into continuous integration pipelines but fail to capture the legally significant dimensions of substantial similarity (Chen et al. 2024; Wei et al. 2024). Consequently, models may pass technical benchmarks while producing outputs that appropriate protected stylistic elements, plot structures, or character archetypes without triggering existing safeguards.

Existing LLM copyright-compliance mechanisms primarily address lawful data ingestion, transparency, and literal memorisation: provenance systems document training sources, unlearning methods suppress known memorised content, and disclosure regimes make dataset use more visible (Bommarito II et al. 2025; Russinovich and

Salem 2025; Wei et al. 2024; Warso and Gahntz 2024). These mechanisms are necessary but insufficient because EU copyright infringement can turn on substantial similarity in protectable expression, not only on exact or near-exact textual reproduction (Court of Justice of the European Union 2009, 2019a; Lucchi and Hunter 2025). Consequently, a model may pass overlap-based memorisation tests while still producing outputs that closely track protected expressive choices such as narrative voice, characterisation, plot architecture, or the selection and arrangement of incidents (Chen et al. 2024; Chun 2024).

The central challenge is therefore methodological: how can LLM outputs be evaluated at scale for legally relevant expressive similarity when the relevant similarities are qualitative, context-dependent, and often non-verbatim? Existing computational metrics, including n-gram overlap, edit distance, longest common substring, and embedding similarity, are useful for detecting literal or semantic correspondence but are poorly suited to assessing stylistic and structural appropriation (Chen et al. 2024; Wei et al. 2024). Manual review by legal experts or trained annotators can capture these subtleties, but it is too slow and costly to support continuous evaluation of deployed LLM systems. This creates a compliance verification gap: copyright law requires assessment of substantial similarity in protectable expression, while current technical pipelines primarily measure surface-form overlap or semantic relatedness.

This paper addresses this deadlock by proposing **PSALM** (Probing Stylistic Appropriation by Language Models), an automated evaluation framework implementing EU copyright doctrine as measurable assessments in three steps. First, it identifies legally relevant dimensions where courts recognise protectable expression: narrative voice, character development, plot structure, world-building, and scene sequencing, derived from Court of Justice of the European Union (CJEU) jurisprudence establishing the originality standard (Court of Justice of the European Union 2009, 2019a), scholarly analysis of AI copyright issues (Lucchi and Hunter 2025; Quintais 2025; Borhi et al. 2025), and narratological theory characterising distinctive authorial choices (Genette 1980; Booth 1961; Palmer 2004). Second, it decomposes each dimension into evaluable sub-dimensions grounded in legal doctrine (e.g., narrative voice $\rightarrow$ point of view, narrative distance, focalisation patterns). Third, it employs LLM-as-a-judge evaluation to assess similarity along these sub-dimensions through structured prompts encoding legal tests, with hierarchical aggregation via directed acyclic graphs (DAG) synthesising evidence through weighted judgement nodes that tries to mirror how courts could potentially weigh multiple factors in infringement determinations. This approach enables scalable qualitative assessment: the dimensional structure could capture legally relevant expressive features, while LLM judges approximate the context-dependent reasoning human evaluators perform.

By bridging the gap between retrospective legal adjudication and proactive technical control, this work advances toward auditable, legally meaningful copyright compliance at machine scale. In particular, we conduct experiments to answer the following research questions:

Research questions.

- **RQ1:** To what extent do base language models exhibit baseline stylistic similarity to protected works across legally relevant dimensions as operationalised by PSALM?
- **RQ2:** Does supervised fine-tuning on a corpus of literary question–answer pairs increase PSALM-assessed stylistic appropriation beyond effects attributable to verbatim memorisation?
- **RQ3:** Does machine unlearning via Negative Preference Optimisation reduce PSALM-assessed stylistic appropriation across copyright-relevant dimensions, or does it primarily suppress literal recall?

Our findings have implications for model developers seeking to demonstrate due diligence, for policymakers crafting enforceable standards, and for rightsholders seeking reliable mechanisms to protect their works in the eras of generative AI.

2 Related Work

2.1 Copyright Law and Large Language Models

The EU legal framework, anchored by the CDSM Directive and the AI Act, establishes a civil-law approach grounded in defined exceptions—rather than the flexible "fair use" doctrine prevalent in common-law jurisdictions—including mandatory text-and-data mining (TDM) exceptions under Articles 3 and 4, permitting reproduction of lawfully accessible works for computational analysis subject to opt-out rights exercised "in an appropriate manner" by rights holders (European Parliament and Council 2019, 2024). However, as Quintais (2025) and Margoni and Kretschmer (2022) observed, core terms such as "lawful access" remain ambiguous in the context of automated web scraping, creating compliance uncertainty for developers and enforcement challenges for rightsholders. Scharrenberg and Sun (2025) identified this as symptomatic of a deeper epistemic misalignment between the qualitative, context-dependent standards of the European copyright law and the quantitative, proactive metrics of LLM development.

Prior litigation has exposed the fragility of these frameworks. The Hamburg District Court's decision in *Kneschke v. LAION* found that inclusion of a photographer's images in a large-scale dataset without explicit consent violated copyright protections, notwithstanding the dataset creators' invocation of TDM exceptions (Hamburg District Court 2024; Havlikova 2025); a concurrent EUIPO study confirmed that petabyte-scale ingestion renders individualised consent or verification practically infeasible (Borhi et al. 2025). The Munich Regional Court's preliminary judgement in *GEMA v. OpenAI* adds a further dimension: the court held that memorisation of song lyrics within model weights constitutes an act of reproduction under Art. 2 of the InfoSoc Directive and § 16 UrhG, regardless of whether the stored representation is explicit or encoded as probability parameters, and that such memorisation falls outside the TDM exception of Art. 4 CDSM because it goes beyond the preparatory acts of data ingestion and directly engages the rightsholder's exploitation rights (Kallert 2025). The court further attributed responsibility to the model operator rather than the end-user, on the basis that the operator controls the architecture and training data

determining what the model reproduces. At the legislative level, the European Parliament's Committee on Legal Affairs recently adopted a report calling for mandatory transparency obligations, including itemised lists of copyrighted works used in training, and for EU copyright law to apply to all GPAI systems placed on the EU market irrespective of where training occurs, with non-compliance with transparency requirements potentially constituting infringement in itself (Axel Voss 2026). Together, these developments suggest that compliance cannot be established from ingestion governance, transparency, or overlap-based memorisation checks alone; rather, they point to the need for evaluations that target protectable expressive choices assessed through substantial-similarity reasoning (Scharrenberg and Sun 2025).

2.2 Memorisation and Extraction in Language Models

Memorisation in LLMs is well documented: Carlini et al. (2021) demonstrated that GPT-2 can be prompted to emit memorised sequences, particularly when training data contain repeated examples. Subsequent work has shown that memorisation correlates with data frequency, model capacity, and the presence of duplicates in training corpora (Lee et al. 2022; Kandpal et al. 2022).

Recent research has extended these findings to current models and specialised domains. A recent study demonstrated that Meta's LlaMA 3.1 could reproduce entire chapters of Harry Potter nearly verbatim through carefully constructed prompts, revealing vulnerabilities even in models subjected to alignment training (Cooper et al. 2025). Another study introduced CopyBench, a benchmark for measuring both literal and non-literal reproduction of copyrighted text, finding that models exhibit varying degrees of memorisation depending on prompt specificity and output length (Chen et al. 2024).

However, existing memorisation studies focus predominantly on verbatim or near-verbatim extraction, with detection methods relying on n-gram overlap, longest common substring (LCS) metrics, or embedding-based similarity measures that capture literal correspondence but struggle to identify structural or stylistic parallels (Ippolito et al. 2023; Shi et al. 2025). This limitation is significant because copyright infringement does not require word-for-word copying, substantial similarity in expressive elements (such as narrative structure, character dynamics or stylistic choices) can constitute infringement even when no literal overlap exists (Lucchi and Hunter 2025).

2.3 Substantial Similarity and Derivative Works

According to the EU copyright doctrine, infringement requires substantial similarity in protectable expressive elements—plot structure, character development, narrative voice, and the selection and arrangement of incidents—rather than verbatim correspondence alone (European Parliament and Council 2001; Court of Justice of the European Union 2009; Lucchi and Hunter 2025; Chun 2024).

AIStorySimilarity (Chun 2024)—the closest methodological antecedent to PSALM—decomposes narrative comparison into plot, character, and thematic dimensions using embedding-based semantic similarity, graph-structured event analysis,

and transformer-based contextual matching. It achieves $r = 0.72$ to $r = 0.85$ with human annotations across plot and character dimensions, demonstrating that automated methods can approximate human judgement on structural narrative parallels. PSALM adopts a similar hierarchical decomposition strategy but diverges from AIStorySimilarity in three fundamental respects.

First, **legal grounding**: AIStorySimilarity measures narrative similarity as a descriptive construct relevant to human perception and information retrieval; PSALM instead operationalises legally-defined constructs grounded in EU copyright doctrine—originality as "author's own intellectual creation", the idea-expression dichotomy, and substantial similarity tests—whose dimensions reflect protectability criteria established through CJEU jurisprudence. Second, **exception integration**: AIStorySimilarity measures *how* similar works are without assessing whether similarity is actionable. PSALM integrates defence evaluators implementing parody (CJEU *Deckmyn* test), pastiche (AG Emiliou criteria), quotation (Art. 5(3)(d) requirements), and scènes à faire doctrine, recognising that high similarity scores may be legally permissible under specific conditions (Court of Justice of the European Union 2014, 2025; European Parliament and Council 2001). Third, **stylistic versus semantic focus**: AIStorySimilarity prioritises semantic content similarity—whether events, characters, and themes align in meaning. PSALM prioritises stylistic and structural similarity—whether expression choices and narrative techniques align—using LLM-as-judge evaluation to capture qualitative patterns resistant to embedding-based methods. Two texts may describe semantically identical events yet exhibit entirely different stylistic execution; the former would score as highly similar under AIStorySimilarity whilst PSALM would score them as dissimilar on writing style and narrative voice. This distinction is legally consequential: copyright protects expression rather than ideas.

2.4 Machine Unlearning for LLMs

Machine unlearning (MU) has emerged as a mechanism to remove or suppress specific knowledge from trained models, driven by copyright takedown obligations and GDPR's right to erasure (Russinovich and Salem 2025; Wei et al. 2024). However, the effectiveness of unlearning in mitigating copyright risk depends on whether it addresses only literal memorisation or also disrupts abstract stylistic patterns that may constitute substantial similarity (Scharrenberg and Sun 2025); LLMs' non-convex architectures preclude exact removal, necessitating approximate techniques (Chien et al. 2024).

Current approaches include gradient-based methods that reduce model likelihood on targeted sequences (Zhang et al. 2024; Russinovich and Salem 2025; jin et al. 2024), self-distillation variants that preserve general utility whilst forgetting specific content (Dong et al. 2025; Vasilev et al. 2025), and inference-time controls applying learned refusal policies or logit adjustments without parameter modification (Bhaila et al. 2025; Ji et al. 2024). Wei et al. (2024) introduced a comprehensive evaluation suite, assessing both forget quality and model utility retention.

However, existing unlearning benchmarks concentrate on verbatim memorisation, evaluating success primarily through exact-match or high-overlap metrics (Maini et al.

2024; Shi et al. 2025). This focus leaves two critical gaps unaddressed. First, unlearning methods remain vulnerable to paraphrase extraction and adversarial prompting strategies that elicit memorized content through indirect means (Wei et al. 2024). Second, current evaluations do not assess whether unlearning mitigates stylistic appropriation. Compounding both gaps, Jia et al. (2025) showed that standard unlearning metrics systematically overestimate forgetting success: evaluations on the explicit forget set may certify unlearning whilst the model retains reproductive capacity over semantically derived content (e.g., fan fiction based on a copyrighted sources work), revealing a fundamental mismatch between benchmark coverage and the knowledge-level erasure copyright expectations demand.

2.5 Automated Evaluation of LLM Outputs

Standard reference-based metrics (BLEU, ROUGE, METEOR) and reference-free alternatives correlate poorly with human judgements for creative or stylistically nuanced generation (Novikova et al. 2017; Sai et al. 2022; Gehrmann et al. 2021). Recent work has demonstrated that LLMs themselves can serve as effective evaluators for complex, subjective properties: Zheng et al. (2023) showed that GPT-4 achieves agreement with human annotators exceeding 80% on multi-dimensional response evaluation, and carefully designed prompts can elicit reliable comparative judgements even for nuanced attributes such as helpfulness and harmlessness (Chiang and Lee 2023).

LLM-as-judge methodologies show particular promise in legal domains, where evaluation requires contextual interpretation of rule-based frameworks and qualitative assessment of multi-factor tests. Cui et al. (2024) found that GPT-3.5 and GPT-4 can identify logical fallacies, assess argument strength, and detect missing premises with moderate reliability ($F_1 = 0.68$ to 0.74) when supplied with structured criteria grounded in legal doctrine, and that explicit definitional grounding—providing formal definitions of concepts such as burden of proof and material fact—significantly improves classification consistency. Relatedly, Katz et al. (2024) demonstrated that GPT-4 scores in the 90^{th} percentile on the Uniform Bar Examination, confirming capability for doctrinal rule application and legal argumentation of the kind required for copyright assessment.

These studies establish that LLM-as-judge reliability depends on three design features: (1)**explicit doctrinal grounding** through prompts encoding legal definitions and multi-factor tests; (2)**hierarchical decomposition** of complex legal determinations into evaluable sub-questions; (3)**structured output formats** constraining models to categorical verdicts with justifications(Cui et al. 2024; Katz et al. 2024). PSALM incorporates all three, extending prior binary or scalar legal LLM applications to coordinated multi-dimensional assessment across thirteen heterogeneous metrics.

2.6 Research Gap

This literature reveals three research gaps. First, evaluation frameworks remain anchored to verbatim similarity metrics that poorly approximate the totality-of-expressive-overlap tests applied in legal adjudication. Second, unlearning methods are validated for literal forgetting but not for the suppression of stylistic and structural

appropriation; the ”erasure illusion” problem compounds this by systematically overstating forgetting success. Third, automated evaluation lacks the doctrinal grounding required for copyright compliance assessment, including recognition of transformative uses and statutory exceptions.

This work addresses these gaps through a legally informed evaluation framework assessing stylistic copyright infringement across the LLM lifecycle, including base models, fine-tuning, and unlearning, implementing EU substantial similarity tests and integrating recognised defences, while acknowledging that automated scores inform rather than replace human legal judgement (Scharrenberg and Sun 2025).

3 Methods

PSALM (Probing Stylistic Appropriation by Language Models) operationalises elements of EU copyright doctrine as measurable assessments over source–target text pairs. Each evaluator is implemented as a directed acyclic graph (DAG) of textual analysis tasks adjudicated by an LLM-as-judge, producing scores along two categories of legally relevant dimensions: infringement-oriented (Section 3.3), measuring similarity across protectable expressive features, and defence-oriented (Section 3.4), assessing whether observed similarities qualify for statutory exceptions or fall within unprotectable elements.

3.1 Directed Acyclic Graph-based LLM-as-a-Judge Architecture

Each evaluator's DAG comprises three node types: task nodes, a single judgement node, and verdict nodes (visualised in Figures 1–10).

Task nodes correspond to sub-dimensions of a legal construct. A task node takes as input the source text x_s, x_t, and an instruction that specifies which aspects of the texts to compare and how to classify the comparison using a fixed five-level scale. The instruction is expressed in natural language and includes three elements: a definition of the sub-dimension, a list of concrete factors to examine in each text, and a request to assign the sub-dimension to one of five verbal categories. For example, the “Lexical Complexity Analysis” task in the writing style evaluator (Section 3.3.1) asks the judge model to analyse vocabulary richness, word length patterns, and formality in both texts, to ignore semantic content, and then to classify the similarity as “identical or near-identical writing style across all dimensions”, “very similar writing style with only minor differences”, “moderately similar writing style with some notable differences”, “somewhat different writing style with limited similarities”, or “clearly different or opposite writing styles”.

The judgement node aggregates the sub-dimensions. Its instruction restates the overall construct (for example: “overall writing style similarity between x_s and x_t”), lists the outputs of the relevant task nodes by name, and specifies explicit weights for each sub-dimension as percentages. The weights aim to encode which aspects are treated as more salient in legal (and narratological) terms. The judge model is instructed to take these weights into account and to select a single verdict, from a

separate set of five verdict nodes, together with a justification that explains how the sub-dimension analyses support the chosen level.

For infringement evaluators, the five verdicts correspond to scores $\{0, 3, 5, 8, 10\}$, mapped to the normalised set $\{0.0, 0.3, 0.5, 0.8, 1.0\}$. These represent, respectively, "clearly different or opposite", "somewhat different with limited similarities", "moderately similar with some notable differences", "very similar with only minor differences", and "identical or near-identical". For defence evaluators, the same numerical scale is reused, but verdict labels reflect defence strength (of genericness) rather than similarity, for example "strong parody (10)" through to "no parody (0)".

The use of a coarse, ordered five-band rubric is motivated by prior work showing that discretisations yield more consistent LLM judgements than fine-grained numerical scoring (Zheng et al. 2023; Chiang and Lee 2023). Within this ordinal rubric (Stevens 1946; Norman 2010), we chose the particular spacing $\{0, 3, 5, 8, 10\}$ to encode a monotone but intentionally non-linear notion of risk contribution for aggregation: shifting into (or out of) the extreme bands is treated as more consequential than moving within the middle bands. Concretely, the endpoints (0 and 10) anchor the lowest- and highest-risk categories, while the intermediate values $(3, 5, 8)$ compress the middle region and expand the tails, so that aggregated scores respond more strongly when judgements cross into near-identical or clearly-different ranges. This is a heuristic design for comparing and aggregating LLM-judge outputs across multiple tasks and evaluators; it is not a claim that rubric categories are separated by equal or empirically calibrated quantitative distances.

Formally, each evaluator E implements a function:

$$E(x_s, x_t) = s \in \{0.0, 0.3, 0.5, 0.8, 1.0\} \tag{1}$$

where x_s and x_t are arbitrary source and target texts, respectively. Evaluation proceeds by presenting each task node's instruction and the two texts to the judge model, obtaining textual analyses and local categorical labels, and then providing these task outputs to the judgement node, which selects one of the five verdicts. The final numerical score is the normalised value associated with that verdict. Sub-dimension labels remain part of the decision-making trace but only the final evaluator-level score is used as a quantitative output. Any sufficiently capable instruction-following LLM may serve as the judge model, though performance varies by model.

3.2 Foundational Legal Principles

Evaluator design draws on two bodies of work: EU copyright law and doctrinal commentary—the CJEU originality standard (Court of Justice of the European Union 2009, 2019a; Rosati 2013), InfoSoc and CDSM exceptions (European Parliament and Council 2001, 2019), and generative AI copyright analyses (Lucchi and Hunter 2025; Quintais 2025; Borhi et al. 2025)—and narratology, stylistics, and authorship attribution, which supply operational definitions of style, voice, character, plot, scene, and world-building (Stamatatos 2009; Juola 2008; Genette 1980; Booth 1961; Palmer 2004; Herrmann et al. 2015; Abbott 2008; Ryan 2007; Bordwell and Thompson 2012; McKee 1997; Wolf 2012; Ekman 2013; Gavins 2007; Chatman 1990; Herman et al. 2012).

Infringement evaluators target expressive dimensions that CJEU originality doctrine treats as protectable through the author's "free and creative choices"(Court of Justice of the European Union 2009, 2019a)—narrative voice, characterisation, plot architecture, and stylistic elaboration(Lucchi and Hunter 2025; Borhi et al. 2025; Hermann et al. 2015)—yielding six evaluators decomposed into narratologically grounded sub-dimensions.

Defence evaluators implement four recognised exceptions: parody/satire (*Deckmyn*(Court of Justice of the European Union 2014; Rendas 2024)), pastiche (Art. 5(3)(k) InfoSoc; AG Emiliou in *Pelham II*(European Parliament and Council 2001; Court of Justice of the European Union 2025)), quotation/citation (Art. 5(3)(d); *Funke Medien*(European Parliament and Council 2001; Court of Justice of the European Union 2019b)), and scènes à faire (idea-expression dichotomy; stock element exclusion(United States Court of Appeals for the Second Circuit 1930; United States District Court, S.D. New York 2008; United States Court of Appeals for the Second Circuit 1980; United States Supreme Court 1936; United States District Court, S.D. New York. 1986; Geiger et al. 2014)).

These sources are translated into evaluators, sub-dimensions, and weights by combining them with narratological and stylistic categories. Genette's taxonomy of focalisation and narrative person, for instance, informs the focalisation and point-of-view sub-dimensions of narrative voice (Genette 1980), while stylometric work on function words, sentence-length distribution, and syntactic patterns underpins the lexical and sentence-structure sub-dimensions of writing style (Stamatatos 2009; Juola 2008; Hoover 1973). Narrative semiotics and film theory support the segmentation into scenes and beats (Bordwell and Thompson 2012; Dancyger 2010; McKee 1997). The general procedure for selecting sub-dimensions and assigning weights is set out in Section 3.2.1.

3.2.1 Dimension Selection and Weight Determination

The selection of evaluators and their sub-dimensions follows a two-stage process. First, for each legally relevant construct (such as substantial similarity of literary expression, parody, or quotation), we identified the expressive phenomena that EU copyright doctrine plausibly treats as protectable or as components of an exception. This step draws on case law and doctrinal commentary to determine which aspects of a text might be relevant for infringement or for the application of a limitation or exception. For example, CJEU decisions and subsequent scholarship emphasise narrative voice, characterisation, plot architecture, and stylistic elaboration as indicators of originality (Court of Justice of the European Union 2009, 2019a; Lucchi and Hunter 2025; Borhi et al. 2025), and identify evocation, noticeable difference, and humorous or critical character as the core elements of parody under EU law (Court of Justice of the European Union 2014; Geiger et al. 2014).

Secondly, for each such legal construct, we map these phenomena onto operationally definable sub-dimensions that can be assessed from the text alone. This mapping uses narratology, stylistics, and authorship-attribution research to obtain theoretically grounded categories such as point of view, focalisation, sentence-length distribution, or conflict escalation patterns (Genette 1980; Juola 2008; Stamatatos

2009; Abbott 2008; Bordwell and Thompson 2012). A sub-dimension is included only if it meets three criteria: it must correspond to a feature that doctrine or commentary treats as relevant to protection or exception; it must admit a clear textual definition that can be expressed in natural-language instructions; and it must be feasible for human or LLM judges to evaluate on the basis of limited excerpts without external information.

The weights attached to sub-dimensions are chosen heuristically. They encode a normative judgement, grounded in doctrine and theory, about the relative contribution of each sub-dimension to the overall legal construct being approximated and about the expected robustness of measurement. Dimensions that are repeatedly emphasised in case law or doctrinal analysis, and that are relatively well-defined in narratology or stylistics, receive higher weights. For example, point of view and narrative distance are weighted more heavily than reader engagement in the narrative voice evaluator (Section 3.3.2), because they are both central to narratological accounts of voice and plausibly more indicative of originality (Genette 1980; Booth 1961). Conversely, dimensions that primarily capture functional roles or genre-typical elements, such as generic scene functions or character roles, receive low weights, because they are closer to unprotectable scènes à faire elements (United States Court of Appeals for the Second Circuit 1930).

3.3 Infringement Evaluators

The infringement evaluators measure similarity along dimensions corresponding to protectable expressive features; DAG node colours follow the convention established in Section 3 (blue: task nodes; red: judgement node; green: verdict nodes).

3.3.1 Writing Style Similarity

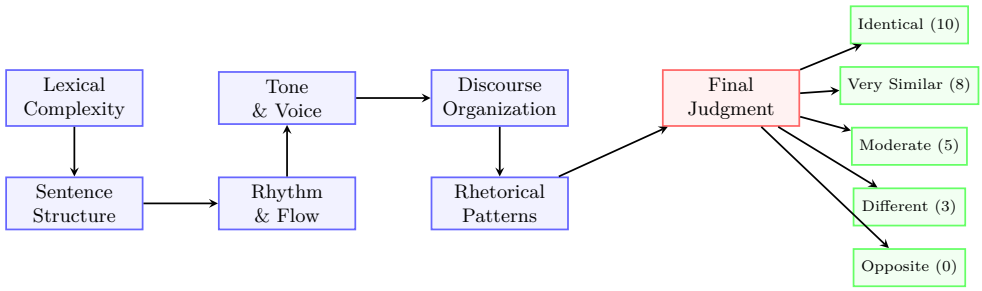


Fig. 1 DAG architecture for the writing style evaluator that shows in which sub-dimension analyses feed into a final judgement of overall writing style similarity

Figure 1 shows the writing style evaluator DAG, comprising six task nodes (Lexical Complexity, Sentence Structure, Rhythm & Flow, Rhetorical Patterns, Discourse Organisation, Tone & Voice) feeding a single judgement node.

Doctrinally, writing style is relevant because originality in literary work may manifest in the author's distinctive way of expressing ideas, not in the ideas themselves (Court of Justice of the European Union 2019a; Herrmann et al. 2015). Stylometry and authorship attribution provide empirical support for treating lexical choice and syntactic patterns as authorial signatures (Stamatatos 2009; Juola 2008; Hoover 1973).

The lexical complexity task instructs the judge model to examine vocabulary richness (patterns of repetition and variety), word length tendencies, and formality markers in each text, ignoring semantic content.

The sentence structure task similarly focuses on sentence length distribution and syntactic complexity. Informed by work showing that authors exhibit stable preferences in sentence structure (Hoover 1973), the instruction asks the model to compare whether both texts favour, for example, short simple sentences or long multi-clause constructions.

The rhythm & flow task examines punctuation patterns, sentence rhythm, and pacing, drawing on analyses of prose rhythm (Hoover 1973) and discourse markers (Biber and Finegan 1988). The model is asked to report whether both texts exhibit similar punctuation density and rhythmic patterns and then to map this to one of the five labels.

The rhetorical patterns task considers distributions of questions, imperatives, repetition, and parallelism, features that stylistic research has linked to authorial habits (Biber and Finegan 1988; Herrmann et al. 2015).

The discourse organisation task targets paragraph structure and use of connectives, grounded in work on discourse cohesion and information sequencing (Biber and Finegan 1988; Gehrmann et al. 2021). The model compares, for instance, whether both texts tend to build long multi-sentence paragraphs with explicit discourse markers or short stand-alone paragraphs.

The tone & voice task captures authorial stance, drawing on rhetorical narratology (Booth 1961; Herrmann et al. 2015). The model assesses the degree of personal versus impersonal voice and assertiveness versus heding, then classifies similarity. Tone is included even though it may be more variable than syntax, because legal and literary discussions often refer to an author's "voice" as part of style.

The judgement node for writing style lists these six analyses and assigns weights: lexical complexity and sentence structure each receive $20 - 25\%$ of the total weight, rhythm, rhetorical patterns, and discourse organisation each $15 - 20\%$, and tone around 5%. These weights are chosen heuristically, but are grounded in two considerations. First, stylometric evidence shows that lexical and syntactic features are highly discriminative for authorship (Stamatatos 2009; Juola 2008), whereas tone is more topic-dependent and less stable. Secondly, from a legal perspective, concrete formal features (wording and construction) are closer to the protected "form of expression" than high-level mood, which may overlap with genre conventions.

3.3.2 Narrative Voice Similarity

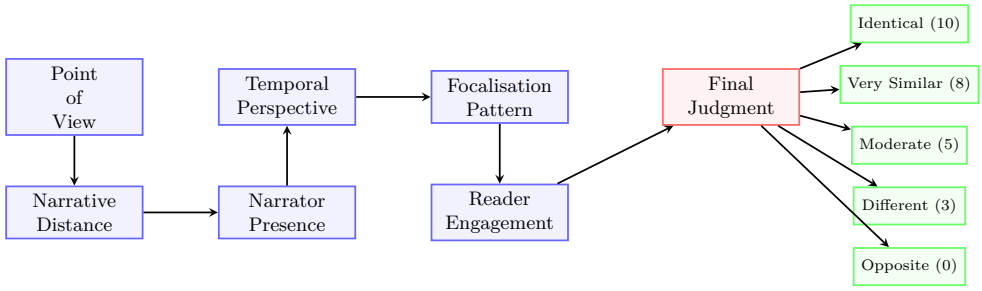


Fig. 2 DAG architecture for the Narrative Voice evaluator that shows how voice-related sub dimensions are aggregated into an overall narrative voice similarity judgement

The narrative voice evaluator (Figure 2) targets perspective-related originality: narrative voice shapes how events and characters are presented and is central to both narratology (Genette 1980; Booth 1961; Palmer 2004) and CJEU originality analysis (Herman et al. 2012).

The point-of-view task, grounded in Genette's classification of narrative person and knowledge scope (Genette 1980), instructs the model to identify, for each text, whether narration is in the first, second, or third person; whether the narrator's knowledge is limited, multiple, omniscient, or objective; and whether this configuration is consistent. This produces a local point-of-view score that reflects whether, for example, both texts use a single limited first-person narrator, or whether they differ fundamentally (such as first-person versus omniscient third-person).

The narrative distance task draws on Booth's and Phelan's work on distance (Booth 1961; Phelan 2005). The instruction asks the model to assess how close the narrator is to character consciousness, how emotionally involved the narrator is, and whether there is a shift in distance.

The narrator presence task distinguishes homodiegetic and heterodiegetic narrators and degrees of intrusiveness (Stanzel 1984; Fludernik 1996). The model compares whether the narrator participates in the events, how often they comment directly, and whether they seem inside or outside the story world.

The temporal perspective task asks the model to identify primary verb tenses and the narrator's temporal position (retrospective, simultaneous, anticipatory), following narratological treatment of time (Genette 1980).

The focalisation pattern task uses Genette's notion of focalisation (Genette 1980). The model is instructed to determine whose perspective filters the information (fixed internal, variable internal, external mixed), what information boundaries exist, and whether the two texts display similar focalisation schemes.

The reader engagement task, informed by discourse studies on audience address (Biber and Finegan 1988), asks whether and how the narrator directly

addresses the reader, how often, and with what assumed relationship (intimate, neutral, formal).

In the judgement node, the point of view and narrative distance are assigned the highest weights (together around 45%, narrator presence and focalisation intermediate weights (around 15% each), temporal perspective a modest weight, and reader engagement the smallest (around 5%). This ordering reflects the centrality of point of view and distance in narratological accounts of voice (Genette 1980; Booth 1961), and the fact that reader address is often more variable and can be governed by genre or discourse tradition (for example, children's literature may more frequently use direct address) rather than authorial individuality.

3.3.3 Character Similarity

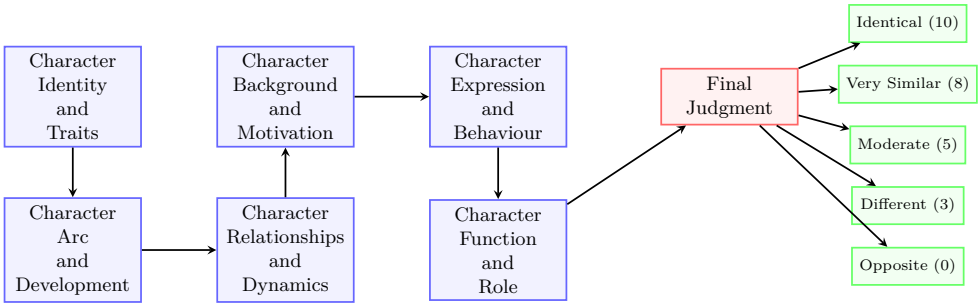


Fig. 3 DAG architecture for the Character Similarity evaluator showing the aggregating character-related sub-dimension analyses into an overall character similarity judgement

Figure 3 illustrates the character similarity evaluator. It reflects legal and narrative recognition that distinctive characters can be protectable elements of work, providing they are sufficiently defined and creative (United States Court of Appeals for the Second Circuit 1930; United States District Court, S.D. New York 2008; Kurtz and Borod 2015; Palmer 2004). Each sub-dimension is designed to distinguish generic archetypes from specific creative elaborations.

The character identity & traits task asks for a comparison of distinctive personality traits, physical idiosyncrasies, mannerisms, and psychological complexity in the two texts. It draws on literary theory emphasising that characters becomes original through detailed, specific embodiment rather than generic labels (Palmer 2004; Herrmann et al. 2015).

The character arc & development task compares the sequence and nature of character change and internal conflict, building on narrative accounts of arcs (McKee 1997; Abbott 2008). The model identifies initial states, triggers, stages, and resolutions.

The character relationship & dynamics task considers interaction patterns, power balances, and emotional textures, in line with work on social cognition in narrative (Palmer 2004; Wolf 2012).

The character background & motivation task distinguishes generic backstory (e.g. "traumatic past") from specific causal histories with detailed events and motivational structures. Motivational architecture is singled out in both legal and literary treatments as a place where originality can reside (Palmer 2004; Lucchi and Hunter 2025).

The character expression & behaviour task looks at behavioural signatures and emotional response patterns. It captures whether both characters exhibit similar specific habits and copying mechanisms.

The character function & role task compares high-level narrative functions such as protagonist, antagonist or mentor. The instruction explicitly frames these as often generic and weakly protectable, consistent with scènes à faire reasoning (United States Court of Appeals for the Second Circuit 1930).

The judgement node weights identity/traits, arc/development, and relationships highest, background/motivation and expression/behaviour moderately, and function/-role minimally. This reflects both doctrinal and doctrinally-informed commentary: courts and commentators stress that appropriation of a character's distinctive, detailed features is more problematic than sharing generic roles or archetypes (United States District Court, S.D. New York 2008; Kurtz and Borod 2015).

3.3.4 Plot Structure Similarity

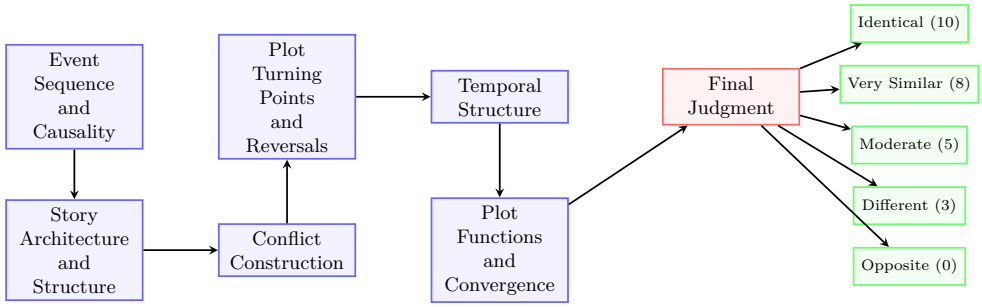


Fig. 4 DAG architecture for the Plot Structure Similarity evaluator combining event conflict and structural sub-dimensions into a final plot structure similarity judgement

Figure 4 illustrates the plot structure similarity evaluator. Plot structure is a central narrative construct and a frequent focal point in infringement disputes (United States Court of Appeals for the Second Circuit 1930, 1980; United States Supreme Court 1936; Kurtz and Borod 2015).

The event sequence & causality task instructs the model to identify and compare specific plot events and their causal links, distinguishing generic beats from detailed sequences and mechanisms.

The story architecture & structure task addresses the higher-level organisation: act division, framing devices, nested narratives, and interwoven subplots (Bordwell

and Thompson 2012; Dancyger 2010; McKee 1997). The model is asked to examine whether the two texts adopt similar structural architectures.

The conflict construction task deals with escalation patterns and obstacle networks, grounded in dramaturgical notions of rising action and complication (McKee 1997).

The plot turning points & reversals task looks at pivotal shifts and revelations. It reflects doctrinal concerns where plaintiffs allege copying of unique twists rather than generic "surprises" (United States Court of Appeals for the Second Circuit 1980; Kurtz and Borod 2015).

The temporal structure task examines chronology, flashbacks, and pacing, building on narratological work on time (Genette 1980).

The plot functions & convergence task compares how threads converge and resolve. It captures whether particular combinations of subplots and resolutions are mirrored. As with character functions, generic high-level functions are less protectable, so this dimension receives the smallest weight in the judgement node.

The judgement node weights event sequence and story architecture most heavily, conflict and turning points moderately, temporal structure lower, and plot functions-convergence lowest. Nichols and subsequence cases illustrate courts' focus on the sequence and organisation of events and on key plot twists (United States Court of Appeals for the Second Circuit 1930; Kurtz and Borod 2015). Event sequence and architecture are weighted highest, reflecting courts' focus on specific causal chains and structural copying over generic narrative functions (United States Court of Appeals for the Second Circuit 1930; Kurtz and Borod 2015).

3.3.5 Scene Sequence Similarity

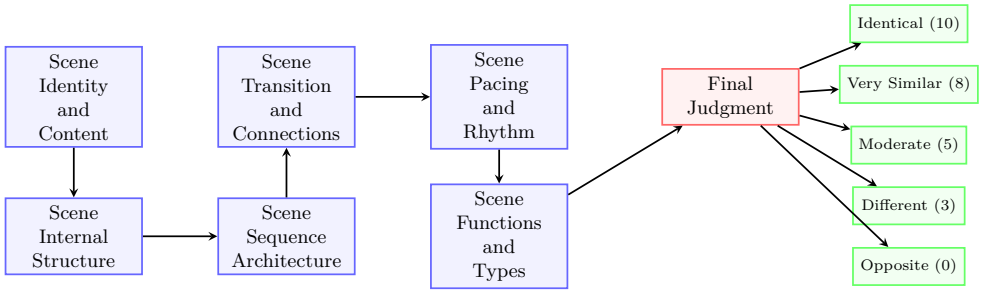


Fig. 5 DAG architecture for the Scene Sequence Similarity evaluator combining scene level sub-dimensions into an overall scene sequence similarity judgement

Figure 5 shows the scene sequence similarity evaluator. It draws on film-based narrative theory (Bordwell and Thompson 2012; Dancyger 2010; McKee 1997), probing distinctive scene construction and ordering beyond generic scene types.

The scene identity & content task asks the model to compare particular scenes in terms of setting, participants, actions, and narrative purpose. It is designed to capture

whether recognisable scenes from the source are recreated in the target at the level of concrete detail.

The scene internal structure task considers beat sequences and staging inside scenes (McKee 1997). The model examines the micro-structure of scenes and classifies how similarly they unfold.

The scene sequence architecture task looks at the ordering and relationship network of scenes, focusing on how one scene leads to another.

The scene transition & connections task focuses on linking devices and continuity. It captures whether the same motif-based or structural transitions occur.

The scene pacing & rhythm task compares patterns of scene length and tempo.

The scene functions & types task examines standard scene categories and their positions. Since scene functions are often determined by genre, this dimension is treated as weakly protectable.

The judgement node weights scene identity and internal structure highest, scene sequence architecture somewhat lower, transitions and pacing lower still, and functions/types lowest. This tries to reflect the view that copying the detailed content and staging of scenes is more indicative of appropriation than sharing generic patterns such as "climactic confrontation near the end".

3.3.6 World Building Similarity

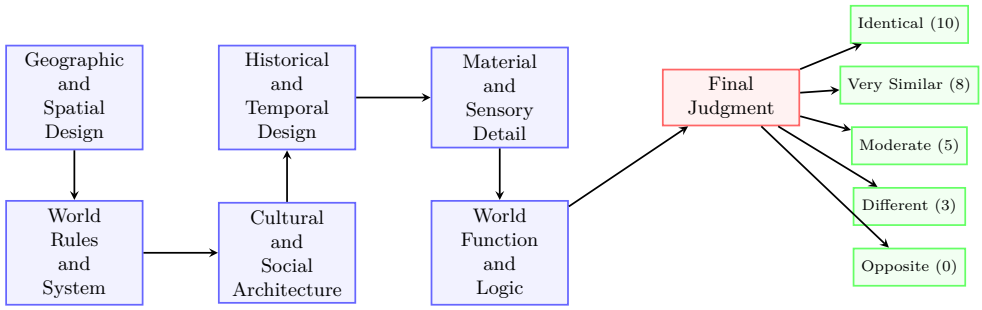


Fig. 6 DAG architecture for the World Building Similarity evaluator combining world related sub-dimensions into an overall world building similarity judgement

The world building similarity evaluator (Figure 6) targets the construction of fictional worlds, discriminating generic genre backdrops from richly elaborated original expression (Wolf 2012; Ekman 2013; Gavins 2007).

The geographic & spatial design task asks the model to compare detailed locations, spatial layouts, and environmental features. It is designed to detect, for example, copying of a distinctive city layout or landscape structure.

The world rules & system task compares magic, technology, or physics systems. The model analyses specific rules, costs, capabilities, and limitations. Fantasy and science-fiction scholarship emphasises that such systems can be highly original and recognisable (Wolf 2012).

The cultural & social architecture task considers customs, rituals, institutions, and hierarchies.

The historical & temporal design task looks at past events and temporal patterns, including cycles and eras (Ekman 2013).

The material & sensory detail task compares objects, flora and fauna, and sensory atmospheres, reflecting text-world theory's emphasis on how readers construct experiential worlds (Gavins 2007). It captures whether distinctive material signatures are shared.

The world function & logic task examines how the world operates and maintains internal coherence, but is treated as more abstract and closer to functional necessity.

The judgement node weights geographic/spatial design and world rules/systems highest, cultural/social and historical/temporal design intermediate, material/sensory detail lower, and world function/logic lowest. This pattern is justified by the observation that copying specific maps and systems is more indicative of appropriation than sharing generic principles such as "the world is internally consistent". From a more legal-oriented perspective, concrete spatial and systematic elaborations are closer to protectable expression, while abstract functional properties are closer to unprotectable ideas.

3.4 Defence Evaluator

Defence evaluators follow the same DAG architecture as infringement evaluators, with the inversion that higher scores indicate stronger exception applicability or greater genericness. Weights are determined heuristically using the same doctrine-informed procedure.

3.4.1 Parody and Satire

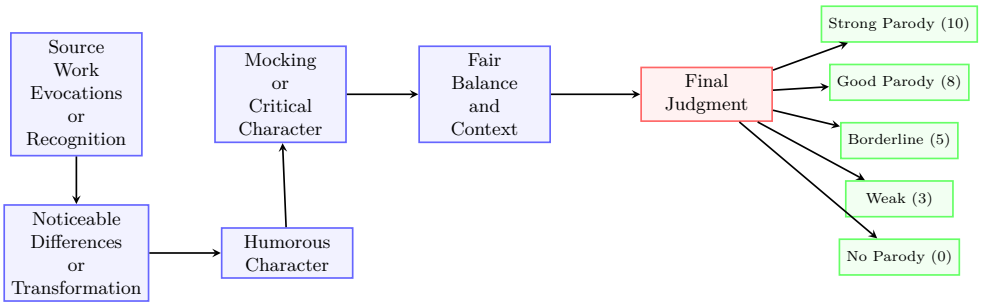


Fig. 7 DAG architecture for the Parody and Satire evaluator aggregating the Deckmyn dimensions into a final parody strength judgement

The parody/satire evaluator in Figure 7 implements the *Deckmyn* criteria (Court of Justice of the European Union 2014), as interpreted by doctrinal analyses (Rendas 2024).

The source work evocation / recognition task asks the judge model to determine how clearly the target evokes the specific source work, by looking for named references, distinctive characters, iconic settings, or recognisable stylistic imitation. It then labels evocation strength on the five-level scale, from "no evocation" to "clear, specific evocation".

The noticeable differences/transformation task asks the model to compare content and tone between source and target, and to assess whether there are substantive transformations in characters, plot, setting, and message, rather than cosmetic changes. The label expresses whether the target is "not noticeably different" or "highly transformed".

The Humorous Character task assesses presence and clarity of humour. The model identifies devices such as exaggeration, irony, absurdity, and wordplay, and classifies whether humour is absent, weak, moderate, or strong. Though Deckmyn does not prescribe a particular comedic form, the task classifies humour strength from absent to strong.

The mocking/critical character task focuses on critical commentary and target. The model determines whether the target critiques the original work, its author, a genre, or broader social phenomena, and labels the strength of this mocking or critical dimension. This is important because parody can be primarily critical, even if not overtly humorous.

The fair balance & context task evaluates proportionality and non-discrimination. It asks whether the extent of copying is proportionate to the parodic purpose and whether the work contains discriminatory content against protected groups. The label indicates whether fair balance is strongly met, borderline, or clearly violated. This tries to reflect Deckmyn's emphasis on balancing rights and the specific warning that discriminatory parody may not receive protection.

The judgement node weights evocation and transformation most heavily, humour and mockery next, and fair balance lowest but still non-negligible. This weighting is justified by the structure of the Deckmyn test: evocation and difference are necessary conditions for parody, whereas fair balance qualifies the scope of the exception. Humour and mockery are essential characteristics, but once present at a clear level, marginal increases have less legal effect.

3.4.2 Pastiche

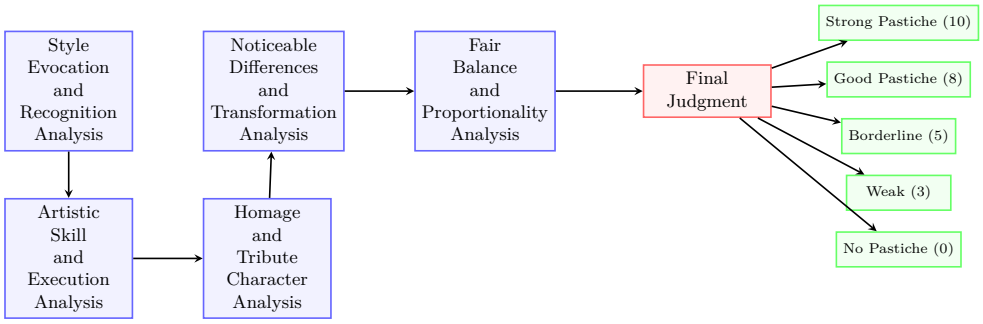


Fig. 8 DAG architecture for the Pastiche evaluator aggregating style evocation artistic skill tribute character transformation and proportionality into a final pastiche strength judgement

The pastiche evaluator in Figure 8 is grounded in Article 5(3)(k) of InfoSoc and AG Emiliou's Opinion in *Pelham II* (European Parliament and Council 2001; Court of Justice of the European Union 2025), which differentiates pastiche from parody and caricature.

The style evocation & recognition task asks the model to assess how convincingly the target imitates the source's distinctive aesthetic language (its sentence rhythm, lexical choices, tonal qualities, and narrative techniques). A high label indicates strong, source-specific stylistic echo, not just generic genre similarity.

The artistic skill & execution task evaluates technical quality and consistency of imitation. The model assesses whether the stylistic imitation is precise, coherent, and adapted naturally to the new content, and labels this from "no or clumsy imitation" to "high artistic mastery". This operationalises AG Emiliou's "artistic creation" requirement (Court of Justice of the European Union 2025).

The homage/tribute character task determines whether the work expresses respect or admiration, without mockery. The model looks for evidence of tribute (tone, framing, dedications) and absence of critical intent, then labels the strength of this homage character. This is the primary criterion that distinguishes pastiche from parody.

The noticeable differences/transformation task asks whether, despite stylistic similarity, the content (characters, plot, themes) substantially differs from the source.

The fair balance & proportionality task considers the extent of stylistic borrowing relative to new content and assesses potential market substitution, in line with the three-step test and doctrinal concerns about "catch-all" misuse of pastiche (Rendas 2024; Geiger et al. 2014).

The judgement node weights style evocation and artistic skill highest, homage character next, and transformation and proportionality lower. This tries to reflect doctrinal emphasis on the combination of recognisable style imitation and artistic

creation, together with the requirement that the work be a tribute rather than a critique. Transformation and proportionality matter, but primarily as safeguards against excessively wide use of the exception.

3.4.3 Quotation and Citation

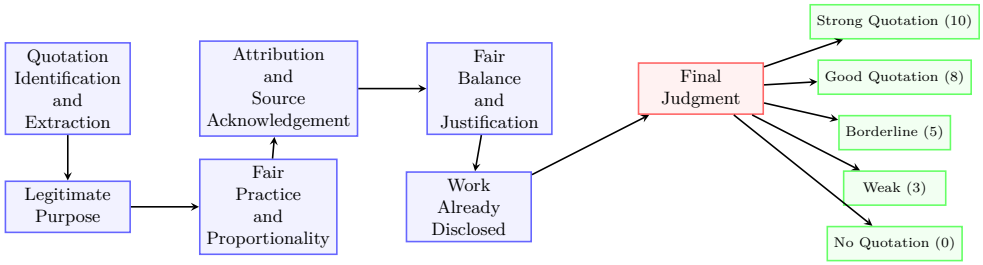


Fig. 9 DAG architecture for the Quotation and Citation evaluator aggregating quotation presence purpose fair practice attribution fair balance and prior disclosure into a final quotation strength judgement

Figure 9 shows the quotation/citation evaluator, grounded in Article 5(3)(d) of InfoSoc and CJEU case law (European Parliament and Council 2001; Court of Justice of the European Union 2019b).

The quotation identification & extraction task instructs the model to identify whether the target reproduces passages from the source verbatim or near-verbatim, whether quotation marks or other conventions are used, and to classify the strength of quotation presence: from "none" through "partial" to "substantial, explicit quotation". This directly reflects the requirement that there be quotation in the first place.

The legitimate purpose task asks the model to infer the purpose of the quotations (criticism, review, news reporting, teaching, research, or similar purposes) and to assess how genuinely the quoted material is used for that purpose, based on surrounding commentary.

The fair practice & proportionality task evaluates extent of taking relative to purpose and integration of quotations into the target's own analysis, as demanded by the "fair practice" language and the three-step test (Geiger et al. 2014).

The attribution & source acknowledgement task determines whether the source work and author are identified, as required "unless impossible" (Court of Justice of the European Union 2019b).

The fair balance & justification task examines necessity and (potential) market impact, asking whether the quotation could plausibly substitute for the source or harm its market.

The work already disclosed task focusses on whether the source appears to have been lawfully made available to the public before the target, reflecting the requirement that quotation applies only to disclosed works. This dimension is necessarily limited,

depending on the model's training data or contextual information provided within the texts.

The judgement node weights quotation identification and legitimate purpose highest, fair practice and attribution next, and fair balance and prior disclosure lower. This weighting tries to correspond to the doctrinal structure: without quotation and legitimate purpose, the exception cannot apply; fair practice and attribution are core conditions; fair balance and prior disclosure operate as background constraints.

3.4.4 Scènes à Faire

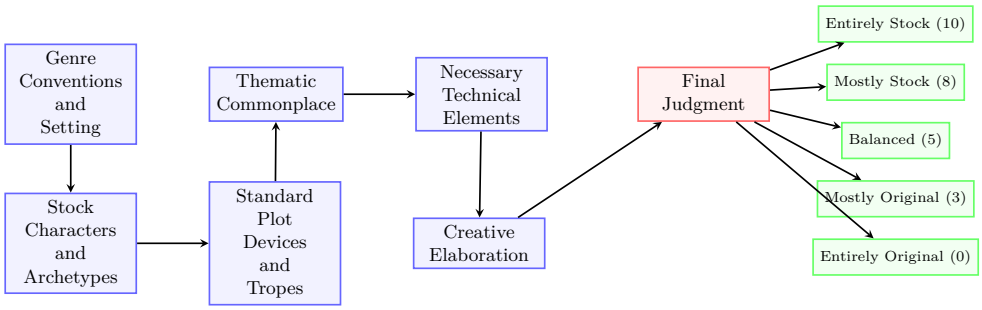


Fig. 10 DAG architecture for the Scènes à Faire evaluator combining genre character plot thematic technical and elaboration dimensions into an overall judgement of stockness or genericness

The scènes à faire evaluator (Figure 10) operationalises the exclusion of standard elements dictated by genre, setting, or technical constraints from protectable expression (United States Court of Appeals for the Second Circuit 1930; United States District Court, S.D. New York 2008; United States Court of Appeals for the Second Circuit 1980; United States District Court, S.D. New York. 1986; Geiger et al. 2014).

The genre conventions & setting task asks the model to identify genre-specific plot structures and standard settings, such as "crime leading to investigation and revelation" or "medieval fantasy kingdom with castles and dragons", and to assess how heavily both texts rely on these standard conventions.

The stock characters & archetypes task examines whether key characters fit archetypal moulds without elaboration. It captures reliance on stock roles like "reluctant hero" or "wise old mentor".

The standard plot devices & tropes task considers common devices, such as love triangles, mistaken identity, or deus ex machina.

The thematic commonplaces task evaluates whether themes such as "good versus evil" or "coming of age" are treated in a generic, conventional way. Themes as such are unprotectable ideas, but highly original thematic treatment may be reflected in lower stock-ness (Stamatatos 2009; Geiger et al. 2014).

The necessary technical elements task looks at elements required by genre or medium (for example, basic investigative steps in a detective story, scene transitions, or exposition).

The creative elaboration task, inversely defined, asks the model to assess the degree of distinctive original expression across style, characterisation, world-building, and structure, and to rate how generic or non-generic this expression is. A high label indicates that both texts are most likely largely generic; a low label indicates potential substantial creative elaboration. In combining scores, this dimension effectively tries to reduce the overall scènes à faire score when originality is high.

The judgement node assigns highest weights to genre conventions and stock characters, standard plot devices somewhat less, thematic commonplaces and functional elements lower again, and creative elaboration as a corrective term with modest weight. This tries to reflect the legal reasoning in scènes à faire cases, where courts first identify elements dictated by genre or stock situations and then ask whether the work contains additional creative elaboration that is protectable (United States Court of Appeals for the Second Circuit 1930; United States District Court, S.D. New York 2008; United States Court of Appeals for the Second Circuit 1980; United States District Court, S.D. New York. 1986).

4 Experimental Setup

The experiments address three research questions through controlled validation and large-scale corpus evaluation; hyperparameter specifications, validation corpus construction, and extended statistical protocols appear in Appendix C and Appendix D.

4.1 Corpus Construction and Preparation

4.1.1 Source Materials

The experimental corpus is drawn from the Digital Library of Dutch Literature (DBNL) (Koninklijke Bibliotheek van Nederland 2025), which provides high-quality editions of pre-1900 Dutch literary works. These works are out of copyright yet exhibit rich and distinctive narrative and stylistic features, making them suitable proxies for contemporary protected literature. Eleven authors spanning prose, drama, and children's literature were selected; approximately three works per author were sampled to ensure diversity of narrative voices and stylistic registers.

4.1.2 Story-Generation Question-Answer Dataset

To fine-tune the model in a way that emphasises stylistic learning rather than verbatim copying of prompts, the DBNL texts were converted into a question-answer (QA) dataset. Each QA pair consists of a natural-language prompt requesting a short scene or passage and an answer containing an excerpt from a source work.

The prompts were automatically generated using the open-source GPTOSS-120B model. We selected GPTOSS-120B for two reasons. First, it was reliably available in our experimental environment at the time. Second, in initial pilot experiments,

GPTOSS-120B produced prompts with better semantic alignment to the corresponding source passages and fewer formatting/artefact issues than other candidate LLMs that we tested, as judged by manual review by the authors.

For each selected passage, the model was asked to write a request that described a scene, dialogue, or monologue "in the style of" the corresponding work or author, mentioning characters, settings, or events that appear in the source text. The answer for each prompt is a contiguous excerpt of approximately $200 - 800$ tokens from the original Dutch work, chosen to match the scenario described in the question. This procedure yields around 500 QA pairs. A random sample of generated prompts and associated passages was inspected and manually edited when necessary to ensure that the prompt semantically matches the answer and that the answer does not contain typesetting artefacts.

The QA format enables the same dataset to be used for both supervised fine-tuning and generation-based evaluation, by prompting the model with the questions and comparing outputs against source passages.

4.1.3 Translation and Modernisation

The DBNL texts are written in historical varieties of Dutch that differ substantially from contemporary usage in orthography, morphology and lexicon. Pilot fine-tuning showed that LLaMA 3.2 1B struggles with archaic Dutch, reverting to modern Dutch even when prompted otherwise. Since the study probes stylistic appropriation rather than historical language acquisition, source passages were translated into modern variants.

For each passage, three variants exist: the original historical Dutch excerpt, a modern Dutch translation, and an English translation. Only the translated variants are used during fine-tuning, unlearning, and generation; original passages are retained as reference material.

The translations were produced with a combination of GPTOSS-120B and OpenAI's GPT-4o-mini. GPTOSS-120B was used as the primary translator, while GPT-4o-mini was employed for a minority of passages (on the order of $50 - 100$) for efficiency reasons when the self-hosted model was unavailable. Both directions (Dutch $\rightarrow$ modern Dutch and Dutch $\rightarrow$ English) were translated independently rather than via cascading translation.

Quality control employed manual back-translation checking on randomly sampled rows, with the acceptance criterion being semantic fidelity (events, characters, and relations preserved; fluent target-language prose), permitting paraphrase and synonymy. No systematic errors were observed.

4.1.4 Forget and Retain Set Partitioning

To study unlearning, the QA dataset is partitioned into *retain* and *forget* sets. The partitioning is performed at the semantic level of works and authors rather than at the level of individual QA pairs, to mimic realistic copyright takedown requests, though sequential takedown scenarios are outside the scope of this study.

Approximately 350 QA pairs form the retain set and remain available throughout training and unlearning, while around 130 pairs form the forget set. Forget sets

are constructed using two strategies. First, is author-level forgetting, where for some authors all books and their associated QA pairs are assigned to the forget set. Second, is book-level forgetting, where for other authors only specific works are designated as forget targets, while their remaining books remain in the retain set.

4.2 Model and Training Configurations

4.2.1 Base Model and Variants

All experiments use Unsloth's optimised Meta LLaMA 3.2 1B Instruct model; the unmodified model serves as the per-language baseline. Fine-tuning and unlearning then produce additional variants as summarised in Table 1.

Model	Language	Stage	Training data	Objective
`en_b`	EN	baseline	–	–
`nl_b`	NL	baseline	–	–
`en_f`	EN	fine-tuned	EN QA (forget + retain)	CE
`nl_f`	NL	fine-tuned	NL QA (forget + retain)	CE
`en_npo`	EN	NPO unlearning	EN QA (forget, retain)	NPO loss
`nl_npo`	NL	NPO unlearning	NL QA (forget, retain)	NPO loss

Table 1 Overview of model variants and training regimes EN and NL denote English and modern Dutch respectively CE denotes cross entropy loss used for fine-tuned training

4.2.2 Fine-Tuning and Unlearning Configuration

Both fine-tuning and unlearning use LoRA adapters. Fine-tuning is performed separately per language on the full QA dataset; NPO unlearning starts from the fine-tuned model and updates adapters using the forget-retain partitions (Section 4.1.4). Table 2 summarises all hyperparameters used in the experiments.

Fine-tuning minimises token-level cross-entropy on QA answer text. NPO updates use the same optimiser and schedule with the loss from Zhang et al. (2024): negative preferences on the forget set combined with KL regularisation on the retain set. PSALM is agnostic to the internal loss form; the relevant artefacts are the model variants in Table 1 produced under the configuration in Table 2.

4.2.3 Output Generation

Standardised prompts elicit narrative text that potentially exhibits stylistic similarity without explicitly requesting copying. For base models, the prompts specify abstract scenarios/themes present in the corpus. For fine-tuned and unlearned models, the prompts include QA dataset questions used during training.

The hyperparameters used during output generation are shown in Appendix Table B1.

Parameter	Fine-tuning	NPO unlearning
Training epochs	5–20 (pilot); chosen config: 10	3–10 (pilot); chosen config: 5
Sequence length	1024	1024
Batch size (per device)	16	8
Gradient accumulation	8	1
Effective batch size	128	8
Learning rate	5×10^{-4}	5×10^{-4}
LR scheduler	cosine with warmup	cosine with warmup
Warmup ratio	0.05	0.05
Optimiser	AdamW	AdamW
Weight decay	0.01	0.01
Max gradient norm	1.0	1.0
Mixed precision	bfloat16	bfloat16
LoRA rank	32	32
LoRA alpha	64	64
LoRA dropout	0.05	0.05
Target modules	all attention and FFN projections	all attention and FFN projections
NPO temperature α		1.0
NPO KL coefficient β		0.1

Table 2 Hyperparameters for LoRA fine tuning and NPO unlearning Values are identical for English and Dutch unless noted otherwise where a parameter is specific to NPO the finetuning column is left blank

4.3 Evaluation Procedure

All PSALM evaluators use GPT-5-nano as the judge model; verbose mode is disabled during corpus evaluation.

4.3.1 Controlled Validation Set

Each evaluator is validated on a purpose-built corpus of 50 text pairs: five test cases per evaluator (TC1–TC5) spanning the full similarity range (Appendix C).

The criteria for determining whether an evaluator is deemed usable in the main experiments are as follows. First, the validated evaluators, where a strict alignment rate (proportion of predictions with ± 0.1 of the expected score) is at least 80% and mean absolute error (MAE) is at most 0.15. Secondly, the partially validated evaluators, where the alignment rate (within ± 0.2) at least 70% and MAE at most 0.25. Lastly, the failed evaluators that do not meet the (partially) validated criteria.

These thresholds are heuristic: a tolerance of ± 0.2 corresponds to one adjacent category on the five-point scale, so an 80% alignment rate requires the evaluator to assign the correct band in the large majority of cases. Validated evaluators are used in all subsequent analyses; partially validated evaluators are retained for exploratory inspection only; failed evaluators are excluded or revised.

4.3.2 Corpus Evaluation Execution

For each QA pair (stratified by language and forget-retain partition), the model generates a response x_t; the corresponding source passage x_s is retrieved; and all validated PSALM evaluators $E(x_s, x_t)$ are executed, yielding up to ten scores per pair alongside the lexical baselines (Section 4.4).

The resulting dataset records, per generated answer, its language, author, forget-retain status, model condition, PSALM scores, and lexical metrics.

4.4 Baseline Computational Metrics

Three standard lexical overlap metrics are computed for every (x_s, x_t) pair as reference points for interpreting PSALM scores (Table 3).

Metric	Implementation	Summary
Exact Match	custom string matching	indicator equal to 1 if the normalised source passage appears as a contiguous substring of the model output, and 0 otherwise
BLEU-4	`sacrebleu`	n-gram precision with brevity penalty using up to 4-grams, as commonly applied in machine translation and memorisation studies
ROUGE-L F_1	`rouge-score`	longest-common-subsequence-based ROUGE-L, converted to an F_1 score to capture overlap in token order and coverage

Table 3 Lexical baseline metrics used alongside PSALM scores

4.5 Statistical Analysis

4.5.1 Metric Category Taxonomy

Aggregated summaries are computed at three levels: the generation level, where each QA prompt contributes one set of scores; the model-condition level, where scores are aggregated across prompts for a fixed language, model stage, and split, forming the primary input to significance testing; and the category level, where metrics are grouped into four conceptual categories, namely, computational (Exact Match, BLEU, ROUGE-L), stylistic (Writing Style, Narrative Voice), content (Character, Plot, Scene, World-building), and exceptions (Parody/Satire, Pastiche, Quotation/Citation, Scènes à Faire), to inspect whether trends are consistent within broader copyright-relevant notions.

4.5.2 Assumption Checks

Because PSALM scores are bounded and often exhibit ceiling effects after fine-tuning, strict normality and equal variance are hard to justify. Before choosing tests for each metric, Shapiro-Wilk tests for normality and Levene's tests for homogeneity of variance are applied as provided in Appendix D. In almost all cases at

least one assumption is violated. Consequently, non-parametric methods (Kruskal-Wallis, Mann-Whitney, Wilcoxon signed-rank, ART-ANOVA) are used as the primary inferential tools. Parametric results, where reported, are interpreted as descriptive summaries and accompanied by appropriate caveats.

4.5.3 Primary Hypothesis Tests

Effect of fine-tuning and unlearning.

For each metric, language and split, an omnibus Kruskal-Wallis test is used to compare the three training stages (baseline, fine-tuned, unlearned). When significant, pairwise Mann-Whitney U-tests with Bonferroni correction probe specific contrasts such as baseline versus fine-tuned (RQ2) and fine-tuned versus unlearned (RQ3).

Within-model reductions.

To quantify how much unlearning changes behaviour for the same prompts, paired Wilcoxon signed-rank tests compare fine-tuned and unlearned scores per QA pair. These tests capture whether NPO produces systematic reductions in similarity on the forget set, and whether any unintended changes occur on the retain set.

Language $\times$ stage interactions

To explore whether training-stage effects differ between English and Dutch, ART-ANOVA is applied with factors language and stage. Main effects test whether one language consistently exhibits higher similarity than the other, while interaction terms test whether, for example, unlearning is more effective in Dutch than in English. When significant interactions are found, simple-effects analyses stratified by language follow the procedure outlined in Appendix D.

4.5.4 Effect Size Estimation

For Mann-Whitney and Wilcoxon tests, Cliff's δ serves as the primary effect size. The values $|\delta| < 0.2$ are interpreted as negligible, $0.2 \leq |\delta| < 0.5$ as small, $0.5 \leq |\delta| < 0.8$ as medium, and $|\delta| \geq 0.8$ as large. Bootstrap confidence intervals for δ are obtained following Appendix D.

For ART-ANOVA and Kruskal-Wallis tests, eta-squared η^2 and omega-squared ω^2 are reported as rank-based analogues of the proportion of explained variance, again following the definitions in the appendix.

Where parametric comparisons are included for completeness (e.g., to facilitate comparison with other work), Cohen's d and Hedges' g are reported as standardised mean differences. Given the discrete scale of PSALM's scores, these values are interpreted cautiously and primarily used for relative ranking rather than as precise metric distances.

4.5.5 Equivalence Testing

For RQ3, it is insufficient to show that NPO reduces stylistic similarity; we additionally test whether unlearned models have returned to baseline levels using two one-sided tests (TOST).

For each metric, language and split, TOST compares the mean score of the unlearned model to that of the baseline model with equivalence bounds of ± 0.1 on the normalised $0-1$ scale. These bounds correspond roughly to half a PSALM category and represent the smallest difference considered practically meaningful. If both one-sided tests reject the null hypothesis that the difference exceeds the bounds, the unlearned model is considered statistically equivalent to the baseline for that metric. If equivalence cannot be established, results are interpreted either as residual over- or under-similarity relative to the baseline.

4.6 Computational Resources

Experiments utilise NVIDIA RTX 4090 GPUs (24GB VRAM) as the primary computational resource, with NVIDIA L40(s) and H100 GPUs available where requirements exceeded capacity. The implementation employs Hugging Face Transformers, Unsloth (optimised LoRA), DeepSpeed/Accelerate (training optimisation), DeepEval (LLM-as-judge infrastructure), and PyTorch.

Estimated computational budget: ≈ 30 GPU-hours total (4 fine-tuning, 4 unlearning, 6 generation, 16+ evaluation). Actual requirements vary with convergence speed, hardware, hyperparameters, optimisation, and dataset sizes.

5 Results

This section presents the quantitative results of the PSALM evaluation under the experimental conditions defined in Section 4.

Throughout the results, we use the following notation for each of the models: English baseline (en_b), English fine-tuned (en_f), English NPO unlearned (en_{npo}), Dutch baseline (nl_b), Dutch fine-tuned (nl_f) and Dutch NPO unlearned (nl_{npo}).

5.1 Controlled Validation of PSALM Evaluators

The validation corpus provides five hand-crafted test cases per evaluator, each instantiating one level of the PSALM scale with five stochastic replications, assessing whether the LLM-as-judge architecture reliably reproduces the qualitative distinctions the legal and narratological design encodes. Additional similar results are reported in Appendix E.

5.1.1 Validation Performance Summary

Figure 11 and 12 summarises overall performance per evaluator in terms of strict alignment rate and MAE. All evaluators comfortably exceed the pre-defined validation thresholds: strict alignment is at least 0.92 and MAE remains well below the 0.15 ceiling for every metric, leaving no evaluator in the partial or not-validated range. On a five-point scale where adjacent categories are separated by 0.2, this means that, on average, the evaluators deviate from the intended level by substantially less than one-third of a category.

The DAG-based prompts successfully implement the full range of constructs defined in Section 3 with no evaluator persistently near threshold. Performance

is also broadly comparable across the three conceptual groups (Stylistic, Content, Exceptions, as examined further in Figure 13).

The two deviations at the 0.80 level are informative: Character Similarity and Pastiche exhibit the highest MAE overall and are the only evaluators producing errors of this magnitude, reflecting the legal and narratological subtlety of their constructs.

5.1.2 Prediction Accuracy and Alignment behaviour

The alignment analysis in Figure 14 provides a complementary view. Across all evaluators and test cases, more than four-fifths of the 250 individual evaluations fall within the EXACT band (± 0.1 of the expected score), fewer than one in twenty fall more than one category away, and POOR cases are rare. Errors manifest predominantly as local boundary shifts—rating a "very similar" pair as "moderately similar"—rather than confusions across qualitatively distinct regions of the scale (Figure 13).

Character Similarity and Pastiche account for all observed POOR alignments: the judge model either underestimates similarity for near-parallel character pairs or oscillates between classifying a stylised homage as strong pastiche versus weak stylistic imitation. Subsequent analyses therefore interpret modest differences on these two metrics more cautiously than comparable differences on, for example, Plot Structure Similarity or Quotation/Citation.

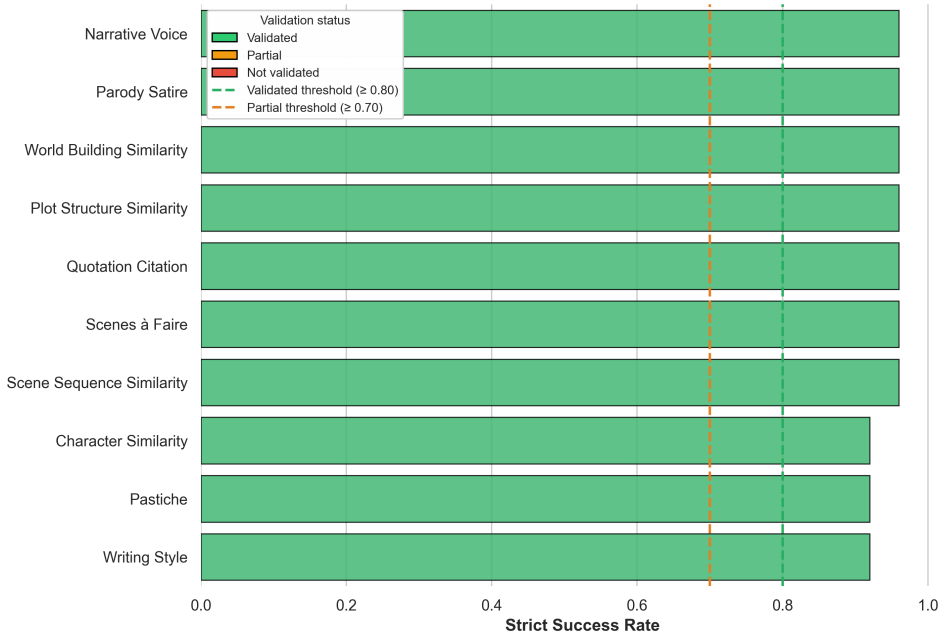


Fig. 11 Strict alignment rates per evaluator across all ten metrics; all evaluators exceed the validated threshold (≥ 0.80), and the partial and not-validated legend entries are included for reference only as no evaluator falls below this threshold

5.2 Stylistic Appropriation Through Fine-Tuning

We compare baseline models (en_b, nl_b) with their fine-tuned variants (en_f, nl_f) on both forget and retain splits to assess whether supervised fine-tuning induces stylistic appropriation beyond verbatim memorisation. Additional results are reported in Appendix F.

5.2.1 Overview of Fine-Tuning Effects

Fine-tuning on the DBNL-derived QA corpus substantially increases similarity between model outputs and the underlying works across all infringement-oriented metrics. Figure 15 shows that, averaged across all 13 metrics, both fine-tuned models consistently achieve lower (better) ranks than their baselines on both splits; the English–Dutch difference is small relative to the baseline–fine-tuned gap.

The heatmap in Figure 16 and the category-level means in Table 4 show how this improvement is distributed across metric families. For both languages and both splits, fine-tuning drives computational metrics (Exact Match, BLEU, ROUGE) from values close to zero to values very close to one. Stylistic metrics (Writing Style, Narrative Voice) and content metrics (Character, Plot, Scene Sequence, World-Building) follow the same pattern: baseline models exhibit low to moderate similarity, whereas fine-tuned models are typically judged by PSALM as "very similar" or "near-identical" to the source passages. In contrast, defence-oriented metrics (Parody/Satire, Pastiche,

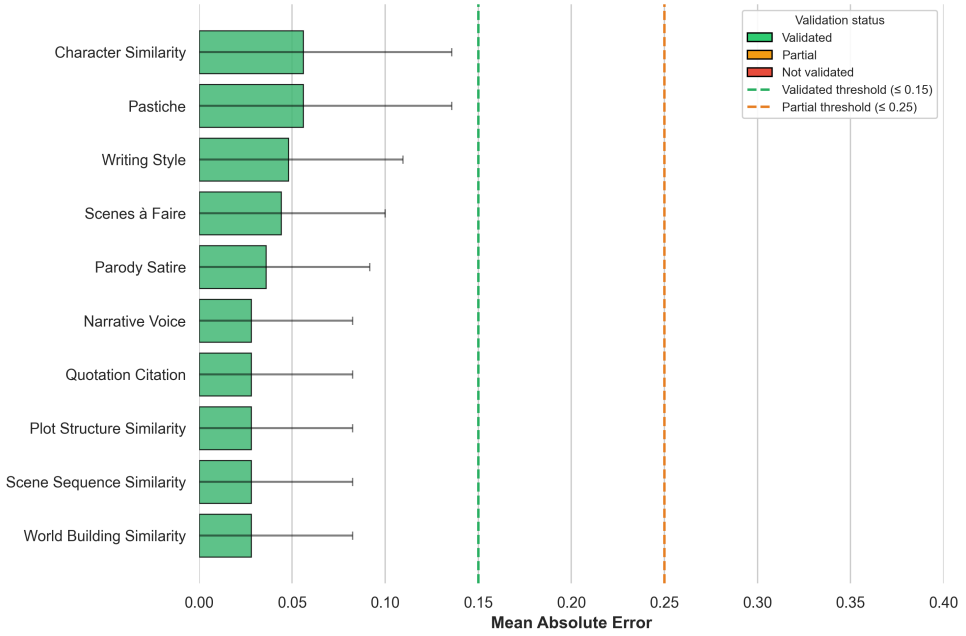


Fig. 12 Mean absolute error per evaluator across all ten metrics; all evaluators remain within the validated threshold (≤ 0.15), and the partial and not-validated legend entries are included for reference only as no evaluator exceeds this threshold

Quotation/Citation, Scènes à Faire) change much less. Quotation/Citation increases moderately, while Parody/Satire and Pastiche remain low and Scènes à Faire stays around the mid-range for all conditions. This is because the corpus does not necessarily address these exceptions, and therefore no questions are asked about aspects such as parody or quotations.

Table 4 also shows that English baselines start from slightly higher similarity than Dutch baselines, reflecting the English-centric pre-training of LLaMA 3.2 1B. After fine-tuning, this difference almost disappears: both languages converge to similarly high scores on computational, stylistic and content categories, and to similar, much lower scores on exceptions. This convergence suggests that the supervised fine-tuning stage is the dominant driver of similarity behaviour, while cross-linguistic differences in the base model matter mainly before training.

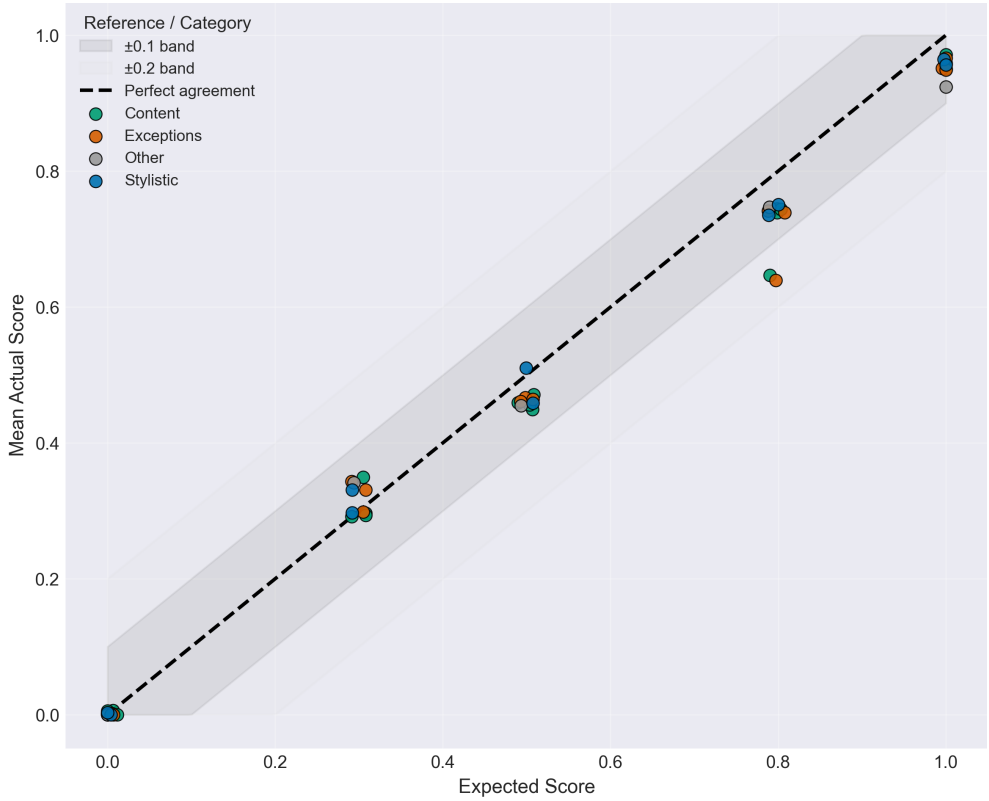


Fig. 13 Mean actual score versus expected score for each evaluator across five experimental scenarios, with each point representing one evaluator coloured by category (Content, Exceptions, Stylistic, Other) and slight horizontal jitter applied within each scenario cluster for legibility; two evaluators at the 0.80 expected score level fall outside the ±0.1 acceptable band

5.2.2 Stylistic and Structural Appropriation

Figure 17 shows the general trend of the forget-split on stylistic and content evaluators, while Appendix Figures F4 and F3 reports similar results for the other evaluators of the forget-split, as well as for the retain-split.

In the baselines, PSALM assigns low-to-moderate similarity scores on these dimensions. For English, the results are somewhat higher than for Dutch, which is consistent with the English bias in pre-training and with RQ1's observation that the English baseline already partially overlaps with stylistic features of late 19th century prose, even when translated. However, even the English baseline tends to fall into PSALM's "somewhat similar" or "moderately similar" bands, not the "very similar" or "near-identical" categories.

Fine-tuning moves all stylistic and structural metrics sharply upward. Writing Style and Narrative Voice scores cluster near one for both languages, with PSALM judging the generated answers as exhibiting near-identical lexical and syntactic patterns, rhythm, rhetorical devices, and narratorial perspective. Character, Plot, Scene Sequence, and World-Building similarities likewise rise from near-zero baselines to above 0.9 on both splits.

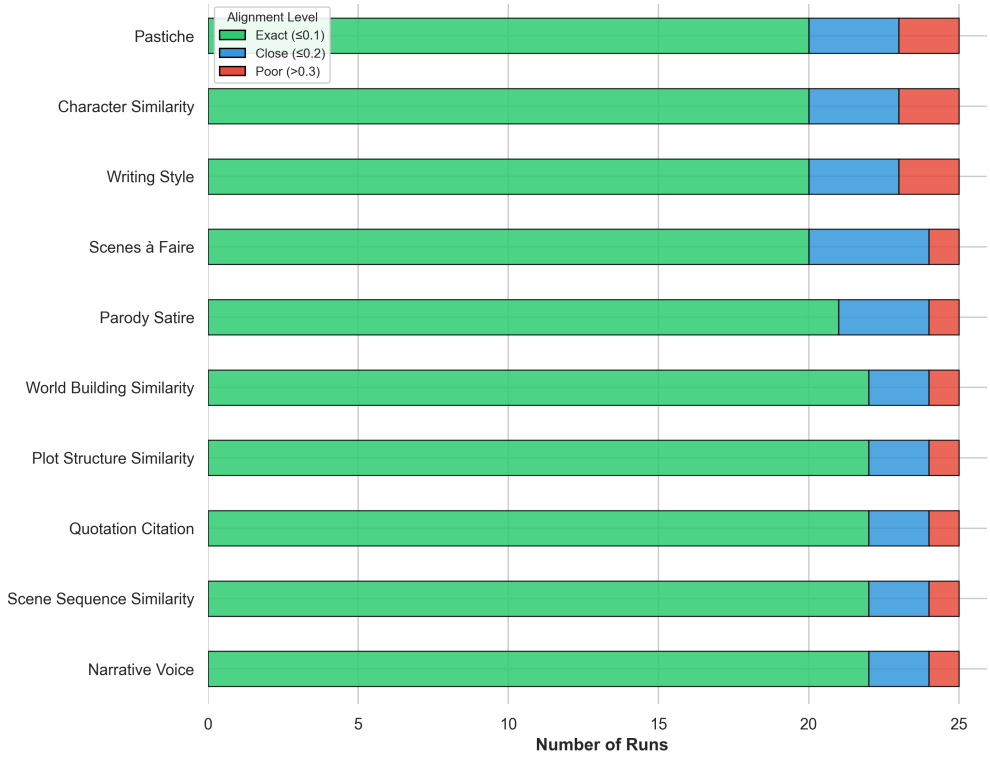


Fig. 14 alignment classification per evaluator with predominant exact or close alignment and few poor alignments

Split	Category	Baseline EN	Fine-tuned EN	Baseline NL	Fine-tuned NL
Forget	Computational	0.05	1.00	0.04	0.97
	Stylistic	0.18	1.00	0.08	0.97
	Content	0.21	0.99	0.04	0.95
	Exceptions	0.21	0.25	0.13	0.37
Retain	Computational	0.08	1.00	0.03	0.99
	Stylistic	0.30	0.99	0.23	0.97
	Content	0.27	0.97	0.25	0.98
	Exceptions	0.24	0.32	0.18	0.33

Table 4 Category-level mean PSALM scores (0 to 1 scale) before and after fine-tuning by language and split

Paired Wilcoxon tests, in Appendix F, confirm that these increases are statistically robust, with large or very large effect sizes across all stylistic and content metrics for both languages and both splits. Crucially, the effect is nearly as strong on the retain split as on the forget split.

Under EU copyright doctrine, these findings are significant because they concern precisely the expressive dimensions courts treat as indicators of originality: narrative voice, character construction, plot architecture, and world-building. PSALM's near-ceiling post-fine-tuning scores indicate that fine-tuned models routinely produce outputs perceived as closely tracking an author's "own intellectual creation" across multiple legally significant dimensions, not merely surface text.

Fig. 15 Average rank (lower is better) across all metrics for each model and split of the baseline and fine-tuned models

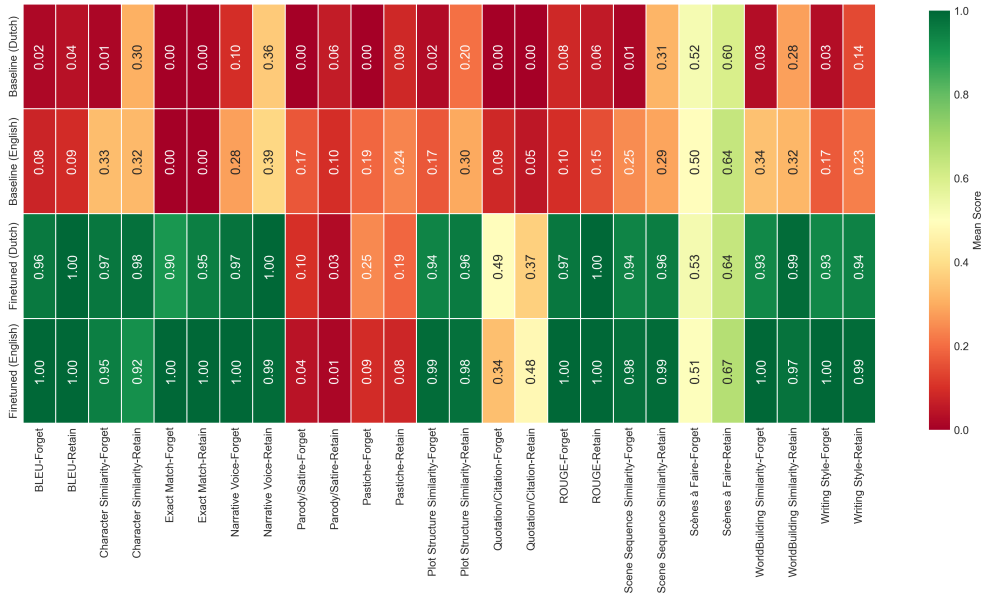


Fig. 16 Mean PSALM scores per metric and model condition (forget and retain splits) where the finetuned models show high mean scores (green) and baseline low mean scores (red) with statutory exceptions being low in both cases

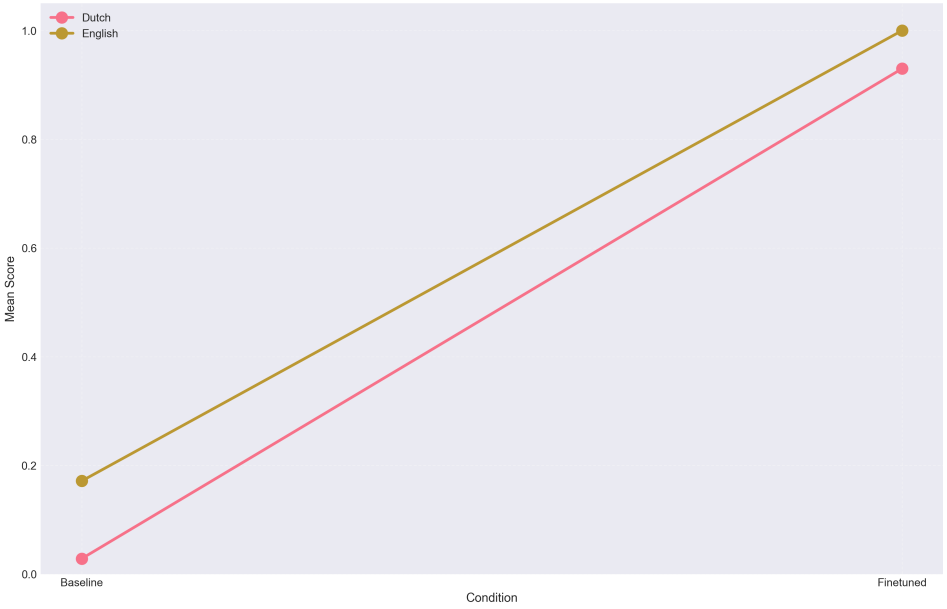
5.2.3 Behaviour of Exceptions

The defence-oriented evaluators behave very differently from the infringement-oriented ones. Figure 16 shows the overall pattern, while additional detailed split-level exception plots are provided in Appendix Figures F5 and F6.

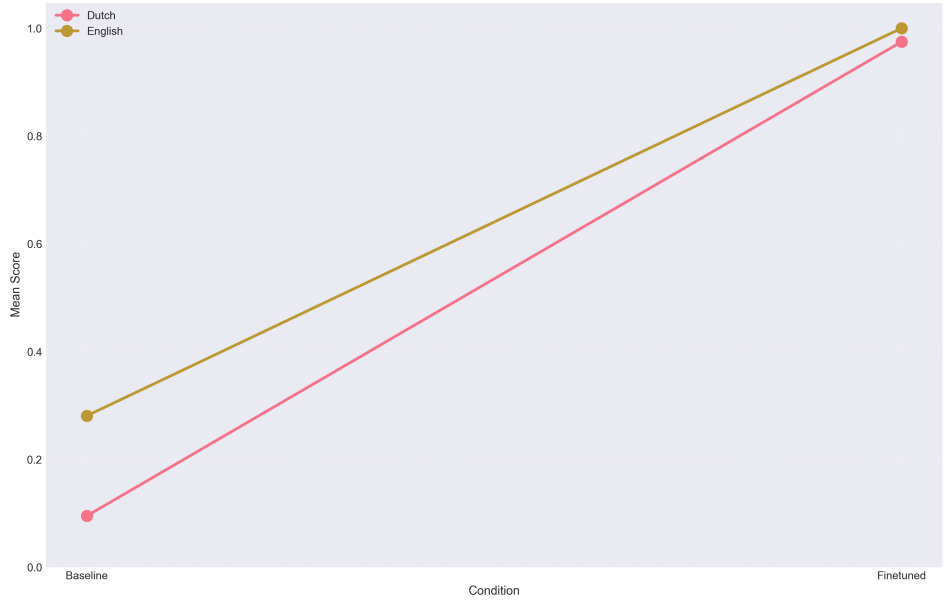
Parody/Satire and Pastiche scores remain low across all conditions. Some paired comparisons reach statistical significance, but absolute differences are negligible relative to shifts on infringement-oriented metrics, with no indication that fine-tuning produces systematically satirical or pastiche-like outputs.

Quotation/Citation is the only exception metric that increases substantially after fine-tuning. In both languages and both splits, scores move from near-zero to moderate levels (roughly one third to one half of the maximum). This reflects the fact that, after fine-tuning, the models often embed verbatim fragments of the source text in their answers in a way that PSALM judges as quotation-like: the answer contains the original wording within a surrounding narrative context. However, PSALM does not model all conditions of the Article 5(3)(d) quotation exception, such as lawful access or proportionality.

Scènes à Faire scores, which estimate how much of the similarity can be attributed to unprotectable stock elements, remain stable across all model conditions. Both baselines and fine-tuned models sit in the middle of the scale (around $0.5 - 0.67$), with no consistent up- or downward trend.



(a) writing style interaction on forget set showing low scores on the baseline and high scores on fine-tuned model



(b) narrative voice interaction on forget set showing low scores on the baseline and high scores on fine-tuned model

Fig. 17 PSALM scores for baseline against finetuned models on the Forget set showing that finetuning increases stylistic and content similarity writing style narrative voice toward 1 across languages and splits suggesting strong imprinting of author style and other evaluators show similar trends Appendix Figure F4

5.3 Unlearning Effectiveness via Negative Preference Optimisation

This section compares fine-tuned models (en_f, nl_f) with their NPO-unlearned counterparts (en_{npo}, nl_{npo}) on both forget and retain splits. Additional results are reported in Appendix G.

5.3.1 Overall Unlearning Behaviour

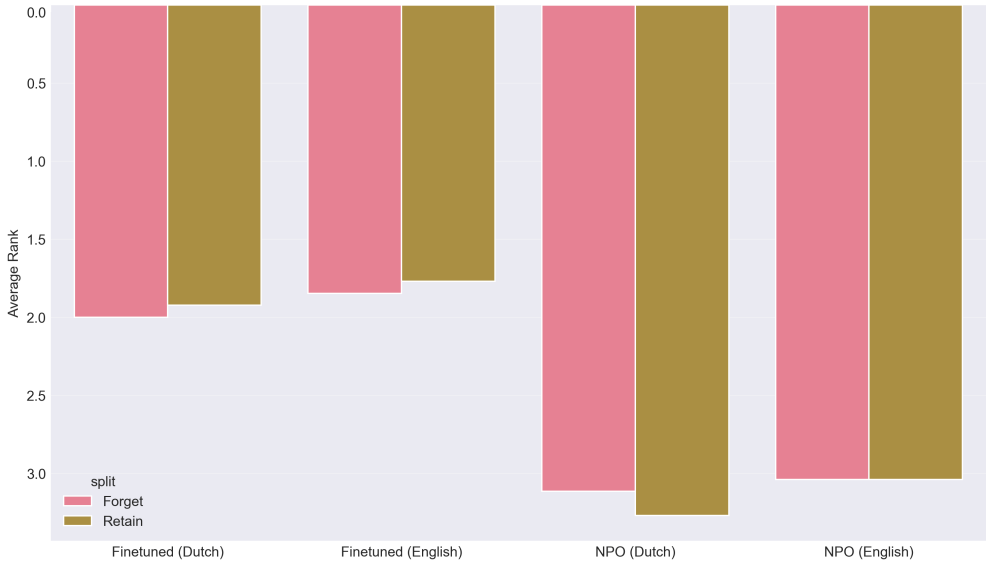


Fig. 18 Average rank (lower is better) across all metrics for each model and split of the fine-tuned and unlearned models

Across both languages, NPO consistently pushes models away from the training works. Figure 18 shows that the average ranks increases from approximately 1.8–2.0 for the fine-tuned models to roughly 3.0 for the NPO variants on both forget and retain splits. However, they do not return to baseline behaviour: even after unlearning, they remain more similar to the sources than the original instruction-tuned models, indicating that some imprint of the fine-tuning corpus persists.

Figure 19 clarifies where these changes occur. On the forget split, computational metrics (exact match, BLEU, ROUGE) drop from almost perfect similarity to values close to zero. Stylistic and content categories also decrease sharply, from near 1.0 to means around $0.2 - 0.3$. Scores in this range correspond roughly to PSALM's "somewhat different" band, so the unlearned models are no longer judged as near-identical to the forget passages, but they are not fully dissimilar either. On the retain split, the same categories drop only part of the way: mean scores remain around 0.8 for stylistic and content dimensions, i.e., still in the "very similar" region.

Exception metrics behave differently as shown in Figure 19 and additionally in Appendix Figures G10 and G11, Pastiche and parody/satire remain low throughout and increase only modestly after unlearning, while quotation/citation decreases on the forget split and scènes à faire increases slightly.

5.3.2 Suppression of Literal Memorisation

Figure 20 (and additionally Appendix Figure G12) shows how NPO affects the three computational baselines. On the forget split, exact match is effectively eliminated: both languages move from near-perfect exact match rates under fine-tuning to zero matches after unlearning. BLEU and ROUGE follow the same pattern, dropping from values close to one to means around $0.15 - 0.20$. Rank-based omnibus tests and Wilcoxon signed-rank comparisons (Appendix G) indicate that these reductions are statistically robust and associated with very large effect sizes.

NPO thus addresses the most obvious class of infringements—exact or near-exact reproduction of training passages—but computational metrics say little about whether the remaining outputs are still too close in terms of character, plot, or world-building, dimensions that EU copyright doctrine treats as protectable expression even in the absence of verbatim copying.

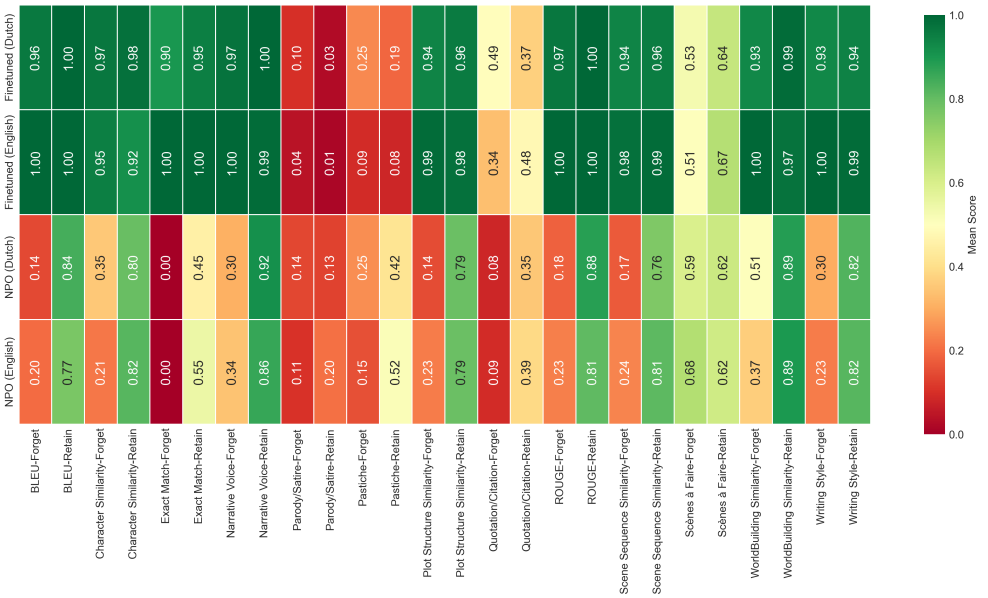
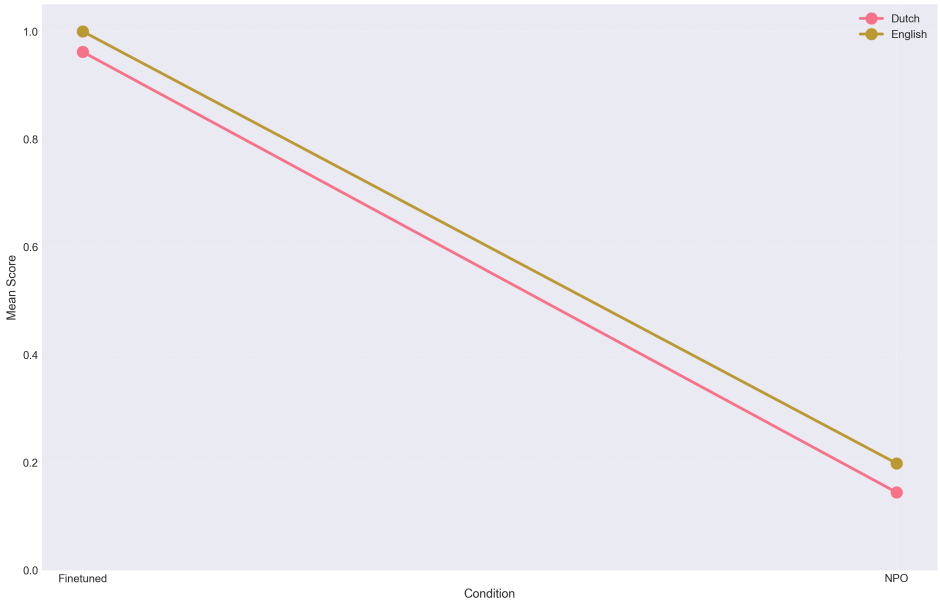
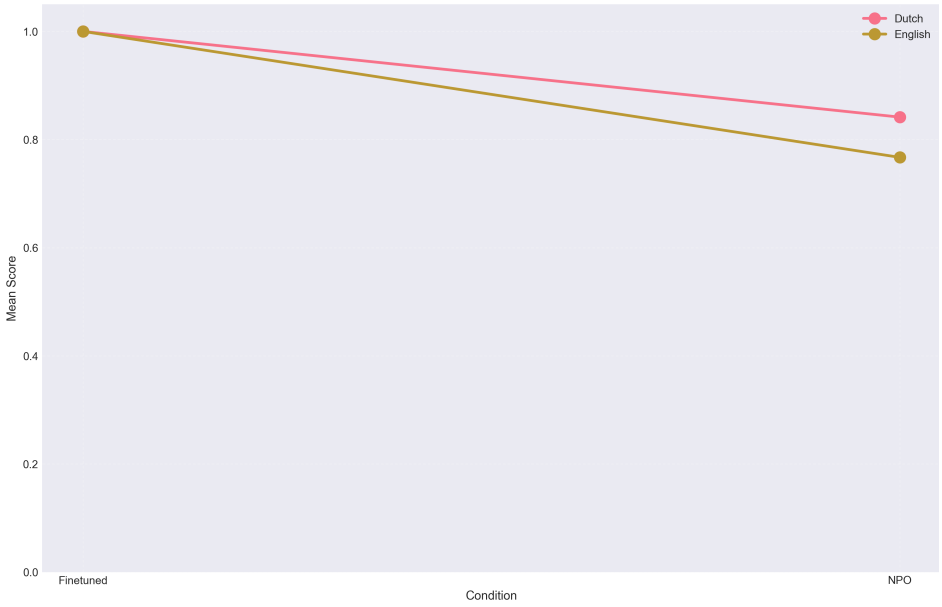


Fig. 19 Mean PSALM scores per metric and model condition (forget and retain splits) where the finetuned models show high mean scores (green) and unlearned low mean scores (red) with statutory exceptions being low in both cases

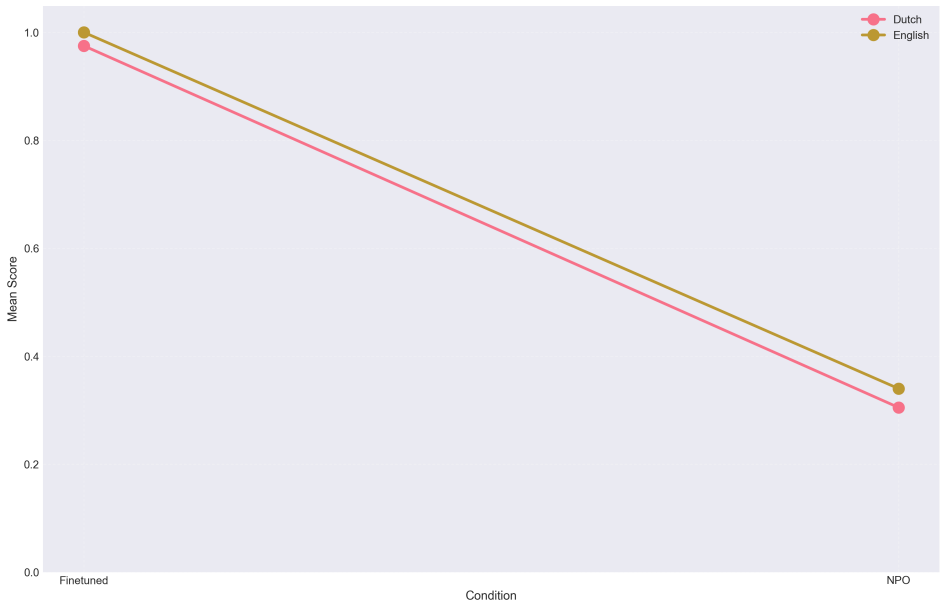


(a) BLEU interaction on forget set showing strong decrease in score from fine-tuned to NPO

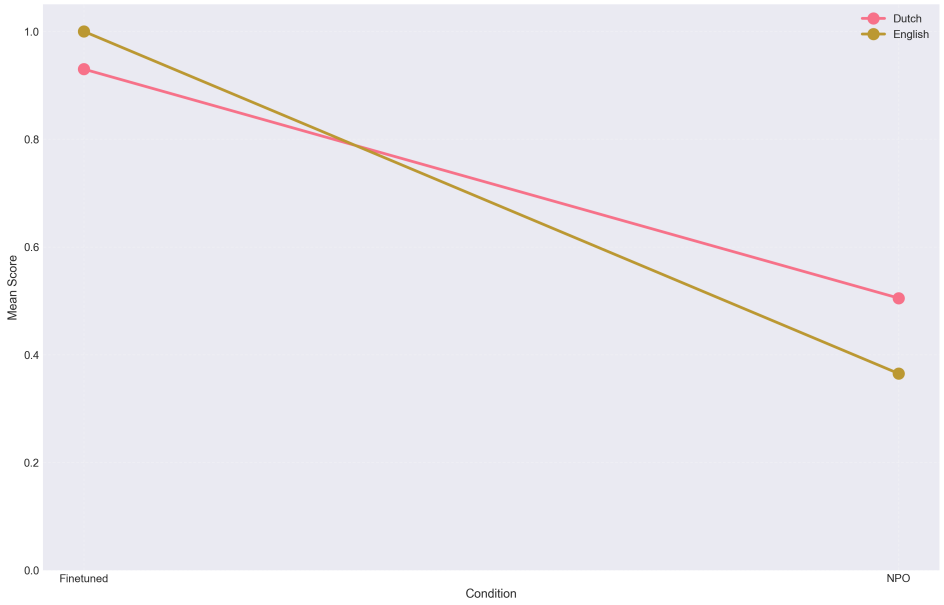


(b) BLEU interaction on retain set showing a slight decrease but still high score from fine-tuned to NPO

Fig. 20 Computational similarity metric results for finetuned and NPO models where NPO almost eliminates verbatim reproduction on forget sets while leaving substantial overlap on retain sets and other evaluators show similar trends Appendix Figure G12



(a) narrative voice interaction on forget set showing stark decline in scores from fine-tuned to NPO



(b) world building interaction on forget set showing stark decline in scores from fine-tuned to NPO with english decreasing more than dutch after NPO

Fig. 21 PSALM scores for fine tuned vs NPO models on the Forget set unlearning drastically decreases stylistic and content similarity across languages and splits and other evaluators show similar trends Appendix Figure G8

5.3.3 Stylistic and Structural Similarity

Figure 21 reports PSALM's forget-split stylistic and content scores before and after unlearning, while Appendix Figure G8 and G9 reports similar results for the other evaluators of the forget-split, as well as for the retain-split. On the forget split, writing style and narrative voice scores fall from near 1.0 under fine-tuning to roughly $0.25 - 0.35$ after NPO in both languages. Character similarity, plot structure, scene sequence and world-building show similar reductions, with means in the $0.15-0.35$ range. Rank-based ANOVA and post-hoc analyses in Appendix G confirm that these decreases are large and statistically reliable.

NPO does not only introduce superficial noise: on forget prompts, PSALM judges outputs as having different vocabulary, sentence structure, narrative voice, and character arcs. Average scores move from "near-identical" or "very similar" to around the boundary between "somewhat different" and "moderately similar", indicating that NPO reduces stylistic appropriation alongside literal memorisation, at least for the corpus studied here.

The reduction is nonetheless incomplete. Even on forget prompts the scores do not fall into the "clearly different" band (which would correspond to values near 0). After unlearning, some forget pairs are still classified as "very similar" or "near-identical" on individual dimensions, especially world-building and character similarity (Appendix Figure G14). The distribution has shifted downwards, but not collapsed to baseline. This means that unlearning substantially lowers the probability that a random output will exhibit high stylistic or structural similarity to the training passage, but it does not guarantee that no such outputs exist.

On the retain split, the picture is even more mixed. Stylistic scores for `en_npo` and `nl_npo` remain high (means around $0.83 - 0.86$), and content scores stay near 0.80. Compared to the fine-tuned models, this represents a noticeable softening of similarity, but the absolute levels are still in PSALM's "very similar" region. From a machine learning perspective, this is expected: NPO is optimised primarily to alter behaviour on the forget set while regularising the retain set. Even after unlearning, the models continue to generate outputs that are, on average, judged as highly similar in style and structure to works that remain in the training set. Unlearning therefore cannot be relied upon as a universal safeguard against stylistic appropriation; it may be more precise than a blanket reduction in capacity, but it does not restore models to a pre-fine-tuning state.

Finally, language effects are limited. Figure 21 shows a modest interaction for world-building similarity on the forget split: Dutch models retain slightly higher world-building similarity than English models after unlearning, despite starting from comparable fine-tuned levels. This suggests that the same NPO recipe may be somewhat less effective at disrupting complex world structures in Dutch than in English, possibly because the underlying pre-training distributions differ. For most other metrics, however, English and Dutch behave in parallel.

5.3.4 Behaviour of Exception Metrics

Figure 19 shows that the exception category changes far less than the computational, stylistic, and content categories, while Appendix Figures G10 and G11 additionally show how unlearning affects the four defence-oriented evaluators individually. Across all settings, PSALM almost never classifies outputs as strong parody, strong pastiche or clear quotation; mean scores stay well below the "strong defence" band. NPO does not change this qualitative picture, but it reshapes the balance between different exceptions.

On the forget split, quotation/citation scores decrease substantially after NPO. This is a direct consequence of the reduction in verbatim overlap: the outputs no longer contain long, clearly recognisable fragments of the source passage surrounded by commentary, so PSALM no longer treats them as quotations.

Pastiche and parody/satire show small upward shifts, especially on the retain split and especially for English. After unlearning, PSALM more often sees outputs as weak homages or as having a slight mocking or humorous character. Scores remain far below the "strong parody" or "strong pastiche" bands; there is little evidence that NPO pushes models into behaviour qualifying under EU parody or pastiche exceptions. Any increase in parody-like behaviour is a side effect of similarity reduction rather than an intentional transformation.

5.3.5 Residual Similarity and Distance to Baseline

Two complementary analyses assess whether NPO returns models to a state comparable to the baseline: paired comparisons between fine-tuned and NPO models, and equivalence tests comparing NPO models with baselines (Appendix G).

Paired Wilcoxon tests show that for forget prompts, NPO induces very large changes in all computational, stylistic and content metrics. For example, on English forget pairs, mean PSALM scores for plot structure drop by more than half, with effect sizes well into the "very large" range; similar patterns hold for narrative voice, character similarity, scene sequence and world-building in both languages. On the retain split, the corresponding changes are smaller but still often statistically significant. This confirms that NPO not only removes literal memorisation, but also substantially reshapes the model's expressive behaviour on the targeted works.

Equivalence tests using a tolerance of ± 0.1 on the $[0, 1]$ scale (roughly half a PSALM category) none of the unlearned models can be considered statistically equivalent to the baselines on any computational, stylistic or content metric (Table 5). Even after unlearning, the models remain noticeably more similar to the training works than the baseline models. For English, for instance, world-building similarity on the retain split decreases after NPO but remains well above the baseline level; the tests reject the hypothesis that the two are practically indistinguishable. This pattern holds for both languages and for both forget and retain sets.

Put differently, NPO moves models in the right direction for copyright compliance (it reduces literal copying and attenuates stylistic and structural overlap) but it does not fully "un-fine-tune" them. Appendix Figure G14 additionally visualises these score

Metric	Category	Lang.	Finetuned mean	NPO mean	Mean diff	TOST_p	Equivalent
Exact Match	Computational	EN	1.00	0.55	−0.45	0.99	No
BLEU	Computational	EN	1.00	0.77	−0.23	0.96	No
ROUGE	Computational	EN	1.00	0.81	−0.19	0.93	No
Writing Style	Stylistic	EN	0.99	0.82	−0.17	0.90	No
Narrative Voice	Stylistic	EN	0.99	0.86	−0.13	0.73	No
Character Similarity	Content	EN	0.92	0.82	−0.10	0.47	No
Plot Structure Similarity	Content	EN	0.98	0.79	−0.19	0.93	No
Scene Sequence Similarity	Content	EN	0.99	0.81	−0.18	0.93	No
World-building Similarity	Content	EN	0.98	0.89	−0.08	0.30	No
Exact Match	Computational	NL	0.95	0.45	−0.50	0.99	No
BLEU	Computational	NL	1.00	0.84	−0.16	0.86	No
ROUGE	Computational	NL	1.00	0.88	−0.12	0.68	No
Writing Style	Stylistic	NL	0.95	0.83	−0.12	0.64	No
Narrative Voice	Stylistic	NL	1.00	0.92	−0.09	0.31	No
Character Similarity	Content	NL	0.98	0.80	−0.18	0.90	No
Plot Structure Similarity	Content	NL	0.96	0.79	−0.17	0.82	No
Scene Sequence Similarity	Content	NL	0.96	0.76	−0.20	0.87	No
World-building Similarity	Content	NL	0.99	0.89	−0.10	0.53	No

Table 5 Two one-sided tests (TOST) for equivalence between fine-tuned and NPO models on the retain split for computational stylistic and content metrics where equivalence bounds are ± 0.1 on the $[0, 1]$ scale and TOST_p is the larger of the two one-sided p-values such that equivalence is rejected when $\text{TOST}_p > 0.05$

distributions and confirms that, although NPO shifts the distributions downward, many retain-set scores remain in the upper half of the scale.

6 Discussion

6.1 Core Findings and Technical Significance

6.1.1 Feasibility of Automated Similarity Assessment

The controlled validation experiments in Section 5.1 show that the ten PSALM evaluators achieve high strict alignment rates and low mean absolute error on purpose-designed test cases (Figures 11 and 12). The scatter of expected versus mean actual scores in Figure 13 is especially informative: almost all evaluator–test-case pairs fall within the ± 0.1 band, corresponding to at most one category difference on the five-point scale.

The DAG-based prompts and GPT-5-nano judge consistently apply the doctrinal and narratological criteria: errors are rare except at boundaries between adjacent similarity bands.

The higher variance and occasional large deviations for Character Similarity and Pastiche (Figures 13 and 14) expose a constraint that some constructs are harder for the judge model than others. This is not purely a modelling issue: even human experts disagree about when a character becomes sufficiently distinctive to be protectable, or when stylistic imitation crosses the line into lawful pastiche. The modest instability observed in these evaluators therefore reflects conceptual difficulty rather than a simple calibration error. For downstream use, this means that PSALM is more reliable as a detector of very high or very low similarity than as an arbiter of borderline cases.

A more fundamental constraint is that the validation corpus consists of researcher-designed test pairs constructed to instantiate prototypical similarity regimes. This procedure establishes internal consistency between the evaluator prompts and the judge's interpretations of those prompts. It does not establish that the resulting scores correspond to the judgements a copyright lawyer or court would reach on real disputed material. We therefore treat the validation results as evidence of measurement stability, not of legal validity. Establishing legal validity requires expert-annotated benchmarks (see also our future-work discussion in Section 6.4.1).

6.1.2 Fine-Tuning as Mechanism for Stylistic Appropriation

After fine-tuning, the English and Dutch models achieve near-ceiling scores across all infringement-oriented metrics (Figure 16). Table 4 demonstrates that similarity is not confined to lexical overlap: computational metrics (Exact Match, BLEU, ROUGE-L), stylistic metrics (Writing Style, Narrative Voice) and content metrics (Character, Plot, Scene Sequence, World-Building) all move from low or moderate values in the baselines to means around 0.95 or higher. The fine-tuned models therefore might not merely "remember" specific passages. When prompted with questions that resemble their training data, they tend to reproduce texts that PSALM judges as almost indistinguishable from those passages in style, structure and narrative content.

The magnitude of the stylistic effect exceeds what a characterisation of fine-tuning as a "light alignment layer" would predict. Here, however, the QA-style fine-tuning regime effectively rewrites the conditional distribution in the prompt region corresponding to the training questions. The model behaves less like a general-language assistant and more like a retrieval system for the specific works included in the fine-tuning corpus. This pattern is visible in the near-ceiling values in Figure 16, in the detailed forget-split stylistic plots in Figure 17, and additionally in the appendix split-level computational and retain-split stylistic plots (Appendix Figures F2 and F3). From a legal perspective, this means that fine-tuning can create a precision tool for reproducing specific works.

Second, fine-tuning largely erases the pre-training differences between languages. Before fine-tuning, English models already exhibit somewhat higher similarity to the training works than Dutch models, reflecting their English-centric pre-training. After tuning, Table 4 shows that English and Dutch converge to almost the same similarity levels, with differences of a few hundredths at most. This cross-lingual convergence is not trivial: the Dutch model must learn stylistic patterns from translations of historical Dutch prose into modern Dutch, whereas the English model sees translations into English. The convergence suggests that the QA format and the relatively small corpus encourage the models to internalise very specific patterns of narrative voice, characterisation and plot irrespective of language. The legal implication is that arguments based on the opacity of the pre-training corpus ("we do not know whether these works were included") become less relevant once clear evidence of post-training exposure exists: even a comparatively small fine-tuning corpus can dominate behaviour.

Equally important is what fine-tuning does *not* change. Figure 16 shows that the exception category changes far less than the infringement-oriented categories, and additionally the appendix split-level exception plots (Appendix Figures F5 and F6)

show that Parody/Satire and Pastiche remain low, Quotation/Citation becomes only moderate, and scènes-à-faire scores stay roughly constant. This means that the strengthened similarity is rarely accompanied by stronger evidence of lawful transformations or unprotectable stock elements. In legal terms, fine-tuning increases the plausibility that outputs will be considered reproductions or derivative works under EU standards, without simultaneously increasing the likelihood that they fall within exceptions or limitations (Lucchi and Hunter 2025; Chun 2024). A caveat applies: the QA dataset contains no parody, satire, or quotation prompts, so the low exception scores partly reflect the absence of relevant training signal. Nevertheless, this confirms that fine-tuning on domain-specific QA data does not automatically instil exception-like safeguards.

The QA format provides a particularly strong memorisation signal and may overstate appropriation relative to more varied instruction-tuning pipelines. However, domain-specific QA fine-tuning is common in commercial workflows; the RQ2 results are best read as a plausible upper bound on stylistic appropriation under strong exposure conditions.

6.1.3 Machine Unlearning: Partial Mitigation and Persistent Residuals

The post-unlearning experiments yield the most legally consequential finding of this study and provide the clean test of whether style-level similarity can be measured separately from verbatim copying. Figure 20 shows that NPO is effective at suppressing near-verbatim memorisation on the forget set: Exact Match scores drop to zero and BLEU/ROUGE-L means fall to levels that correspond to substantial paraphrasing. Stylistic and content metrics on the forget split also decrease appreciably (Figure 21), moving from the "near-identical" band to much lower values. This suggests that NPO may be changing the way the model tells the stories or retaining certain aspects, instead of unlearning the stories.

These declines, however, do not return the model to baseline behaviour (Table 5, Appendix Figure G14). Outputs on the forget split remain highly similar according to at least some PSALM dimensions, particularly world building and character similarity, and none of the metrics on the retain split become statistically equivalent to the baseline within a tolerance of ± 0.1. NPO therefore operates as a blunt instrument: it significantly shifts the score distributions but cannot surgically excise the influence of the forget set.

Quotation-Citation scores decline on the forget split after NPO (Appendix Figure G10), indicating that the unlearned models are less likely to embed recognisable fragments of the source passages with surrounding commentary. This is technically consistent with NPO's goal of reducing verbatim fidelity, but legally ambivalent. Quotations, when properly attributed and proportionate, can be lawful under Article 5(3)(d) of InfoSoc (European Parliament and Council 2001; Court of Justice of the European Union 2019b); by pushing outputs away from quotation-like behaviour, unlearning may remove one potential defence without fully eliminating expressive similarity. At the same time, scènes à faire scores rise moderately, which can be read as the model falling back on generic genre patterns once specific expressive choices have

been attenuated. From a doctrinal perspective this shift may be positive, as it moves more shared elements into the unprotectable domain; from a creative perspective it risks flattening outputs into formulaic narratives.

Cross-linguistic differences under NPO are small but revealing. The Dutch models retain slightly higher world-building similarity on the forget split than the English models, despite starting from comparable fine-tuned levels (Figure 21). One plausible explanation is that the Dutch translations preserve more of the spatial and institutional structure of the original works, making these features harder to disrupt without also harming general coherence. This highlights a tension: the aspects of works that are most structurally central (precisely those likely to be part of the author's "own intellectual creation") may also be the hardest to forget without significant utility loss.

NPO effectively reduces the most blatant forms of memorisation and shifts the overall distribution of stylistic similarity favourably, but non-trivial residual similarities remain, particularly for narrative structure and world-building. Unlearning should not be regarded as a guarantee of compliance; it demonstrates a reduction of risk, but not its complete removal.

6.1.4 Stylistic Memorisation as Distinct Phenomenon

A central conceptual contribution of the experiments is the empirical separation between verbatim memorisation and stylistic memorisation. The results show that NPO reduces computational metrics more strongly than stylistic and content metrics. Fine-tuning followed by unlearning can therefore produce models that rarely emit long verbatim sequences yet still generate outputs that PSALM judges as highly similar in narrative voice, characterisation or plot structure.

This finding has two important implications. First, it helps to explain why technical efforts focused solely on exact-match metrics, such as those captured by BLEU or longest common substrings, may give a false sense of security (Ippolito et al. 2023). A model can pass all computational overlap tests and still, in the sense operationalised by PSALM, "tell the same story in the same way". Second, it reinforces the need for evaluation frameworks that align more directly with legal tests. EU originality and substantial similarity standards (Court of Justice of the European Union 2009, 2019a; Lucchi and Hunter 2025) focus precisely on expressive choices in plot, character and style, not on verbatim copying as such.

6.2 Legal Interpretation and Compliance Practice

6.2.1 The Measurement-Determination Gap and Its Implications

Although PSALM delivers structured measurements, legal infringement remains a context-dependent determination. This gap between measurement and legal decision-making is already visible in the validation stage. For example, a Writing Style score of 0.8 indicates a high degree of similarity in lexical complexity, sentence structure and rhetorical patterns, but it does not by itself tell us whether the output competes with the original in the marketplace, whether it was produced with access to the protected work or whether the use is justified by a limitation or exception.

The gap is especially apparent in relation to exceptions and limitations. Article 5(5) InfoSoc (European Parliament and Council 2001; Geiger et al. 2014) requires that exceptions be applied only in certain special cases, not conflict with normal exploitation and not unreasonably prejudice legitimate interests. PSALM's parody, pastiche and quotation evaluators encode elements of the corresponding doctrinal tests, such as evocation, noticeable differences and legitimate purpose, but they do not incorporate evidence about market substitution, the existence of licensing schemes or the broader cultural context. For instance, the moderately increased scènes-à-faire scores after unlearning (Figure G10) suggest that more of the retained similarity lies in stock elements, but they do not by themselves prove that the use is free from conflict with normal exploitation.

A further limitation is that PSALM models only a subset of the exceptions and limitations potentially relevant under EU copyright law. It does not include dedicated evaluators for, among others, illustration for teaching or scientific research under Article 5(3)(a) InfoSoc, or incidental inclusion of a work in other material under Article 5(3)(i) InfoSoc. As a result, an output that contains a short excerpt for a legitimate teaching or research purpose, or that includes protected material only incidentally, may receive high scores on infringement-oriented dimensions without PSALM recognising a potentially applicable defence. The system may therefore overstate legal risk if its scores are read as a comprehensive assessment of all exceptions and limitations. Its exception-related outputs should instead be understood as limited to the specific doctrines operationalised in the current framework.

Accordingly, PSALM should be read as a diagnostic tool for similarity and exception-like features, consistent with our validation caveat in Section 6.1.1. Human judgment remains necessary for infringement determinations that depend on context such as market impact and the Article 5(5) three-step test.

6.2.2 TDM Exceptions, the Three-Step Test and Residual Similarity

The empirical results cast new light on ongoing debates about the scope of the TDM exceptions in the CDSM Directive and their interaction with the three-step test in Article 5(5) InfoSoc. If, as some commentators argue (Margoni and Kretschmer 2022; Quintais 2025; Lucchi and Hunter 2025), training LLMs on lawfully accessible works is covered by Articles 3 or 4 CDSM, a separate question is whether outputs that closely resemble those works remain within the shelter of the exception.

The fine-tuning experiments show that, when exposed to specific works, LLMs can and do generate outputs that are highly similar across multiple expressive dimensions (Figure 17; Appendix Figure F3) while exhibiting only weak parody, pastiche or quotation characteristics (Figure 16; Appendix Figures F5 and F6). The unlearning experiments then show that, even after targeted mitigation, a non-negligible tail of high-similarity outputs persists (Appendix Figure G14). Under a strict reading of Article 5(5), such residual similarity may still be considered conflicting with normal exploitation, especially if outputs can serve as substitutes for the original works. Under a more permissive reading, residual similarities that mainly involve scènes à faire and lack quotation-like reproduction might be treated as within the analytical scope of

TDM, particularly if models are combined with contractual or technical measures limiting harmful uses.

6.2.3 Interpretive Positions on Residual Similarity

The empirical patterns can be seen as compatible with several doctrinally grounded interpretive positions, rather than pointing to a single legal conclusion.

A strict liability position would treat any unauthorised reproduction of protected expression as infringing regardless of the technical mitigation steps taken. On this view, the high-similarity outliers that remain after NPO (Appendix Figure G14) are sufficient to impose liability for infringing outputs, though unlearning might influence the proportionality of remedies.

A materiality-based position would focus on whether residual similarities are substantial enough to matter. The movement of mean scores from around 1.0 to around 0.3 after NPO on the forget split (Figure 21) could be interpreted as a shift from clearly infringing behaviour towards minimal similarity, especially if the remaining overlap is concentrated in generic narrative patterns rather than detailed expressive choices. From this perspective, the main concern would be the extent of the high-similarity outliers, rather than the average.

A reasonable efforts position would emphasise process and due diligence. Here, PSALM could be used to argue that developers have taken meaningful steps to reduce the most problematic forms of similarity (verbatim copying, highly specific stylistic mimicry) while accepting that some residual risk remains. Such a position might be appealing to regulators seeking incentives for continuous improvement rather than binary judgments.

6.2.4 PSALM within Compliance Architectures

PSALM could become useful when embedded in broader compliance architectures. For model developers, Figures 16 and 19 show how it can be used to quantify the impact of design choices such as fine-tuning regimes and unlearning parameters. For regulators and auditors, PSALM provides a standardised basis for comparing models or mitigation methods against legally relevant dimensions. For rightsholders, high similarity scores after claimed compliance efforts could serve as a basis for contesting the adequacy of those efforts. Courts and legal experts could draw on PSALM analyses as one form of technical evidence among others when assessing substantial similarity and the applicability of exceptions.

6.3 Methodological Contributions and Limitations

6.3.1 Hierarchical Implementation of Legal Constructs

A methodological contribution of this work is the translation of legal constructs into hierarchical, DAG-based evaluators. By decomposing notions such as narrative voice, character similarity or parody into sub-dimensions with explicit prompts and weights (Section 3.2), PSALM tries to make the otherwise opaque judgement of an LLM-as-a-judge more transparent. The validation results in Figures 11 and 12 show that this architecture leads to coherent and controllable behaviour across diverse test cases.

The selection of sub-dimensions and their weights reflects a specific interpretive reading of EU copyright doctrine and narratology—one that is explicit and open to critique. These weights are heuristic rather than empirically learned or legally authoritative. They were not calibrated against expert-labelled infringement decisions, nor do they claim to estimate the relative importance that a court would assign to each factor in a concrete dispute. Different legal scholars might reasonably prioritise other sub-dimensions, assign greater weight to particular forms of protectable expression, or include additional contextual factors. PSALM's methodological advantage is therefore not that its weighting scheme is definitive, but that the scheme is explicit, inspectable and reconfigurable. This supports a more rigorous dialogue between developers and legal experts about how doctrine is operationalised.

Relatedly, the current DAG structure is static during inference and relies on fixed weighted aggregation. This is a deliberate simplification. Legal reasoning is not always additive: some elements operate as threshold conditions, some are relevant only if other conditions are satisfied, and some may be defeated by contextual considerations. A high score on one sub-dimension can mathematically compensate for a low score on another, whereas in legal reasoning the absence of a necessary element may defeat a claim or defence altogether. Conversely, a contextual factor not represented in the graph may be legally decisive despite having no effect on the aggregate score. PSALM's aggregate outputs should therefore be interpreted as diagnostic similarity measurements, not as formal models of legal syllogism or defeasible legal reasoning.

Future versions could combine DAG-based measurement with rule-based gates, expert-calibrated thresholds or argumentation-based models that better represent conditional and defeasible doctrinal structures. A particularly interesting direction is to extend the evaluation with multi-agent or argumentation-based components to improve audibility and to better represent conditional and defeasible aspects of legal reasoning.

6.3.2 Corpus and Model Constraints

The choice of corpus and model architecture introduces several limitations. The texts are drawn from historic public-domain Dutch literature, translated into modern Dutch and English using GPTOSS-120B and, for a subset, GPT-4o-mini (Section 4.1). This translation layer may itself impose stylistic regularities on the corpus. For instance, if the translators prefer certain syntactic structures or discourse markers, the fine-tuned models may learn to reproduce those patterns, and GPT-5-nano may treat them as strong signals of similarity. Manual spot-checking mitigates obvious errors, but subtler systematic biases introduced by the translators remain uncontrolled.

The reliance on LLaMA 3.2 1B Instruction is also a constraint. The model's small size was necessary given the training and unlearning workload and the available hardware (Section 4.6). The qualitative patterns of memorisation and unlearning are likely to generalise, but quantitative magnitudes may not: larger models might memorise more, require more aggressive unlearning, or respond differently to NPO.

6.3.3 LLM-as-a-Judge Considerations

Using GPT-5-nano as the judge model introduces dependencies on its training data, alignment, and internal biases. The controlled experiments in Section 5.1 indicate that GPT-5-nano can apply the PSALM prompts consistently, but they do not guarantee alignment with human legal reasoning. A related limitation is that the current implementation uses a single judge architecture across the nodes of an evaluator. The same underlying agent is therefore responsible for transforming, analysing, assessing, and judging. This design improves consistency and experimental control, but it also limits role specialisation. In legal and literary analysis, these dimensions may require different forms of expertise, different evidentiary standards, or different reasoning procedures that suit different agents.

A central limitation is therefore the absence of independent expert or human ground truth. The validation set tests whether the judge model applies PSALM's own evaluative criteria in a stable manner; it does not test whether those scores correlate with the assessments that copyright lawyers, literary experts, or courts would reach on real disputes. Accordingly, the reported alignment rates and error measures indicate internal reliability with respect to PSALM's prompt-defined criteria, not agreement with legal decisions (see Section 6.1.1).

Moreover, the overall pipeline employs different models for translation (GPTOSS-120B, GPT-4o-mini), generation (LLaMA) and judgement (GPT-5-nano). This heterogeneity reduces some risks of self-evaluation, but it also means that model-specific quirks at the translation stage may interact with the judge's preferences in ways that are hard to predict.

A further limitation is the lack of adversarial testing. All evaluations use relatively straightforward prompts. It remains unknown how robust GPT-5-nano's judgements are to strategically crafted inputs that attempt to manipulate similarity scores.

6.3.4 Experimental Design Trade-offs

The experimental design involves several trade-offs that influence how the results should be interpreted. The QA format strongly associates each question with a specific passage, accentuating memorisation but also approximating realistic use-cases where domain experts fine-tune models on QA pairs. The resulting evaluations are primarily passage-level rather than book-level. This makes the experiments tractable and allows controlled comparison between source passages and generated outputs, but it limits the framework's ability to detect similarities that emerge only across longer narrative spans, such as chapter-level pacing, cumulative character development, recurring motifs, or whole-work plot architecture.

The forget/retain partition combines author-level and book-level forgetting, exposing both coarse and fine-grained unlearning, yet it does not simulate long sequences of heterogeneous takedown requests. Nor does it test whether a generated chapter or full-length work remains substantially similar to a source book when similarity is distributed across many non-contiguous passages. Generation parameters are chosen to produce coherent, moderate-variance outputs; different settings might yield different memorisation profiles.

These trade-offs emphasise that the study examines a deliberately demanding but localised configuration. The results indicate what can happen under strong exposure to specific works and passage-level prompting, rather than constituting precise predictions for full-book generation, long-context comparison or large-scale deployment pipelines.

6.4 Future Research

6.4.1 Legal Expert Validation Studies

Validation against human copyright lawyers, literary scholars, and judges is the most pressing next step. A benchmark of source–target pairs drawn from the RQ2/RQ3 distributions, annotated along PSALM's dimensions and labelled for likely infringement under specified legal assumptions, would clarify whether current prompts and weights align with expert practice, expose under-specified doctrinal concepts, and address the normative question underlying RQ1: to what extent is LLM-based similarity assessment acceptable in legal decision-making?

6.4.2 Generalisation Across Architectures and Scales

Replicating the RQ2 and RQ3 experiments on larger models and different architectures is important for assessing generality. If stylistic convergence and partial unlearning behave similarly across model families, this strengthens the case for PSALM as a broadly applicable assessment tool. If not, the results would show that certain architectures are intrinsically more prone to problematic memorisation or more amenable to effective unlearning, which could inform regulatory guidance and model selection.

6.4.3 Multi-Agent and Ensemble-Based Judgement

A further direction is to replace the current single-judge architecture with multi-agent and ensemble-based evaluation. In the present implementation, each node in the PSALM graph is evaluated by the same judge configuration. This supports consistency, but it does not exploit possible specialisation across legal, literary and computational forms of assessment. Future work could assign different agents to different evaluative roles.

Ensemble judging could also improve robustness. Multiple judges, or multiple independently prompted instances of the same judge, could assess each source–target pair and expose disagreement between evaluators. Such disagreement would be especially informative for borderline cases, where a single scalar score may give a misleading impression of certainty. Aggregation could be performed through majority voting, weighted averaging, calibrated confidence intervals or explicit deliberation between agents. More ambitious versions could require judges to produce structured rationales or proof objects before a final score is assigned, allowing PSALM to distinguish between high-confidence similarity findings and cases where the apparent result depends on contestable assumptions. This would move the framework closer to an auditable evidentiary system rather than a single-pass scoring tool.

6.4.4 Cross-Domain and Cross-Lingual Extension

Extending PSALM beyond narrative fiction presents both technical and legal challenges. For journalism or academic writing, evaluators would need to capture argumentative structure, citation practices and factual overlap. For poetry or song lyrics, metrics of metre, rhyme and imagery would become relevant. At the same time, legal standards for originality and substantial similarity differ across domains, so doctrinal mapping would need to be revisited.

Cross-lingual extension beyond English and Dutch would test PSALM's doctrinal generality and expose latent language biases in the judge model, particularly for languages with morphological or script-level differences that affect how stylistic features are encoded.

6.4.5 Long-Form and Corpus-Scale Evaluation

Future work should also extend PSALM from passage-level comparison to chapter-level, whole-work and corpus-scale evaluation. The present experiments compare relatively short source passages with generated responses, which is appropriate for controlled validation and for measuring local memorisation. However, copyright-relevant similarity may be distributed across longer spans. A generated chapter may reproduce the pacing, sequence of revelations, character development or world-building logic of a source chapter without containing high lexical overlap with any single passage. Similarly, a generated book may appropriate the architecture of a source work through recurring motifs, cumulative plot structure or the arrangement of scenes across chapters.

Long-form evaluation raises technical challenges that are not addressed by the current implementation. Full books may exceed the context window of the judge model, making direct source–target comparison infeasible. Future versions could therefore combine retrieval, segmentation and hierarchical aggregation: first identifying potentially relevant source passages or chapters, then evaluating local similarities, and finally aggregating those findings into chapter-level or book-level assessments. Such an approach would also need to distinguish between isolated local overlap and systematic structural similarity across a work as a whole.

A related challenge is corpus-scale source discovery. In the present experiments, the relevant source passage is known in advance. Real compliance settings may instead require checking a generated text against a large library of books to identify possible sources of appropriation. This would require a retrieval stage capable of narrowing a large corpus to candidate works or passages before applying the more expensive PSALM evaluators. Future systems could combine embedding-based retrieval, approximate nearest-neighbour search, stylometric indexing and metadata-aware filtering with PSALM's legally structured similarity assessment. This would move the framework from pairwise evaluation toward large-scale auditing, where the central question is not only how similar two known texts are, but whether a given output is suspiciously similar to any protected work in a reference corpus.

6.4.6 Alternative Mitigation Approaches and Adversarial Evaluation

Finally, the NPO experiments suggest that unlearning can move models in the right direction but cannot serve as a complete solution. Comparing NPO with other unlearning techniques, influence-based data removal and inference-time controls under the same PSALM evaluation would create a more systematic picture of the trade-offs between forget quality, utility preservation and exception engagement. In parallel, adversarial prompting strategies aimed at maximising PSALM similarity scores would provide a worst-case analysis complementing the average-case view in this work. For legal risk assessment, understanding the worst case is often more relevant than understanding the typical case.

7 Conclusion

This research demonstrates that copyright-relevant stylistic appropriation in large language models can be measured systematically using automated evaluation frameworks grounded in EU legal doctrine. The PSALM framework operationalises substantial similarity assessments across computational, stylistic, and content dimensions and incorporates statutory defences for parody, pastiche, quotation and scènes à faire. Controlled validation on purpose-designed test cases yields strict alignment rates of about 95% and a mean absolute error of 0.037, indicating that PSALM can reliably distinguish clearly different, moderately similar and near-identical text pairs in a manner consistent with its doctrinal design.

The empirical findings show that supervised fine-tuning on a corpus of literary works induces pronounced stylistic appropriation extending well beyond verbatim memorisation. Fine-tuned Llama 3.2 1B models attain near-ceiling similarity scores across writing style, narrative voice, character construction, plot structure, scene sequences and world building. This high expressive overlap persists even when exact textual overlap is negligible. Machine unlearning via Negative Preference Optimisation substantially reduces similarity on the designated forget set and eliminates exact matches, but residual stylistic patterns remain detectable across multiple PSALM dimensions and the models do not revert to baseline behaviour. Unlearning therefore mitigates, but does not fully remove, the influence of the fine-tuning corpus.

These results suggest that technical safeguards focusing exclusively on verbatim memorisation may provide an incomplete picture of copyright risk. At the same time, PSALM's similarity scores are measurements rather than legal verdicts: the mapping between quantitative assessments and infringement determinations remains untested in the absence of validation by legal experts and judicial precedent. Future work should therefore compare PSALM's evaluations with expert legal judgements, replicate the experiments across larger architectures, domains and languages, examine adversarial robustness, and benchmark alternative mitigation approaches under the same legally informed metrics.

PSALM contributes to measurable copyright-compliance assessments and clarifies which aspects of current practice pose the greatest legal uncertainty. Technical measurement serves law rather than replacing it. PSALM offers structured evidence to inform human judgement. Closing the gap between verbatim-focused technical

practice and the broader substantial-similarity standards of EU copyright law will require sustained collaboration across legal, regulatory, and AI research communities, supported by evaluation infrastructure of the kind this work seeks to provide.

Acknowledgements

This research was supported by the Dutch Research Council (NWO) under the NWO-TDCC programme, project ICT.001.TDCC.014 (budget number 20656). We would like to express our gratitude to Rohan Nanda for reviewing this paper and providing valuable comments. We further thank Steven Claeyssens, Michel de Gruijter, and Mirjam Raaphorst (Koninklijke Bibliotheek, KB) for their expertise and continued support.

Data Availability

The source code for the PSALM framework, including all experimental pipelines, is publicly available at https://codeberg.org/nscharrenberg/PSALM (branch: paper-source-code). The corpus used for fine-tuning is publicly available on Hugging Face at https://huggingface.co/datasets/nscharrenberg/DBNL-public. The fine-tuned model is available at https://huggingface.co/nscharrenberg/LLaMA-3.2-1B-DBNL-Public-Domain-Finetuned and the unlearned model at https://huggingface.co/nscharrenberg/LLaMA-3.2-1B-DBNL-Public-Domain-Unlearned.

References

Abbott HP (2008) The Cambridge Introduction to Narrative, 2nd edn. Cambridge University Press, Cambridge, https://doi.org/10.1017/9781108913928

Axel Voss (2026) Report on copyright and generative artificial intelligence – opportunities and challenges. Official Journal of the European Union URL https://www.europarl.europa.eu/doceo/document/A-10-2026-0019%5FEN.html

Bhaila K, Van MH, Wu X (2025) Soft prompting for unlearning in large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 4046–4056, https://doi.org/10.18653/v1/2025.naacl-long.204

Biber D, Finegan E (1988) Adverbial stance types in english. Discourse Processes 11(1):1–34. https://doi.org/10.1080/01638538809544689

Bommarito II MJ, Bommarito J, Katz DM (2025) The kl3m data project: Copyright-clean training resources for large language models. arXiv https://doi.org/10.48550/arXiv.2504.07854, URL https://arxiv.org/abs/2504.07854v1, arXiv:2504.07854 [cs.CL]

Booth WC (1961) The Rhetoric of Fiction. University of Chicago Press, Chicago, Illinois, https://doi.org/10.1215/00382876-61-3-427

Bordwell D, Thompson K (2012) Film Art: An Introduction, 10th edn. McGraw-Hill, Columbus, Ohio

Borhi M, Khan B, Arnaudo A, et al (2025) The development of generative artificial intelligence from a copyright perspective. Tech. rep., European Union Intellectual Property Office (EUIPO), https://doi.org/10.2814/3893780, URL https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025, study

Carlini N, Tramèr F, Wallace E, et al (2021) Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, pp 2633–2650, URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Chatman S (1990) Coming to Terms: The Rhetoric of Narrative in Fiction and Film. Cornell University Press, Ithaca, New York

Chen T, Asai A, Mireshghallah N, et al (2024) CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 15134–15158, https://doi.org/10.18653/v1/2024.emnlp-main.844, URL https://aclanthology.org/2024.emnlp-main.844/

Chiang CH, Lee Hy (2023) Can large language models be an alternative to human evaluations? In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 15607–15631, https://doi.org/10.18653/v1/2023.acl-long.870, URL https://aclanthology.org/2023.acl-long.870/

Chien E, Wang H, Chen Z, et al (2024) Certified machine unlearning via noisy stochastic gradient descent. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 38852–38887, https://doi.org/10.52202/079017-1228, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/448abd486677165ceedfa790e9a61802-Paper-Conference.pdf

Chun J (2024) AIStorySimilarity: Quantifying story similarity using narrative for search, IP infringement, and guided creativity. In: Barak L, Alikhani M (eds) Proceedings of the 28th Conference on Computational Natural Language Learning. Association for Computational Linguistics, Miami, FL, USA, pp 161–177, https://doi.org/10.18653/v1/2024.conll-1.13, URL https://aclanthology.org/2024.conll-1.13/

Cooper AF, Gokaslan A, Ahmed A, et al (2025) Extracting memorized pieces of (copyrighted) books from open-weight language models. arXiv 2505.12546. https://doi.org/10.48550/arXiv.2505.12546, URL https://arxiv.org/abs/2505.12546, v2, arXiv:2505.12546 [cs.CL]

Court of Justice of the European Union (2009) Case c-5/08, infopaq international a/s v danske dagblades forening. EUR-Lex, URL https://curia.europa.eu/juris/liste.jsf?num=C-5/08, eCLI:EU:C:2009:465

Court of Justice of the European Union (2014) Case c-201/13, johan deckmyn and vrijheidsfonds vzw v helena vandersteen and others. EUR-Lex, URL https://curia.europa.eu/juris/liste.jsf?num=C-201/13, eCLI:EU:C:2014:2132

Court of Justice of the European Union (2019a) Case c-683/17, cofemel – sociedade de vestuário sa v g-star raw cv. EUR-Lex / CURIA, URL https://curia.europa.eu/juris/liste.jsf?num=C-683/17, eCLI:EU:C:2019:721

Court of Justice of the European Union (2019b) Funke medien nrw gmbh (Case C-469/17). URL https://curia.europa.eu/juris/liste.jsf?num=C-469/17, eCLI:EU:C:2019:623

Court of Justice of the European Union (2025) Case c-590/23, cg and yn v pelham gmbh and others. EUR-Lex, URL https://curia.europa.eu/juris/liste.jsf?num=C-590/23, eCLI:EU:C:2025:452

Cui J, Ning M, Li Z, et al (2024) Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. https://doi.org/10.48550/arXiv.2306.16092, URL https://arxiv.org/abs/2306.16092, arXiv:2306.16092

Dancyger K (2010) The Technique of Film and Video Editing: History, Theory, and Practice, 5th edn. Focal Press, Waltham, Massachusetts

Dong YR, Lin H, Belkin M, et al (2025) UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 8827–8840, https://doi.org/10.18653/v1/2025.naacl-long.444, URL https://aclanthology.org/2025.naacl-long.444/

Ekman S (2013) Here Be Dragons: Exploring Fantasy Maps and Settings. Wesleyan University Press, Middletown, Connecticut

European Parliament and Council (2001) Directive 2001/29/ec of the european parliament and of the council of 22 may 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. Official Journal of the

European Union URL https://eur-lex.europa.eu/eli/dir/2001/29/oj

European Parliament and Council (2019) Directive (eu) 2019/790 of the european parliament and of the council of 17 april 2019 on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec. Official Journal of the European Union URL https://eur-lex.europa.eu/eli/dir/2019/790/oj

European Parliament and Council (2024) Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union URL http://data.europa.eu/eli/reg/2024/1689/oj

Fludernik M (1996) Towards a 'Natural' Narratology. Routledge, Milton Park

Gavins J (2007) Text World Theory: An Introduction. Edinburgh University Press, Edinburgh

Gehrmann S, Adewumi T, Aggarwal K, et al (2021) The GEM benchmark: Natural language generation, its evaluation and metrics. In: Bosselut A, Durmus E, Gangal VP, et al (eds) Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). Association for Computational Linguistics, Online, pp 96–120, https://doi.org/10.18653/v1/2021.gem-1.10, URL https://aclanthology.org/2021.gem-1.10/

Geiger C, Gervais D, Senftleben M (2014) The three-step test revisited: How to use the test's flexibility in national copyright law. American University International Law Review 29(3):581–626. https://doi.org/10.2139/ssrn.2356619

Genette G (1980) Narrative Discourse: An Essay in Method. Cornell University Press, Ithaca, NY, translated by Jane E. Lewin; original French 1972

Hamburg District Court (2024) Germany – hamburg district court, 310 o 227/23, robert kneschke v LAION e.v. WIPO Lex, URL https://www.wipo.int/wipolex/en/text/592042, landgericht Hamburg, 310 O 227/23 (2024-09-27)

Havlikova S (2025) Technical challenges of rightsholders' opt-out from gen ai training after robert kneschke v. laion. Journal of Intellectual Property, Information Technology and Electronic Commerce Law 16(1). URL https://www.jipitec.eu/jipitec/article/view/422

Herman D, Jahn M, Ryan ML (eds) (2012) Routledge Encyclopedia of Narrative Theory. Routledge, Milton Park

Herrmann JB, van Dalen-Oskam K, Schöch C (2015) Revisiting style, a key concept in literary studies. Journal of Literary Theory 9:25–52. https://doi.org/10.1515/jlt-2015-0003, URL https://www.degruyterbrill.com/document/doi/10.1515/jlt-2015-0003/html

Hoover RM (1973) Prose rhythm: A theory of proportional distribution. College Composition and Communication 24(5):366–374. https://doi.org/10.58680/ccc197317628, URL 10.58680/ccc197317628

Ippolito D, Tramer F, Nasr M, et al (2023) Preventing generation of verbatim memorization in language models gives a false sense of privacy. In: Keet CM, Lee HY, Zarrieß S (eds) Proceedings of the 16th International Natural Language Generation Conference. Association for Computational Linguistics, Prague, Czechia, pp 28–53, https://doi.org/10.18653/v1/2023.inlg-main.3, URL https://aclanthology.org/2023.inlg-main.3/

Ji J, Liu Y, Zhang Y, et al (2024) Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 12581–12611, https://doi.org/10.52202/079017-0400, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/171291d8fed723c6dfc76330aa827ff8-Paper-Conference.pdf

Jia H, Li T, Guan J, et al (2025) The erasure illusion: Stress-testing the generalization of llm forgetting evaluation. https://doi.org/10.48550/arXiv.2512.19025, URL https://arxiv.org/abs/2512.19025, arXiv:2512.19025

jin Z, Cao P, Wang C, et al (2024) Rwku: Benchmarking real-world knowledge unlearning for large language models. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 98213–98263, https://doi.org/10.52202/079017-3117, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b1f78dfc9ca0156498241012aec4efa0-Paper-Datasets_and_Benchmarks_Track.pdf

Juola P (2008) Authorship attribution. Foundations and Trends in Information Retrieval 1(3):233–334. https://doi.org/10.1561/1500000005, URL https://doi.org/10.1561/1500000005, https://www.emerald.com/ftinr/article-pdf/1/3/233/11047725/1500000005en.pdf

Kallert C (2025) Urteil gema gegen open ai. Justiz Bayern URL https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php

Kandpal N, Wallace E, Raffel C (2022) Deduplicating training data mitigates privacy risks in language models. In: Chaudhuri K, Jegelka S, Song L, et al (eds) Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 162. PMLR, pp 10697–10707, URL https://proceedings.mlr.press/v162/kandpal22a.html

Katz DM, Bommarito MJ, Gao S, et al (2024) Gpt-4 passes the bar exam. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 382(2270):20230254. https://doi.org/10.1098/rsta.2023.0254, URL https:

//doi.org/10.1098/rsta.2023.0254, https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2023.0254/1328474/rsta.2023.0254.pdf

Koninklijke Bibliotheek van Nederland (2025) Collectie publiek domein - dbnl. URL https://www.dbnl.org/letterkunde/pd/index.php

Kurtz LA, Borod RS (2015) A Practical Approach to Licensing Motion Pictures, Video, and Other Audiovisual Works, 3rd edn. American Bar Association, Chicago, Illinois

Lee K, Ippolito D, Nystrom A, et al (2022) Deduplicating training data makes language models better. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 8424–8445, https://doi.org/10.18653/v1/2022.acl-long.577, URL https://aclanthology.org/2022.acl-long.577/

Lucchi N, Hunter S (2025) Generative ai and copyright: Training, creation, regulation. Study PE 774.095, European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs (JURI), Brussels, URL https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST%5FSTU(2025)774095%5FEN.pdf, requested by the Committee on Legal Affairs (JURI)

Maini P, Feng Z, Schwarzschild A, et al (2024) Tofu: A task of fictitious unlearning for llms. arXiv https://doi.org/https://doi.org/10.48550/arXiv.2401.06121, URL http://arxiv.org/abs/2401.06121, arXiv:2401.06121 [cs.LG]

Margoni T, Kretschmer M (2022) A deeper look into the eu text and data mining exceptions: Harmonisation, data ownership, and the future of technology. GRUR International 71(8). https://doi.org/10.1093/grurint/ikac054, URL https://academic.oup.com/grurint/article/71/8/685/6650009

McKee R (1997) Story: Substance, Structure, Style and the Principles of Screenwriting. ReganBooks, New York City

Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. Advances in Health Sciences Education 15(5):625–632. https://doi.org/10.1007/s10459-010-9222-y, URL https://link.springer.com/article/10.1007/s10459-010-9222-y

Novikova J, Dušek O, Cercas Curry A, et al (2017) Why we need new evaluation metrics for NLG. In: Palmer M, Hwa R, Riedel S (eds) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 2241–2252, https://doi.org/10.18653/v1/D17-1238, URL https://aclanthology.org/D17-1238/

Palmer A (2004) Fictional Minds. University of Nebraska Press, Lincoln, Nebraska

Phelan J (2005) Living to Tell about It: A Rhetoric and Ethics of Character Narration. Cornell University Press, Ithaca, New York

Quintais JaP (2025) Generative ai, copyright and the ai act. Computer Law & Security Review 56. https://doi.org/10.1016/j.clsr.2025.106107, URL https://www.sciencedirect.com/science/article/pii/S0267364925000020

Rendas T (2024) The pastiche exception after Pelham II: Implications for generative AI. European Intellectual Property Review 46(12):721–730

Rosati E (2013) Chapter 3: Originality in a work, or a work of originality: The effects of the Infopaq decision: Full Harmonization through Case Law, Edward Elgar Publishing, Cheltenham, UK. https://doi.org/10.4337/9781782548942.00013, URL https://www.elgaronline.com/view/9781782548935.00013.xml

Russinovich M, Salem A (2025) Obliviate: Efficient unmemorization for protecting intellectual property in large language models. arXiv https://doi.org/10.48550/arXiv.2502.15010, URL https://arxiv.org/abs/2502.15010, arXiv:2502.15010 [cs.LG]

Ryan ML (2007) Toward a definition of narrative. In: Herman D (ed) The Cambridge Companion to Narrative. Cambridge University Press, Cambridge, p 22–35

Sai AB, Mohankumar AK, Khapra MM (2022) A survey of evaluation metrics used for nlg systems. ACM Comput Surv 55(2). https://doi.org/10.1145/3485766, URL https://doi.org/10.1145/3485766

Scharrenberg N, Sun C (2025) Copyright infringement by large language models in the EU: Misalignment, safeguards, and the path forward. In: Aletras N, Chalkidis I, Barrett L, et al (eds) Proceedings of the Natural Legal Language Processing Workshop 2025. Association for Computational Linguistics, Suzhou, China, pp 125–134, URL https://aclanthology.org/2025.nllp-1.9/

Shi W, Lee J, Huang Y, et al (2025) Muse: Machine unlearning six-way evaluation for language models. In: Yue Y, Garg A, Peng N, et al (eds) International Conference on Learning Representations, pp 27797–27818, URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4556f5398bd2c61bd7500e306b4e560a-Paper-Conference.pdf

Stamatatos E (2009) A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60:538–556. https://doi.org/10.1002/ASI.21001, URL https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21001https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001https://onlinelibrary.wiley.com/doi/10.1002/asi.21001

Stanzel FK (1984) A Theory of Narrative. Cambridge University Press, Cambridge, c. Goedsche, Trans. Original work published 1979

Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680. https://doi.org/10.1126/science.103.2684.677, URL https://www.science.org/doi/abs/10.1126/science.103.2684.677, https://www.science.org/doi/pdf/10.1126/science.103.2684.677

United States Court of Appeals for the Second Circuit (1930) Nichols v. universal pictures corp. 45 F.2d 119 (2d Cir. 1930), URL https://law.justia.com/cases/federal/appellate-courts/F2/45/119/1489834/

United States Court of Appeals for the Second Circuit (1980) Hoehling v. universal city studios, inc. 618 F.2d 972 (2d Cir. 1980), URL https://law.justia.com/cases/federal/appellate-courts/F2/618/972/407386/

United States District Court, S.D. New York. (1986) Walker v. time life films, inc. 784 F.2d 44 (2d Cir. 1986), URL https://law.justia.com/cases/federal/district-courts/FSupp/615/430/1515188/

United States District Court, S.D. New York (2008) Warner bros. entertainment inc. v. rdr books. 575 F. Supp. 2d 513 (S.D.N.Y. 2008), URL https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2007cv09667/315790/92/

United States Supreme Court (1936) Sheldon v. metro-goldwyn pictures corp. 81 F.2d 49 (2d Cir. 1936), URL https://supreme.justia.com/cases/federal/us/309/390/

Vasilev S, Herold C, Liao B, et al (2025) Unilogit: Robust machine unlearning for LLMs using uniform-target self-distillation. In: Che W, Nabende J, Shutova E, et al (eds) Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 22453–22472, https://doi.org/10.18653/v1/2025.findings-acl.1154, URL https://aclanthology.org/2025.findings-acl.1154/

Warso Z, Gahntz M (2024) Advancing training data transparency in the eu ai act. Open Future Blog, URL https://openfuture.eu/blog/advancing-training-data-transparency-in-the-eu-ai-act/, blog post

Wei B, Shi W, Huang Y, et al (2024) Evaluating copyright takedown methods for language models. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 139114–139150, https://doi.org/10.52202/079017-4415, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/faed4276b52ef762879db4142655c699-Paper-Datasets_and_Benchmarks_Track.pdf

Wolf MJP (2012) Building Imaginary Worlds: The Theory and History of Subcreation. Routledge, Milton Park

Zhang R, Lin L, Bai Y, et al (2024) Negative preference optimization: From catastrophic collapse to effective unlearning. URL http://arxiv.org/abs/2404.05868

Zheng L, Chiang WL, Sheng Y, et al (2023) Judging llm-as-a-judge with mt-bench and chatbot arena. In: Oh A, Naumann T, Globerson A, et al (eds) Advances in Neural Information Processing Systems, vol 36. Curran Associates, Inc., pp 46595–46623, URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

Appendix A Supplementary Material

The following Online Resources are submitted as supplementary files alongside this article.

Online Resource 1 `ESM_1.csv` — Evaluator-level validation summary statistics for all ten PSALM evaluators, including MAE, RMSE, EXACT/CLOSE/ACCEPTABLE/POOR counts, strict alignment rate, and coefficient of variation of absolute error.
Online Resource 2 `ESM_2.csv` — Per-evaluator, per-test-case statistics from the controlled validation, including expected score, mean and standard deviation of actual scores over five replications, and minimum and maximum observed values.
Online Resource 3 `ESM_3.csv` — Raw per-run validation results for all evaluator–test-case combinations, including individual scores, alignment band classifications, and free-text reasoning fields.
Online Resource 4 `ESM_4.csv` — Shapiro–Wilk normality and Levene homogeneity test results for all RQ2 metrics on the forget split.
Online Resource 5 `ESM_5.csv` — Shapiro–Wilk normality and Levene homogeneity test results for all RQ2 metrics on the retain split.
Online Resource 6 `ESM_6.csv` — Kruskal–Wallis omnibus statistics and per-condition descriptive statistics for all thirteen metrics across the four model conditions (RQ2, both splits).
Online Resource 7 `ESM_7.csv` — Cohen's d and Cliff's δ effect sizes for all pairwise model contrasts on RQ2 metrics, forget split.
Online Resource 8 `ESM_8.csv` — Cohen's d and Cliff's δ effect sizes for all pairwise model contrasts on RQ2 metrics, retain split.
Online Resource 9 `ESM_9.csv` — Per-category descriptive statistics (mean, standard deviation) for all four model conditions across RQ2 metric categories (Computational, Stylistic, Content, Exceptions).
Online Resource 10 `ESM_10.csv` — Rank-based overall model rankings aggregated across all thirteen RQ2 metrics for forget and retain splits.
Online Resource 11 `ESM_11.csv` — Category-level differential-effect contrasts (baseline minus fine-tuned) for RQ2, forget split.
Online Resource 12 `ESM_12.csv` — Rank-based ART-ANOVA results for language × training-stage interactions on RQ2 metrics, forget split.
Online Resource 13 `ESM_13.csv` — Rank-based ART-ANOVA results for language × training-stage interactions on RQ2 metrics, retain split.

Online Resource 14 `ESM_14.csv` — Paired Wilcoxon signed-rank test results comparing baseline and fine-tuned models on shared prompts, RQ2 forget split.
Online Resource 15 `ESM_15.csv` — Paired Wilcoxon signed-rank test results comparing baseline and fine-tuned models on shared prompts, RQ2 retain split.
Online Resource 16 `ESM_16.csv` — Shapiro–Wilk normality and Levene homogeneity test results for all RQ3 metrics on the forget split.
Online Resource 17 `ESM_17.csv` — Shapiro–Wilk normality and Levene homogeneity test results for all RQ3 metrics on the retain split.
Online Resource 18 `ESM_18.csv` — Kruskal–Wallis omnibus statistics and per-condition descriptive statistics for all thirteen metrics across model conditions (RQ3, both splits).
Online Resource 19 `ESM_19.csv` — Rank-based ART-ANOVA results for language × unlearning interactions on RQ3 metrics, forget split.
Online Resource 20 `ESM_20.csv` — Rank-based ART-ANOVA results for language × unlearning interactions on RQ3 metrics, retain split.
Online Resource 21 `ESM_21.csv` — Rank-based overall model rankings aggregated across all thirteen RQ3 metrics for forget and retain splits.
Online Resource 22 `ESM_22.csv` — Per-category descriptive statistics for all model conditions across RQ3 metric categories.
Online Resource 23 `ESM_23.csv` — Category-level differential-effect contrasts (fine-tuned minus NPO) for RQ3, forget split.
Online Resource 24 `ESM_24.csv` — Paired Wilcoxon signed-rank test results comparing fine-tuned and NPO models on shared prompts, RQ3 forget split.
Online Resource 25 `ESM_25.csv` — Paired Wilcoxon signed-rank test results comparing fine-tuned and NPO models on shared prompts, RQ3 retain split.
Online Resource 26 `ESM_26.csv` — Cohen's d and Cliff's δ effect sizes for pairwise model contrasts on RQ3 metrics, forget split.
Online Resource 27 `ESM_27.csv` — Cohen's d and Cliff's δ effect sizes for pairwise model contrasts on RQ3 metrics, retain split.
Online Resource 28 `ESM_28.csv` — Two one-sided tests (TOST) for equivalence between baseline and NPO models on all metrics, retain split.

Appendix B Hyperparameter Specifications

B.1 Output Generation

The output generation parameters are shown in Table B1.

Parameter	Value
Sampling Method	Nucleus (top-p)
Top-P	0.9
Temperature	0.1
Max New Tokens	1024
Repetition Penalty	1.1

Table B1 Output Generation Parameters

Appendix C Controlled Validation Corpus

C.1 Test Case Design Principles

The controlled validation set comprises 50 unique text pairs (10 evaluators × 5 test cases) designed to isolate individual PSALM evaluators and verify dimensional specificity. Each test case instantiates one point on the evaluator's five-level taxonomy:

- **TC1 (0.0):** Texts exhibit fundamental opposition or complete absence of similarity on the target dimension while potentially sharing characteristics on orthogonal dimensions.
- **TC2 (0.3):** Texts share superficial/minimal overlap on the target dimension while diverging substantially in execution depth or specific elaborations.
- **TC3 (0.5):** Texts demonstrate balanced presence of both similarities and differences, occupying the decision boundary region where classification becomes ambiguous.
- **TC4 (0.8):** Texts exhibit strong alignment on the target dimension with only small variations insufficient to alter overall classification.
- **TC5 (1.0):** Texts match precisely or near-precisely on all sub-dimensions of the evaluator's assessment criteria.

C.2 Text Pair Construction Protocol

For each evaluator-test case combination:

1. **Source text generation:** Manually author or curate a passage (150-300 words) demonstrating distinctive characteristics on the target dimension. For example, for Writing Style TC5, create text with highly specific lexical patterns, sentence structures, and rhetorical devices.
2. **Target text engineering:** Construct target text through controlled modifications:
   - **TC1:** Invert or oppose dimensional characteristics (e.g., first-person → third-person omniscient for Narrative Voice)
   - **TC2:** Preserve 1-2 sub-dimensional features whilst altering others (e.g., maintain sentence length distribution but change lexical complexity for Writing Style)
   - **TC3:** Preserve half of sub-dimensions, alter half
   - **TC4:** Preserve all but 1-2 minor sub-dimensional features
   - **TC5:** Create near-identical text with minimal cosmetic changes (character name substitutions, setting transpositions)
3. **Orthogonality verification:** Ensure modifications isolate the target dimension without introducing confounding changes on other dimensions. For example, when testing Character Similarity TC3, plot structure and world-building should remain constant between source and target.
4. **Internal review:** Independent reviewers assess whether text pairs achieve intended similarity level and dimensional isolation. Ambiguous pairs are revised until consensus is reached.

C.3 Validation Metrics and Criteria

Each text pair receives five independent evaluations (different API calls to GPT-5-nano with temperature = 0). This replication enables assessment of:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **Alignment classifications:**
  - EXACT: $|\Delta| \leq 0.1$
  - CLOSE: $0.1 < |\Delta| \leq 0.2$
  - ACCEPTABLE: $0.2 < |\Delta| \leq 0.3$
  - POOR: $|\Delta| > 0.3$
- **Success rates:**

$$\text{Alignment Rate} = \frac{N_{\text{EXACT}} + N_{\text{CLOSE}} + N_{\text{ACCEPTABLE}}}{N_{\text{total}}} \tag{C1}$$

$$\text{Strict Alignment Rate} = \frac{N_{\text{EXACT}} + N_{\text{CLOSE}}}{N_{\text{total}}} \tag{C2}$$

C.4 Validation Status Assignment

Evaluators are classified as:

- **Validated:** Strict Alignment Rate ≥ 0.8 **and** MAE ≤ 0.15
- **Partially Validated:** Alignment Rate ≥ 0.7 **and** MAE ≤ 0.25
- **Failed:** Does not meet Partially Validated criteria

Only Validated evaluators proceed to corpus evaluation. Partially Validated evaluators may be used for exploratory analysis with explicit acknowledgment of elevated measurement error. Failed evaluators are excluded until re-designed and re-validated.

C.5 Iterative Refinement Protocol

Evaluators failing validation undergo:

1. **Error analysis:** Inspect POOR-aligned cases to diagnose failure modes:
   - Systematic bias (consistent over-/under-estimation)
   - Boundary confusion (misclassification near decision thresholds)
   - Sub-dimension imbalance (over-weighting specific DAG nodes)
   - Stochastic instability (high replication variance)
2. **Prompt refinement:** Modify evaluator prompts to clarify dimensional definitions, add explicit boundary-case examples, or adjust sub-node weighting instructions
3. **Re-evaluation:** Re-run validation with modified configuration
4. **Documentation:** Archive refinement history with rationale for design decisions

All validation data (text pairs, expected scores, observed scores, alignment classifications) are archived in supplementary materials.

Appendix D Extended Statistical Analysis Details

D.1 Multiple Testing Correction Strategies

D.1.1 Family Definitions

Correction families are defined as:

- **Per-metric families:** All pairwise comparisons within one metric (e.g., all Writing Style comparisons: en_b vs. en_f, en_f vs. en_{npo}, etc.)
- **Per-category families:** Differential effect comparisons within categories (e.g., Computational vs. Stylistic vs. Content vs. Exceptions)
- **Independent families:** Across research questions (RQ2 and RQ3 comparisons use separate families)

D.2 Reporting Standards

All statistical results report:

1. **Test identification:** Name and type (e.g., "Mann-Whitney U test", "Wilcoxon signed-rank test")
2. **Test statistic:** Numerical value with symbol (e.g., $U = 234.5$, $W = 120$, $H = 45.32$)
3. **Degrees of freedom:** Where applicable (e.g., $df = 78$ for paired t-test)
4. **p-value:** Exact when $p > 0.001$, otherwise "$p < 0.001$"
5. **Effect size:** Point estimate with 95% CI (e.g., "$d = -2.34$ [95% CI: -2.89, -1.79]")
6. **Interpretation:** Explicit significance statement (e.g., "significant at $\alpha = 0.05$" or "non-significant")

Non-significant results are explicitly reported rather than omitted to ensure transparency about null findings.

Appendix E RQ1: Detailed Validation Statistics

This appendix summarises the core supplementary findings from the controlled validation experiments for RQ1, using the supplementary data for RQ1. The aim is to document the main numerical patterns that are not already discussed in subsection 5.1.

E.1 Evaluator-level performance

Table E2 gives a compact category-level view of the evaluator statistics. For each group of evaluators it reports the range of mean absolute error (MAE) and of the strict alignment rate (EXACT+CLOSE, i.e. predictions within ±0.2 of the expected score), computed directly from Online Resource 1.

Across all ten evaluators, MAE lies between 0.028 and 0.056, with RMSE between 0.082 and 0.136. Strict alignment is either 0.92 or 0.96 for every evaluator, and the overall distribution derived from Online Resource 1 shows 211 EXACT and 26 CLOSE alignments out of 250 runs. The remaining 13 runs are POOR, with no

Table E2 Category-level summary of RQ1 validation results derived from Online Resource 1. Strict alignment is the proportion of runs in EXACT or CLOSE bands

Category	Evaluators	MAE range	Strict alignment range
Stylistic (Narrative Voice, Writing Style)	2	0.028–0.048	0.92–0.96
Content (Character, Plot, Scene, World-building)	4	0.028–0.056	0.92–0.96
Exceptions (Parody/Satire, Pastiche, Quotation/Citation)	3	0.028–0.056	0.96–0.96
Other (Scènes à faire)	1	0.044	0.96

ACCEPTABLE-only cases. Thus, 237/250 evaluations fall within ±0.2 of the intended similarity level and 13/250 fall more than one full category away.

Within categories, the two stylistic evaluators differ only modestly: Narrative Voice has the lowest MAE (0.028) and highest strict alignment (0.96), while Writing Style has slightly higher MAE (0.048) and strict alignment of 0.92. The content evaluators share MAE values between 0.028 and 0.056; Character Similarity is the least accurate of this group with MAE 0.056 and two POOR alignments, while Plot Structure, Scene Sequence and World-building each attain MAE 0.028 and a single POOR run. Among the exception evaluators, Quotation/Citation matches the best-performing content metrics (MAE 0.028, strict alignment 0.96), Parody/Satire sits in the middle (MAE 0.036), and Pastiche mirrors Character Similarity (MAE 0.056, two POOR runs). The Scènes à faire evaluator, categorised as *Other* in the CSVs, has MAE 0.044 and strict alignment 0.96, situated between the most and least accurate evaluators.

The last column of Online Resource 1 reports the coefficient of variation of absolute error, which ranges from approximately 2.08 to 2.83 across evaluators. This combination of small MAE and relatively high coefficients of variation reflects the skewed error distribution already described in the main text: most runs have zero error, with occasional deviations of 0.3 or 0.5.

E.2 Test-case behaviour

Online Resource 2 records, for each evaluator and each of the five test cases, the expected score, the mean and standard deviation of the actual scores across the five replications, and the minimum and maximum values observed.

For all evaluators, the extreme conditions behave as intended. When the expected score is 0.0 (TC1, "clearly different or opposite") the mean actual score is exactly 0.0 with zero standard deviation and all runs classified as EXACT. When the expected score is 1.0 (TC5, "identical or near-identical"), the mean actual score is 0.96 for every evaluator, with standard deviation 0.089…, minimum 0.8 and maximum 1.0; the resulting mean absolute error is 0.04 and strict alignment remains EXACT.

The mid-range conditions show the small systematic deviations that are only qualitatively described in the main text. For the "very similar" test cases (TC4, expected

score 0.8), the mean actual score lies between 0.64 and 0.74 depending on the evaluator. Character Similarity and Pastiche are the most conservative, with mean 0.64 and standard deviation 0.230 . . ., whereas Narrative Voice, Plot Structure, Scene Sequence, World-building, Writing Style, Parody/Satire and Quotation/Citation all have mean 0.74 and standard deviation 0.134 . . .. In all these cases the mean absolute error is at most 0.16, and the summary file labels them as EXACT overall because the deviations remain within the ±0.2 band.

For the "moderately similar" cases (TC3, expected 0.5), most evaluators have mean actual scores close to 0.46 with standard deviation 0.089 . . ., again leading to a mean absolute error of 0.04 and EXACT alignment at the test-case level. Writing Style is slightly different: its "moderately similar" case shows mean actual 0.52, standard deviation 0.178 . . ., and mean absolute error 0.10, indicating more spread and a small upward bias relative to 0.5. For the "somewhat different" cases (TC2, expected 0.3), mean actual scores are either exactly 0.3 (e.g. Narrative Voice, Plot Structure, Scene Sequence, World-building, Quotation/Citation and several exception metrics) or 0.34 with standard deviation 0.089 . . . (Character Similarity, Parody/Satire, Pastiche, Scenes à faire, Writing Style). These values correspond to the alignment label EXACT in the summary CSV because they stay within 0.04 of the target.

E.3 Alignment distribution and failure modes

The raw per-run data (Online Resource 3) allow the 13 POOR alignments reported in Online Resource 1 to be localised to particular evaluators and test cases. All POOR cases occur in the middle or upper-middle of the scale and none appear in the "clearly different" or "identical" conditions.

For Character Similarity, both POOR runs arise in the "very similar characters with only minor differences" test case. The reasoning fields attribute these errors to "weighting noise" and to over-emphasis on "lexical and environmental divergence" when structural and arc-level features were near-identical. Narrative Voice has a single POOR case, also in the "very similar" condition, where the judge model reports that a collective voice pattern was not recognised and the verdict was demoted to "moderately similar". Parody/Satire's only POOR alignment occurs in the "good parody" test case, with the explanation that subtle humour was under-recognised.

Pastiche has two POOR alignments in the "good pastiche" condition. In one run the reasoning notes that the Artistic Skill node was down-weighted; in the other, tonal warmth was interpreted as divergence from the source rather than as part of a respectful homage. For Plot Structure, Scene Sequence and World-building, the single POOR alignment in each case is again attached to the "very similar" test case; the reasoning refers to omitted crisis–repair links, pacing variance and phrasing simplification in specific sub-nodes. Quotation/Citation has one POOR run in the "good quotation" test case, where prior disclosure or identification is under-weighted. Writing Style has two POOR cases, one where a change in rhythm leads to a downgrade from "very similar" to "moderately similar", and one where rhythmic correspondence is over-weighted and a "moderately similar" pair is pushed up to "very similar".

Across all evaluators, these POOR cases share two features that are not explicitly spelled out in the main text. First, they always concern borderline distinctions between

adjacent verbal bands, such as “very similar” versus “moderately similar”, and never represent confusions between opposite ends of the scale. Secondly, the reasoning strings consistently point to shifts in the relative salience of sub-dimensions—lexical versus structural features in Character Similarity, rhythmic versus syntactic features in Writing Style, or artistic execution versus tribute character in Pastiche—rather than to complete misinterpretations of the task instructions. This supports the view that the residual errors are primarily boundary and weighting effects within otherwise well-calibrated evaluators.

E.4 Test-case Difficulty and Boundary Effects

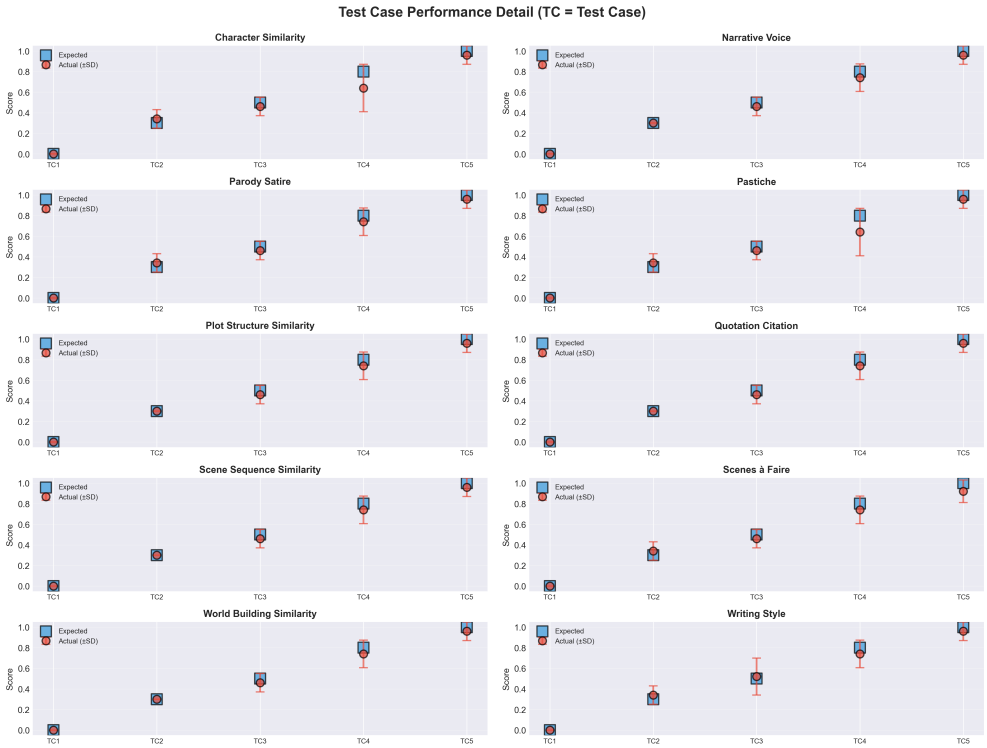


Fig. E1 Test case performance across evaluators where for each evaluator blue bars denote expected scores for test cases TC1 through TC5 and red circles with error bars show mean and standard deviation of actual scores over five replications

Figure E1 disaggregates performance by test case. The extreme conditions behave as intended: all evaluators return scores at or indistinguishable from 0 for TC1 (unrelated texts) and near 1 for TC5 (near-identical texts), confirming PSALM reliably separates obviously non-infringing from near-verbatim pairs.

The more interesting behaviour emerges in the intermediate test cases. The intermediate cases (TC2–TC4) show noticeably higher replication variance, with scores occasionally encroaching on adjacent bands — TC4 pairs pushed into TC3, or TC3 pairs lifted slightly. This is most pronounced for Writing Style and Character Similarity, where subtle gradations are inherently less discrete than, for example, the presence or absence of a quotation.

This behaviour justifies the five-level rubric: evaluators reliably distinguish the ends of the spectrum and generally separate "moderate" from "very" similarity, but finer partitioning would amplify instability in the central region. Mid-range spread is better understood as a property of the underlying construct (borderline cases are genuinely ambiguous) than as a measurement defect.

Appendix F RQ2: Detailed Statistical Analyses

This appendix summarises the core supplementary findings for RQ2, using the supplementary data. The aim is to document the main numerical patterns that are not already discussed in subsection 5.2.

F.1 Literal Reproduction

We first examine literal reproduction as captured by Exact Match, BLEU and ROUGE. Figure F2 displays the mean scores for these metrics before and after fine-tuning, for both languages and splits. Baseline models produce near-zero scores on all three metrics, with the English baseline showing marginally higher overlap than the Dutch (consistent with LLaMA 3.2's English-centric pre-training) but both well below any threshold associated with systematic memorisation.

After fine-tuning, Exact Match and ROUGE means approach one for both languages and splits, with BLEU similarly near ceiling, implying that almost all generated answers are exact or near-exact reproductions of the corresponding corpus passages. Because Exact Match is binary at the level of individual outputs, a mean of 0.9 for the Dutch forget set, for example, means that more than 90% of outputs are exact substring copies of the source passage used as the answer in the QA pair.

Kruskal–Wallis tests confirm that model condition has a highly significant effect on all three computational metrics for both splits, and post-hoc Mann–Whitney tests show that each fine-tuned model differs strongly from each baseline. Paired Wilcoxon tests on shared prompts show very large effect sizes when comparing each baseline with its fine-tuned counterpart.

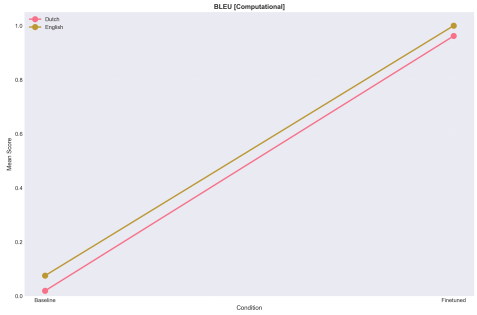
From a copyright perspective, these results indicate that fine-tuning turns the models into near-perfect retrievers of the training passages rather than models that merely internalise style. Were the underlying works protected rather than public domain, such behaviour would present a clear risk of verbatim infringement.

F.2 Stylistic and Structural Appropriation Extended

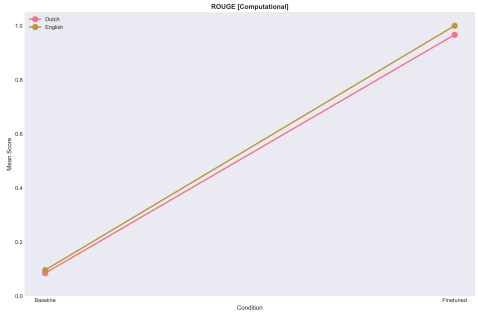
The results of stylistic and content similarity before and after fine-tuning on the forget and retain sets can be seen in Figure F3 and Figure F4.

F.3 Behaviour of Exceptions Extended

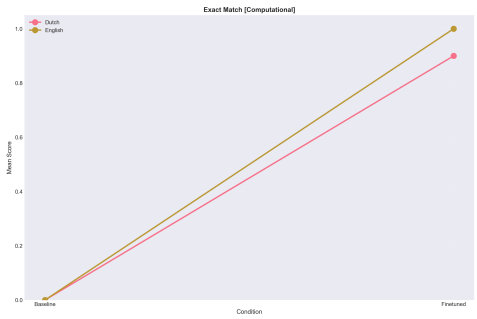
The results of exception metrics before and after fine-tuning on forget and retain sets can be seen in Figure F5 and Figure F6.



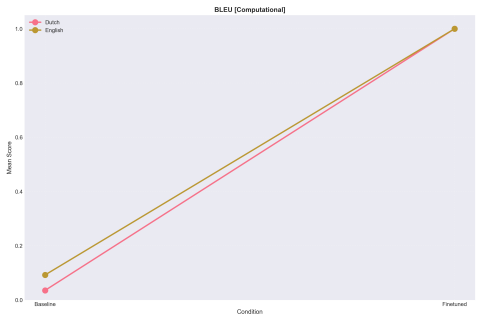
(a) BLEU Interaction of Forget Set



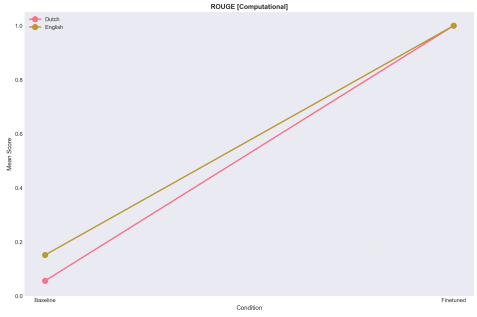
(b) ROUGE-L Interaction of Forget Set



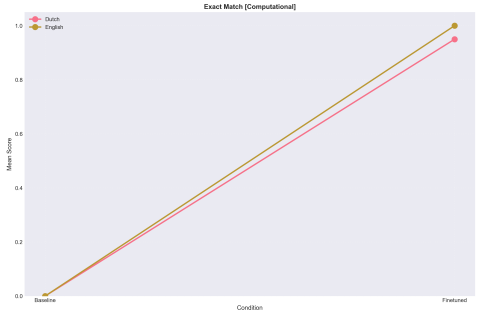
(c) Exact Match Interaction of Forget Set



(d) BLEU Interaction of Retain Set

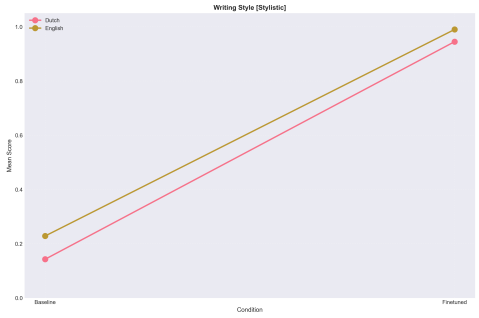


(e) ROUGE-L Interaction of Retain Set

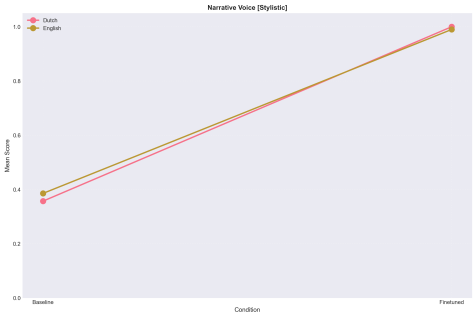


(f) Exact Match Interaction of Retain Set

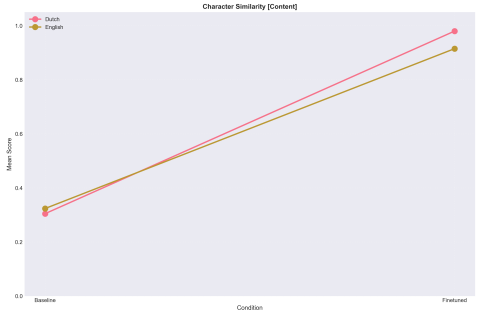
Fig. F2 Computational similarity metrics (Exact Match BLEU and ROUGE) for baseline and fine-tuned models where fine-tuning almost saturates all three metrics in both languages and splits indicating near-perfect literal reproduction of training passages



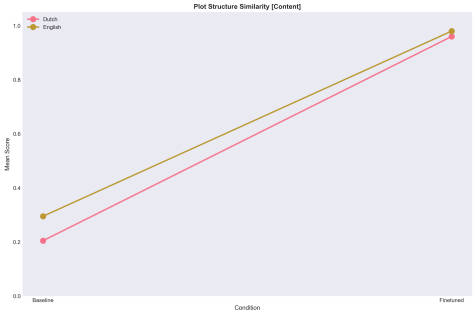
(a) Writing Style Interaction on Retain Set



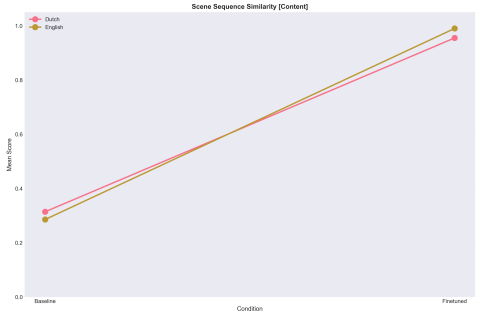
(b) Narrative Voice Interaction on Retain Set



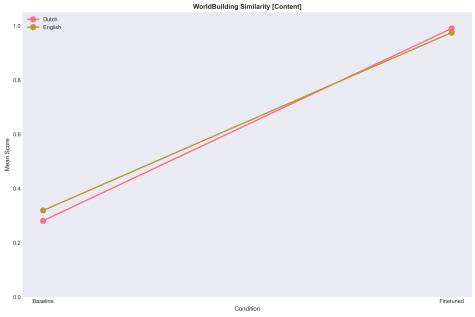
(c) Character Similarity Interaction on Retain Set



(d) Plot Structure Similarity Interaction on Retain Set



(e) Scene Sequence Similarity Interaction on Retain Set



(f) World Building Similarity Interaction on Retain Set

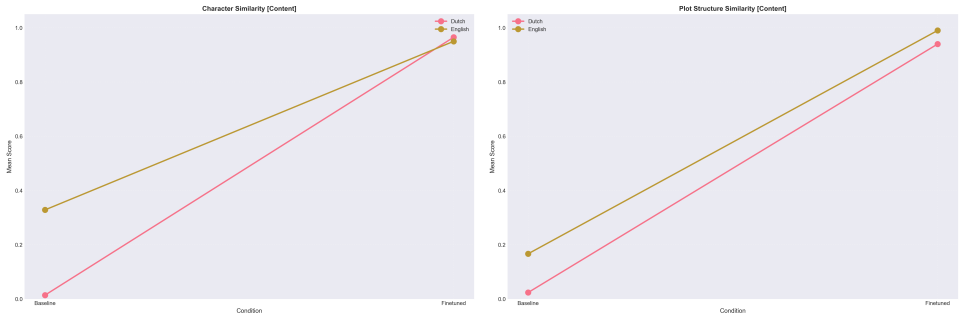
Fig. F3 PSALM evaluator results for baseline and fine-tuned models where fine-tuning raises stylistic and content similarity from low to moderate levels to values close to one across both languages and splits indicating strong imprinting of the authors' style and narratorial choices

F.4 Comparing Metric Categories Extended

A category-level pairwise heatmap can be seen in Figure F7.

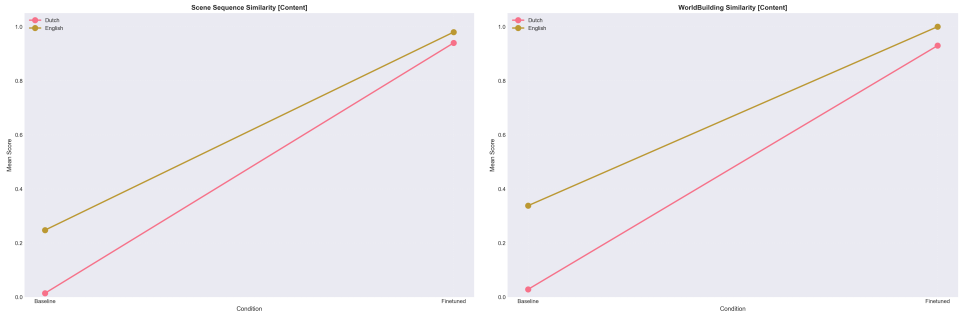
F.5 Assumption Checks and Test Selection

Shapiro–Wilk tests reported in Online Resources 4 and 5 show that none of the thirteen metrics followed a normal distribution in any model–split combination. This is expected given the strong ceiling and floor effects: baseline models often concentrate scores at or near 0, whereas fine-tuned models concentrate scores at or near 1 for infringement-related metrics. Levene's tests indicate that variance homogeneity also fails for most metrics, especially on the retain split, where all but four metrics exhibit significant heteroscedasticity. Consequently, all main analyses use non-parametric procedures: Kruskal–Wallis tests for omnibus comparisons across the four model conditions, Mann–Whitney U-tests for independent pairwise contrasts,



(a) Character Similarity Interaction on Forget Set

(b) Plot Structure Similarity Interaction on Forget Set



(c) Scene Sequence Similarity Interaction on Forget Set

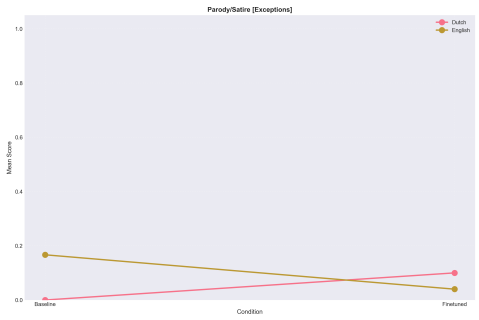
(d) World Building Similarity Interaction on Forget Set

Fig. F4 PSALM evaluator results for baseline and fine-tuned models where fine-tuning raises stylistic and content similarity from low to moderate levels to values close to one across both languages and splits indicating strong imprinting of the authors' style and narratorial choices

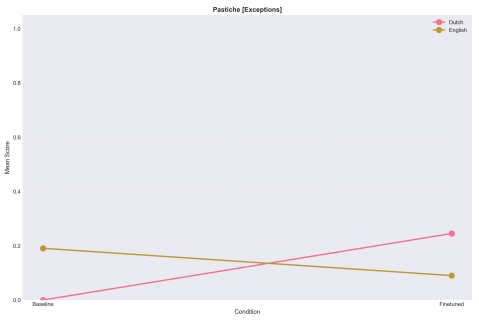
and Wilcoxon signed-rank tests for paired baseline–fine-tuned comparisons on shared prompts. Rank-based ART-ANOVA is used for language × training-stage interactions.

F.6 Omnibus Tests and Effect Size Patterns

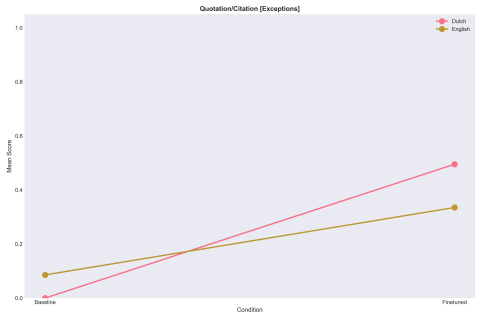
Kruskal–Wallis statistics in Online Resource 6 confirm that training stage has a very strong effect on all infringement-oriented metrics. For all computational, stylistic and content metrics on both forget and retain splits, the omnibus p-values are smaller than 10^{-11}. Among the exception metrics, Quotation/Citation also shows highly significant omnibus differences on both splits (again with $p \ll 10^{-10}$), Parody/Satire and Pastiche are significant only on the forget split, and Scènes à Faire never reaches significance (forget: $p \approx 0.97$, retain: $p \approx 0.53$). This pattern agrees with the qualitative picture in the main text: fine-tuning produces large and systematic changes for the infringement dimensions and for Quotation/Citation, while other exceptions are only weakly affected and Scènes à Faire is essentially unchanged.



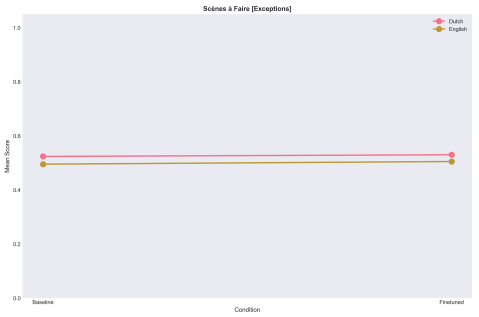
(a) Parody/Satire Interaction on forget set



(b) Pastiche Interaction on forget set



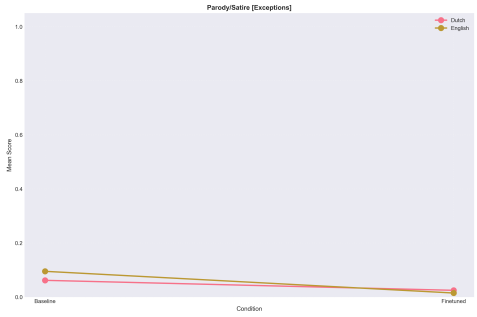
(c) Quotation/Citation Interaction on forget set



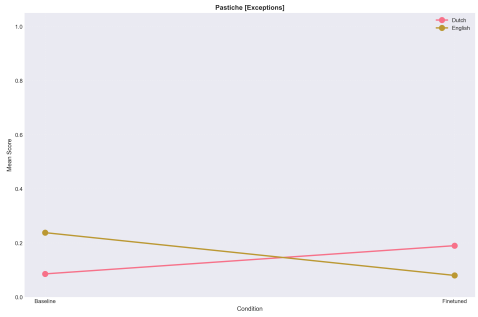
(d) Scènes à Faire Interaction on forget set

Fig. F5 Exception metrics (Parody/Satire Pastiche Quotation/Citation and Scènes à Faire) for baseline and fine-tuned models on the forget set where fine-tuning modestly increases Quotation/Citation scores but leaves Parody/Satire Pastiche and Scènes à Faire largely unchanged

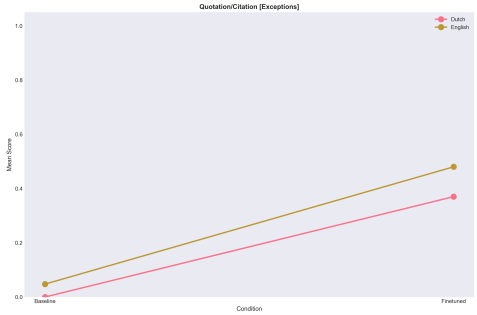
The effect size CSVs (Online Resources 7 and 8) show that almost all baseline–fine-tuned contrasts on computational, stylistic and content metrics are associated with very large Cohen's $|d|$ and Cliff's $|\delta|$. It is common for $|d|$ to exceed 2 for these metrics, and several computational comparisons reach values above 10 (for example BLEU and ROUGE on the retain split, where baseline means are below 0.1 and fine-tuned means are exactly 1). Cliff's δ for these comparisons is typically very close to -1, reflecting almost complete separation between baseline and fine-tuned distributions. By contrast, most exception metrics have substantially smaller effect sizes. For Parody/Satire and Pastiche, $|d|$ values usually lie below 1, with positive and negative signs depending on the language and split. Quotation/Citation is the only exception metric that exhibits large effects, especially when moving from baseline to fine-tuned models on the retain split, where Cohen's d exceeds 2 in several comparisons. Scènes à Faire effect sizes are close to zero in both splits, which matches the non-significant omnibus tests.



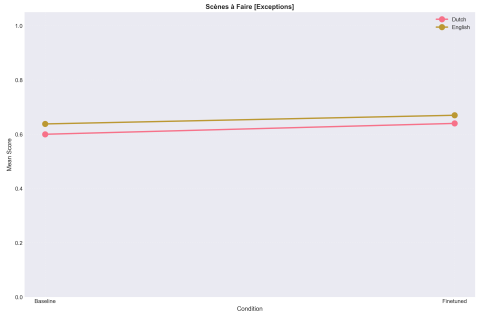
(a) Parody/Satire Interaction on retain set



(b) Pastiche Interaction on retain set



(c) Quotation/Citation Interaction on retain set



(d) Scènes à Faire Interaction on retain set

Fig. F6 Exception metrics (Parody/Satire Pastiche Quotation/Citation and Scènes à Faire) for baseline and fine-tuned models on the retain set where fine-tuning modestly increases Quotation/Citation scores but leaves Parody/Satire Pastiche and Scènes à Faire largely unchanged

F.7 Category-Level Summaries and Ranks

The per-category statistics in Online Resource 9 provide a compact view of these patterns. On the forget split, the Dutch baseline has mean scores of approximately 0.04 (Computational), 0.08 (Stylistic), 0.04 (Content), and 0.13 (Exceptions), whereas the English baseline is somewhat higher on all categories, particularly Content (mean ≈ 0.21) and Exceptions (mean ≈ 0.21). After fine-tuning, both languages converge to very similar category means: around 0.97–1.0 for Computational, Stylistic and Content, and mid-range values for Exceptions (roughly 0.25–0.37 depending on language and split). The retain split shows the same pattern, with baselines starting slightly higher than on the forget split for Content and Stylistic categories, and fine-tuned models again close to 1 on all infringement categories.

These differences are also reflected in the rank-based summary in Online Resource 10. When all thirteen metrics are ranked jointly per split, the two fine-tuned models always achieve the best average ranks, around 1.6–1.8, while the baselines occupy the third and fourth positions, with average ranks between about 2.8 and 3.8. This holds on both forget and retain splits and confirms that, viewed across the full metric set, the fine-tuned models are consistently judged more similar to the sources than either baseline.

Category-level differential-effect analyses in Online Resource 11 further support the claim in the main text that fine-tuning acts as a broadly symmetric amplifier across infringement categories. For Dutch on the forget split, the mean reduction (baseline minus fine-tuned) is almost identical for Computational, Stylistic and Content categories (roughly -0.91 in each case), and significantly larger in magnitude than the reduction for Exceptions (around -0.21). For English, reductions for Computational, Stylistic and Content are again large and negative, and all three differ markedly from the Exceptions category. There is some evidence that computational metrics increase slightly more than content metrics in English (the Computational–Content contrast is

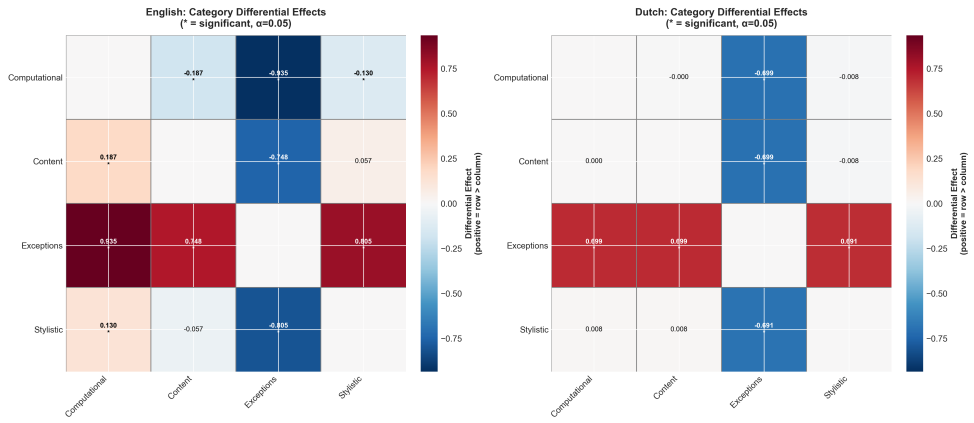


Fig. F7 Category-level pairwise contrasts for the forget split where cells show differences in reductions between categories and asterisks mark statistically significant contrasts

Table F3 Summary of baseline-to-fine-tuned changes by category (forget split). Values are mean score differences (baseline minus fine-tuned) aggregated over metrics within each category

Language	Computational	Stylistic	Content	Exceptions
English	≈ -0.95	≈ -0.82	≈ -0.76	≈ -0.01
Dutch	≈ -0.91	≈ -0.90	≈ -0.91	≈ -0.21

statistically significant), but the absolute differences in magnitude are small compared to the gap between the infringement categories and the Exceptions category.

Table F3 synthesises these differential effects. The precise values come directly from Online Resource 11. In both languages the three infringement categories show reductions close to -1, while the Exceptions category shows only small shifts, particularly in English.

F.8 Language–Stage Interactions

The rank-based ART-ANOVA results in Online Resources 12 and 13 provide a systematic view of interactions between language (English vs. Dutch) and training stage (baseline vs. fine-tuned). On the forget split, significant language main effects appear for several metrics, including BLEU, Writing Style, Narrative Voice, Character Similarity, Scene Sequence, World Building, Parody/Satire, Pastiche and Quotation/Citation. In each of these cases, inspection of the per-metric descriptive statistics shows the same qualitative pattern: the English baseline starts with higher similarity than the Dutch baseline, but the two languages converge after fine-tuning, with English and Dutch fine-tuned models attaining similar means. For BLEU, Character Similarity, Scene Sequence, World Building, Parody/Satire, Pastiche and Quotation/Citation, the ART-ANOVA also reports significant language $\times$ training-stage interactions, which again reflect this "convergence" pattern rather than a reversal of direction.

On the retain split, interactions are sparser but still present. BLEU and ROUGE exhibit significant language effects and significant interactions. Here, English baselines are higher than Dutch baselines, and both fine-tuned models reach the same ceiling (BLEU $= 1$, ROUGE $= 1$), so the interaction is driven by larger English-to-English changes than Dutch-to-Dutch changes. Pastiche also shows a significant interaction on the retain set; the underlying means indicate that English models experience a small decrease in Pastiche scores after fine-tuning, while Dutch models show a small increase, but the absolute magnitudes are low in both cases and remain well below the levels associated with strong pastiche characteristics.

These interaction results do not contradict the main conclusion that fine-tuning largely erases pre-training language differences. Rather, they make that conclusion more precise: for several metrics, English models begin closer to the target works, especially on content dimensions, whereas Dutch models start further away; fine-tuning then aligns both languages to similar similarity levels. Exceptions behave differently:

Quotation/Citation shows strong increases in both languages, Parody/Satire and Pastiche change only modestly, and Scènes à Faire remains stable.

F.9 Paired Analyses on Shared Prompts

The paired Wilcoxon tests in Online Resources 14 and 15 compare baseline and fine-tuned models on exactly the same prompts within each language. For the English forget split, the baseline means are well below 0.5 on all infringement metrics and the fine-tuned means lie very close to 1. The resulting mean differences range from about 0.51 (Character Similarity) to 0.90–0.93 (computational metrics), and all p-values are well below 10^{-5}. Cohen's d values for these paired comparisons reach extremely large magnitudes, for example $d \approx 15.6$ for BLEU and $d \approx 11.1$ for ROUGE. Dutch paired comparisons on the forget split show the same pattern with similar effect sizes. On the retain split, paired differences are slightly smaller but remain large for all infringement metrics; for example, English BLEU increases by about 0.89 on average, and Dutch BLEU by about 0.97. For several exception metrics, the paired tests reveal small but significant changes (e.g. a decrease in English Parody/Satire means on the retain split), whereas others such as Scènes à Faire show no significant paired change.

The paired analyses therefore confirm, at the level of individual prompts, the aggregate picture provided by the Kruskal–Wallis and effect-size summaries: fine-tuning consistently increases similarity across computational, stylistic and content dimensions, both for forget and retain examples, while producing more modest and metric-specific changes in exception dimensions.

F.10 Distributional Characteristics and Atypical Cases

The descriptive statistics in Online Resource 6 also reveal how the shape of the score distributions changes across model conditions. For most infringement metrics, baseline medians on the forget split are exactly 0.0 for both languages, whereas fine-tuned medians are exactly 1.0. The standard deviations for fine-tuned models are small (often less than 0.05 on the retain split and slightly larger on the forget split), indicating that almost all outputs are scored at or very near the maximum similarity level once fine-tuning is applied. Baseline standard deviations are much larger and medians sometimes fall in the mid-range on the retain split, especially for English content metrics, reflecting the fact that the pre-trained model already exhibits some generic similarity to the historical fiction corpus.

Infrequent cases where fine-tuned scores fall below the ceiling can be seen, for example, in Dutch Exact Match on the forget split (mean 0.90, standard deviation ≈ 0.31), where a few prompts do not lead to exact reproductions, and in Dutch Character Similarity and World Building on the forget split, where the means are above 0.93 but standard deviations around 0.18–0.23 indicate some dispersion below 1. Manual inspection (reported in the main RQ2 discussion) attributes these deviations to translation noise, sampling variance, or prompts referring to story elements that are only weakly represented in the selected passages. These atypical cases do not alter

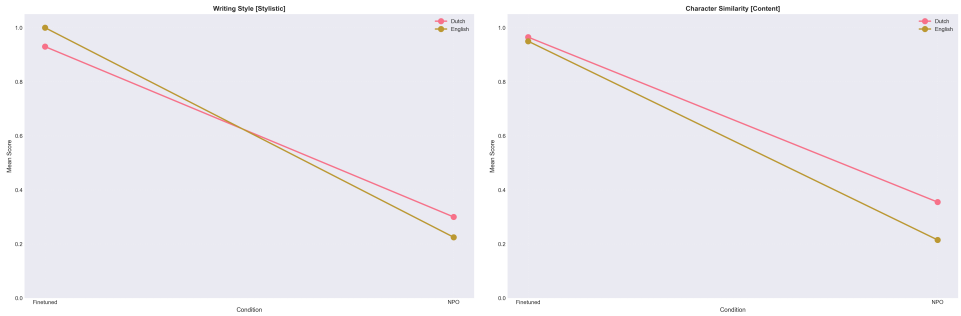
the overall pattern, but they are relevant when interpreting ceiling effects: the near-1.0 means are not artefacts of a single extreme example, but reflect a general shift of the entire distribution towards very high similarity.

Appendix G RQ3: Detailed Statistical Analyses

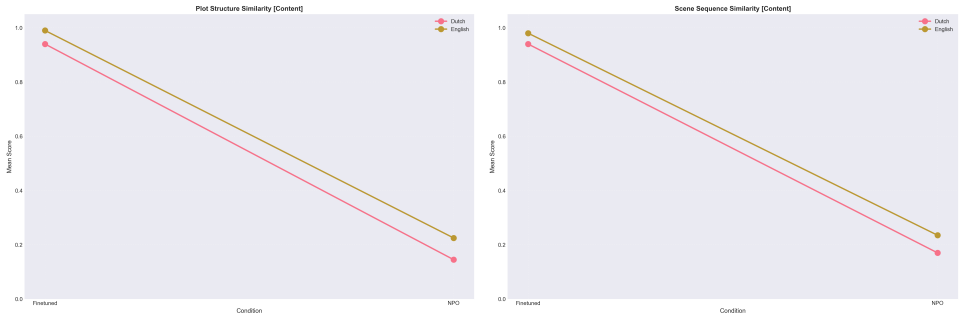
This appendix summarises the core supplementary findings for RQ3, using the supplementary data. The aim is to document the main numerical patterns that are not already discussed in subsection 5.3.

G.1 Stylistic and Structural Similarity Extended

The results of stylistic and content similarity before and after unlearning on the forget and retain sets can be seen in Figure G8 and Figure G9.

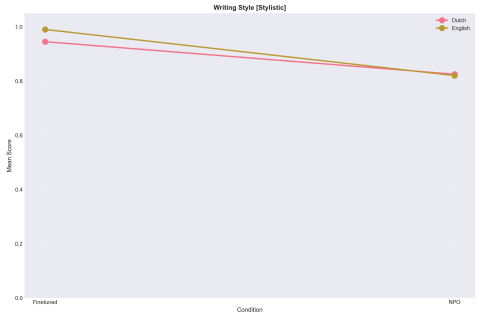


(a) Writing Style Interaction on Forget Set **(b)** Character Similarity Interaction on Forget Set

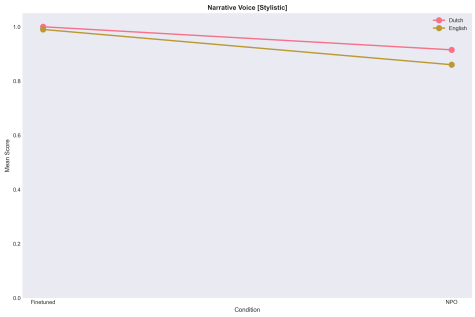


(c) Plot Structure Similarity Interaction on Forget Set **(d)** Scene Sequence Similarity Interaction on Forget Set

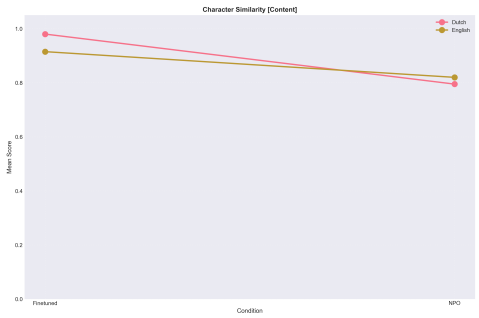
Fig. G8 PSALM evaluator results for fine-tuned and NPO models where fine-tuning raises stylistic and content similarity from low to moderate levels to values close to one across both languages and splits indicating strong imprinting of the authors' style and narratorial choices



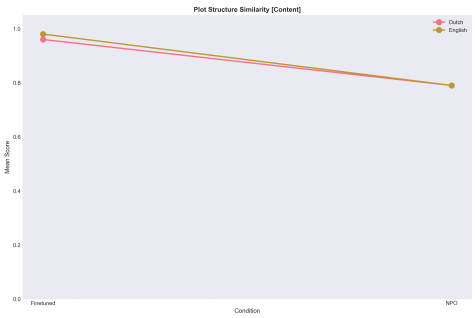
(a) Writing Style Interaction on Retain Set



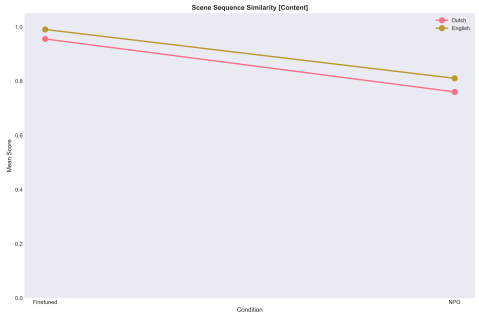
(b) Narrative Voice Interaction on Retain Set



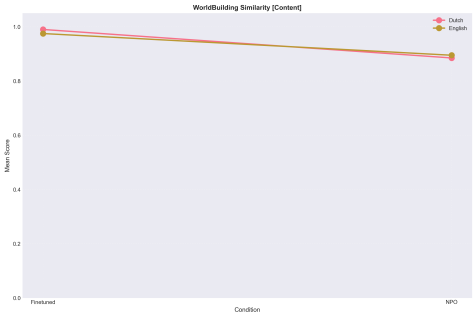
(c) Character Similarity Interaction on Retain Set



(d) Plot Structure Similarity Interaction on Retain Set



(e) Scene Sequence Similarity Interaction on Retain Set



(f) World Building Similarity Interaction on Retain Set

Fig. G9 PSALM evaluator results for fine-tuned and NPO models

G.2 Behaviour of Exceptions Extended

The results of the exception metrics before and after unlearning on the forget and retain sets can be seen in Figure G10 and Figure G11.

G.3 Suppression of Literal Memorisation

The results of the computational metrics before and after unlearning on the forget and retain sets can be seen in Figure G12.

G.4 Comparing Metric Categories Extended

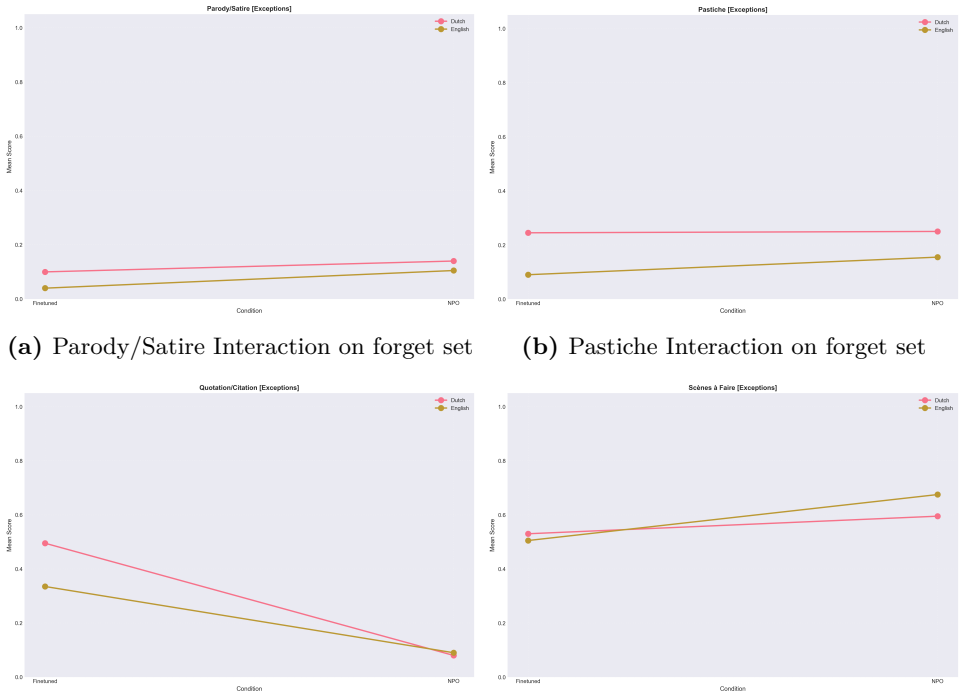
The category-level pairwise heatmap can be seen in Figure G13.

G.5 Residual Similarity and Distance to Baseline

The score distribution after unlearning can be found in Figure G14.

G.6 Diagnostic tests and omnibus effects

Normality and homogeneity checks (Shapiro–Wilk and Levene; Online Resources 16 and 17) indicate that none of the metrics are safely modelled with standard Gaussian assumptions. For both forget and retain splits, all metrics fail normality tests, and



(a) Parody/Satire Interaction on forget set **(b)** Pastiche Interaction on forget set

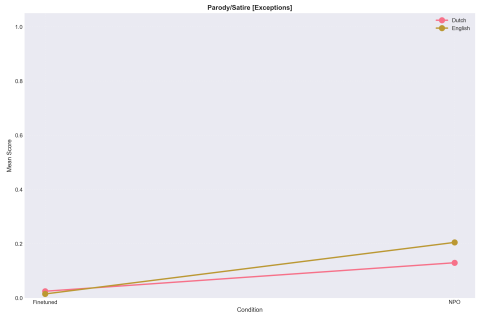
(c) Quotation/Citation Interaction on forget set **(d)** Scènes à Faire Interaction on forget set

Fig. G10 Exception metrics (Parody/Satire Pastiche Quotation/Citation and Scènes à Faire) for fine-tuned and NPO models on the forget set

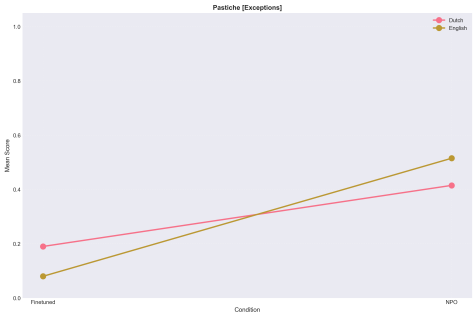
most also violate homoscedasticity. Consequently, the main analyses rely on rank-based methods throughout.

Kruskal–Wallis omnibus tests for each metric and split (Online Resource 18) show highly significant effects of model condition for all computational, stylistic and content metrics on both splits (all $p \leq 5 \times 10^{-3}$). Among the exception metrics, Parody/Satire and Pastiche show non-significant or only weakly significant omnibus results on the forget split, and Quotation/Citation and Scènes à Faire show non-significant omnibus effects on the retain split. This confirms that unlearning substantially affects the infringement-oriented metrics, while its impact on exception-related dimensions is smaller and more heterogeneous.

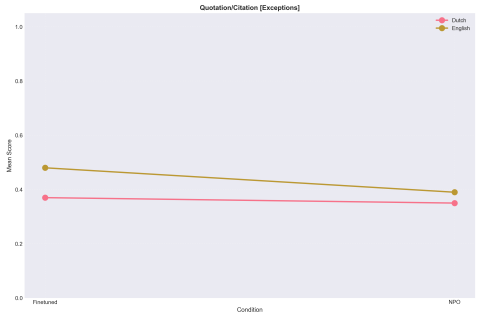
Rank-based ART-ANOVA models (Online Resources 19 and 20) provide consistent evidence that unlearning has a strong main effect on all metrics. For the forget split, the unlearning factor attains F-statistics well above 60 for most infringement-oriented metrics, with p-values below 10^{-11}. On the retain split, unlearning remains significant for every metric except Quotation/Citation and Scènes à Faire. Language main effects and language×unlearning interactions are generally absent; world-building similarity



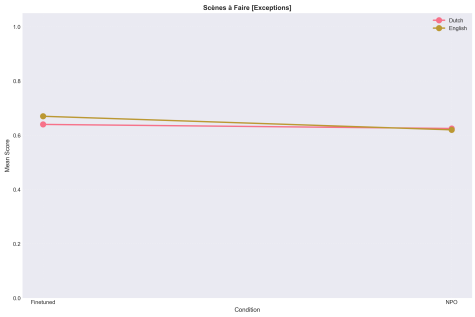
(a) Parody/Satire Interaction on retain set



(b) Pastiche Interaction on retain set



(c) Quotation/Citation Interaction on retain set



(d) Scènes à Faire Interaction on retain set

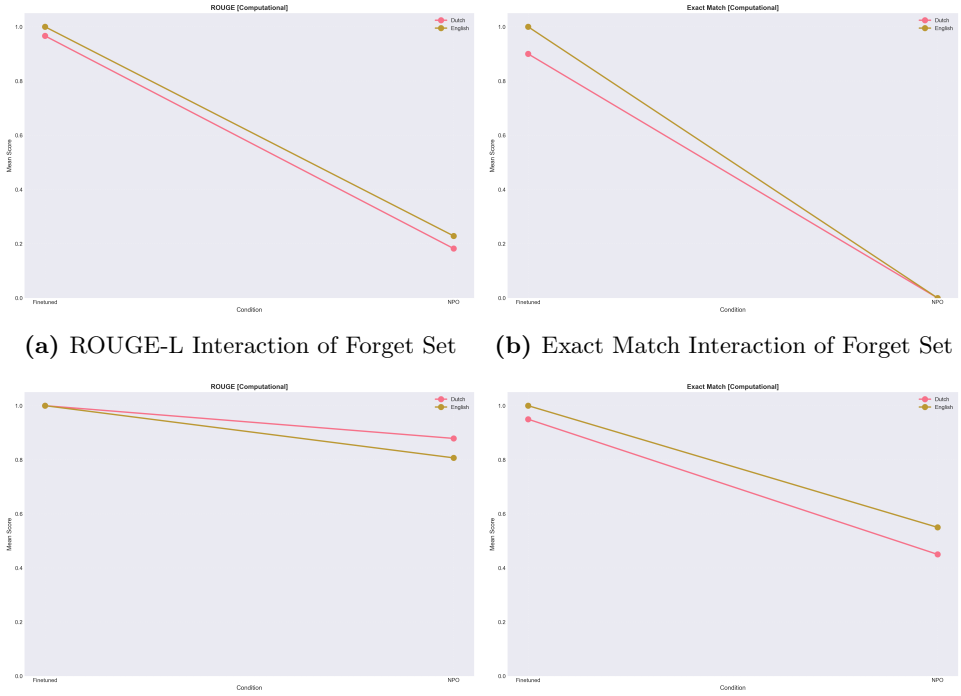
Fig. G11 Exception metrics (Parody/Satire Pastiche Quotation/Citation and Scènes à Faire) for fine-tuned and NPO models on the retain set

on the forget split is the only metric where an interaction reaches significance, matching the qualitative discussion in subsection 5.3.3. No interaction is detected on the retain split.

G.7 Overall and category-level patterns

Average ranks across all thirteen metrics (Online Resource 21) confirm that the fine-tuned models remain closest to the source passages, and that NPO consistently moves models away from the training works. On the forget split, the English and Dutch fine-tuned models have average ranks of approximately 1.85 and 2.00, respectively, whereas the corresponding NPO models have average ranks of about 3.04 and 3.12. On the retain split, the fine-tuned models again occupy the top positions (around 1.77–1.92), with both NPO variants clustered near rank 3.0 or higher. The rank differences between English and Dutch within the same training regime are modest compared to the differences between fine-tuned and unlearned models.

Category-level means (Online Resource 22) give additional context to the figures reported in the main text. On the forget split, fine-tuned models in both languages



(a) ROUGE-L Interaction of Forget Set **(b)** Exact Match Interaction of Forget Set

(c) ROUGE-L Interaction of Retain Set **(d)** Exact Match Interaction of Retain Set

Fig. G12 Computational similarity metrics (Exact Match BLEU and ROUGE) for fine-tuned and NPO models where NPO almost eliminates verbatim reproduction on forget sets while leaving substantial overlap on retain sets

show category means above 0.94 for computational, stylistic and content metrics, while exception means are substantially lower (between roughly 0.25 and 0.37). After NPO, forget-set means fall to about 0.11–0.13 for computational metrics and to the 0.23–0.31 range for stylistic and content metrics, whereas exception means remain in

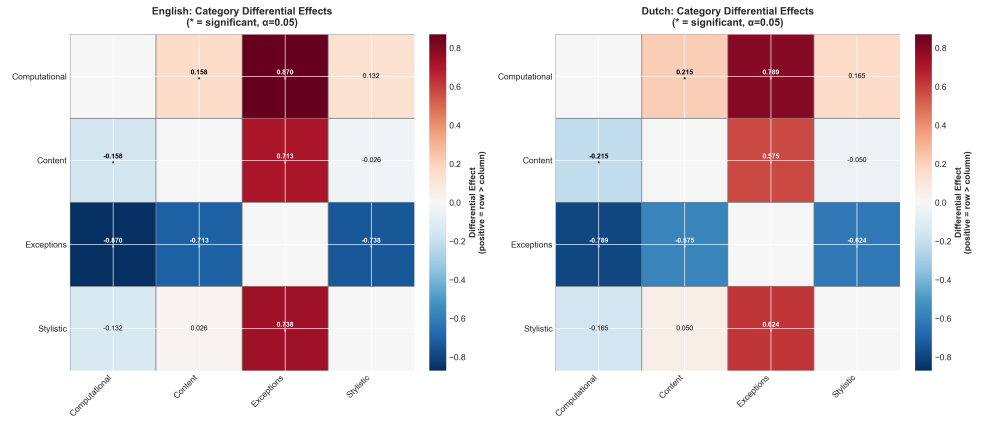


Fig. G13 Category-level pairwise contrasts for the forget split where cells show differences in reductions between categories and asterisks mark statistically significant contrasts

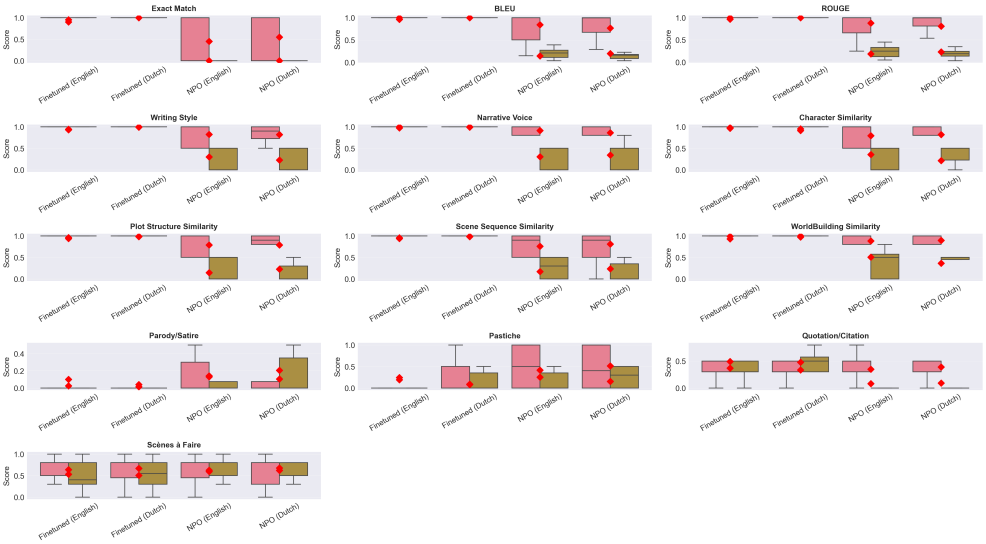


Fig. G14 Score distributions for all PSALM metrics across fine-tuned and NPO models, separated by forget and retain splits. Fine-tuned scores concentrate near the maximum; NPO spreads the distributions and lowers their medians, but many retain-set scores remain high

the 0.24–0.28 range. On the retain split, the fine-tuned models again have category means close to one for computational, stylistic and content dimensions, while exceptions sit around 0.32. The unlearned models keep retain-set means high for stylistic and content categories (roughly 0.83–0.86 for English and 0.79–0.79 for Dutch), whereas computational retain means drop to approximately 0.69–0.71. Exception means on the retain split increase slightly for both languages after NPO (to around 0.39–0.43), but remain well below the infringement-oriented categories.

Differential-effect contrasts between categories on the forget split (Online Resource 23) show that, in both languages, reductions are large and similar for computational, stylistic and content categories, whereas exception metrics change substantially less. For English, the mean reduction in the computational category exceeds that in the content category by about 0.16 (paired $t \approx 3.29$, $p \approx 0.017$), with no meaningful difference between stylistic and content reductions. For Dutch, the computational category is reduced more than content by approximately 0.21, with a highly significant contrast ($t \approx 7.15$, $p < 10^{-4}$), while stylistic and content reductions are statistically indistinguishable. In both languages, all three infringement-oriented categories are reduced markedly more than exceptions.

G.8 Pairwise comparisons and effect sizes

Wilcoxon signed-rank tests comparing fine-tuned and NPO models on the same prompts (Online Resources 24 and 25) refine the picture at the metric level. On the forget split, NPO produces statistically significant changes for every computational, stylistic and content metric in both languages (all $p < 10^{-6}$), with paired Cohen's d values well above 1.5. Among the exception metrics, Quotation/Citation and Scènes à Faire show significant changes on the forget split, whereas Parody/Satire and Pastiche do not. On the retain split, English models exhibit significant changes for all computational, stylistic and content metrics except Character Similarity, and for Parody/Satire and Pastiche; however, Quotation/Citation and Scènes à Faire on the English retain split show no statistically significant change. For Dutch retain data, only the three computational metrics differ significantly between fine-tuned and NPO models; all stylistic, content and exception metrics have Wilcoxon p-values above 0.09, indicating that the Dutch unlearning configuration leaves most retain-set behaviour statistically similar to the fine-tuned state.

Effect-size summaries from independent-sample comparisons (Online Resources 26 and 27) underline the magnitude of unlearning on the forget sets. For computational metrics, Cohen's d comparing fine-tuned and NPO models on the forget split ranges from about 4.1 (Exact Match, Dutch) to over 11 (BLEU, English), with Cliff's δ at or near ± 1, indicating almost complete separation of the distributions. Stylistic and content metrics on the forget split also show very large effects, with Cohen's d typically between 2.0 and 5.5. Exception metrics exhibit smaller and more variable effects: Quotation/Citation differences on the forget split reach d values between roughly 1.35 and 2.07, while Scènes à Faire and Pastiche are associated with d values around 0.3–0.7, and Parody/Satire effects are close to zero or modest in magnitude.

On the retain split, effect sizes are more moderate. For English models, Cohen's d comparing fine-tuned and NPO models is around 1.2 for Exact Match and approximates 1.1 for BLEU and ROUGE, and sits near 1.0 for Writing Style and Scene Sequence Similarity. For Dutch retain data, effect sizes for computational metrics are again large (around 1.3 for Exact Match and approximately 0.9–1.0 for BLEU and ROUGE), whereas stylistic and content metrics generally show medium or small effects, and many exception metrics produce negligible effect sizes. These patterns align with the Wilcoxon results: computational metrics are altered substantially on both splits, while many non-computational metrics on the retain split change only modestly or not at all, especially for Dutch.

G.9 Equivalence to baseline models

Two one-sided tests for equivalence (TOST) between baseline and NPO models on the retain split (Online Resource 28) assess whether unlearning returns the models to a state practically indistinguishable from the original instruction-tuned models. Equivalence bounds are set at ± 0.1 on the normalised 0–1 scale, corresponding to roughly half a PSALM category. For all computational, stylistic and content metrics in both languages, the combined TOST p-values exceed 0.05, and equivalence is therefore rejected. For example, the English Plot Structure Similarity mean drops from 0.98 in the baseline to 0.79 in the unlearned model, with a TOST p-value of about 0.94; similarly, the Dutch World-Building Similarity mean moves from 0.99 to 0.89, with a TOST p around 0.54. Even in cases where numerical differences are smaller, such as Dutch Narrative Voice on the retain split (1.0 vs. 0.979), the tests do not establish equivalence under the chosen bounds.

For exception metrics, the picture is mixed. Some comparisons, such as English Quotation/Citation and both languages' Scènes à Faire on the retain split, also fail the equivalence tests, despite relatively modest mean differences. Others, particularly where baseline and unlearned means are very close, produce TOST p-values that are not directly informative because of the small sample sizes and high variance. Overall, across all metrics evaluated, none of the tested retain-set comparisons between baseline and NPO models meet the predefined equivalence criterion, supporting the conclusion in subsection 5.3.5 that unlearning does not restore models to their pre-fine-tuning similarity levels.