Estranged Predictions: Measuring Semantic Category Disruption with Masked Language Modelling

Yuxuan Liu[1], Haim Dubossarsky[2], and Ruth Ahnert[1]

[1] School of Arts, Queen Mary University of London, London, United Kingdom
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

Abstract

This paper examines how science fiction destabilises ontological categories by measuring conceptual permeability across the terms *human*, *animal*, and *machine* using masked language modelling (MLM). Drawing on corpora of science fiction (Gollancz SF Masterworks) and general fiction (NovelTM), we operationalise Darko Suvin's theory of estrangement as computationally measurable deviation in token prediction, using RoBERTa to generate lexical substitutes for masked referents and classifying them via Gemini. We quantify conceptual slippage through three metrics: retention rate, replacement rate, and entropy, mapping the stability or disruption of category boundaries across genres. Our findings reveal that science fiction exhibits heightened conceptual permeability, particularly around *machine* referents, which show significant cross-category substitution and dispersion. *Human* terms, by contrast, maintain semantic coherence and often anchor substitutional hierarchies. These patterns suggest a genre-specific restructuring within anthropocentric logics. We argue that estrangement in science fiction operates as a controlled perturbation of semantic norms, detectable through probabilistic modelling, and that MLMs, when used critically, serve as interpretive instruments capable of surfacing genre-conditioned ontological assumptions. This study contributes to the methodological repertoire of computational literary studies and offers new insights into the linguistic infrastructure of science fiction.

Keywords: masked language model, science fiction, distant reading, conceptual permeability

1 Introduction

Science fiction has long served as a speculative mirror for our assumptions about identity and existence, and as a crucible in which such categories are actively contested, dismantled, and reconstituted. The boundary between *human* and *Other* emerges not as a stable demarcation but as a zone of negotiation, hybridisation, and semantic leakage. In this space, estrangement functions not only as a narrative device but as a cognitive operation, one that defamiliarises hegemonic epistemologies and opens possibilities for alternative recognition and disidentification [1].

This ontological disturbance finds its linguistic correlate in the destabilisation of language itself, where familiar referents are defamiliarised and perception is slowed into reflective, non-automatic engagement, a process theorised in Viktor Shklovsky's concept of defamiliarisation [2]. Estrangement thus operates not only at the level of macro-narrative architecture, but at the micro-level of syntax and lexical substitution.

Yuxuan Liu, Haim Dubossarsky, and Ruth Ahnert. "Estranged Predictions: Measuring Semantic Category Disruption with Masked Language Modelling." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 637–653. https://doi.org/10.63744/GUArdaU1Y92u.

While substantial critical work has illuminated the speculative imagination of the *posthuman* [3], the *cyborg* [4], and the *animal* [5], these contributions have tended to privilege philosophical, ethical, or narratological perspectives. Less attention has been paid to the linguistic structure through which ontological categories are produced, troubled, or undone. A solely thematic approach risks obscuring the microstructural mechanisms, such as syntactic juxtapositions, semantic proximities, and predictive patterns of substitution, by which science fiction reconfigures the conceptual intelligibility of alterity, for such readings remain predominantly qualitative in scope and ill-equipped to chart statistical patterns across large corpora.

This study therefore poses three interrelated questions: How is the conceptual boundary surrounding the *human* rendered porous or unstable in literary language? With which kinds of entities, including *animal*, *machine*, or others, is this permeability most frequently negotiated? And how do these patterns of semantic proximity vary between speculative and non-speculative fiction? These questions are critical for rethinking how science fiction not only narrates difference, but actively reorganises the linguistic scaffolding through which difference is rendered legible.

2 Background and Related Work

Distributional semantics offers a powerful framework for interrogating such phenomena. As J. R. Firth famously proposed, "you shall know a word by the company it keeps" [6], a foundational principle further developed by Zellig Harris [7] and formalised in vector-based models of semantics [8; 9]. From this perspective, semantic categories such as *human*, *animal*, and *machine* are not defined by essential features but emerge through mutable contextual associations that vary by genre, historical moment, and discursive environment [10; 11].

Recent advances in language modelling, particularly through contextualised models such as BERT [12] and RoBERTa [13], have expanded the empirical tools available for tracking these dynamics. Unlike traditional distributional methods, either static embeddings or those that rely on raw co-occurrence frequencies, contextualised models can predict masked or "missing" tokens based on their sentence context in a paradigm called masked language modelling (MLM). This experimental paradigm allows for the reconstruction of latent semantic expectations and substitution probabilities in sentence-level contexts. When applied to science fiction, MLMs reveal how referential expectations, for example, around *human*, *machine*, and *animal* shift across discursive regimes, and how meaning becomes unstable in moments of predictive uncertainty.

The *Living with Machines* (LwM) project has demonstrated the potential of MLMs to surface latent tensions in language use, particularly by repurposing a characteristic property of large-scale language models, namely their tendency to default to high-probability, statistically dominant completions, as an analytical lens through which to detect linguistic departures from normativity. In this context, the project explored how language models could be used to detect instances of linguistic usage that would appear "surprising" to a model trained on a specific historical corpus, particularly in relation to depictions of animate machines [14; 15].

3 Approach

This study builds on the approach employed in the LwM project. Drawing from a corpus of nineteenth-century texts, the LwM team selected sentences concerning *machines* and masked the *machine*-related terms. They then prompted a historical language model that they had fine-tuned [16], to generate probable lexical substitutions for the masked words, as in their first example:

> **Original sentence**: And why should one say that the machine does not live?
>
> **Masked sentence**: And why should one say that the `[MASK]` does not live?

Predictions with scores: *man* (5.0788), *person* (4.4484), *other* (4.1866), *child* (4.1600), *king* (4.1510), *patient* (4.1249), *one* (4.1141), *stranger* (4.1067), ...

The method provided a powerful way of detecting atypical animacy, as well as revealing deep-seated cultural and ideological biases embedded within the nineteenth-century training corpus. [15].

Building on this precedent, our study extends the application of MLM to a comparative corpus comprising science fiction and general fiction. It focuses on the conceptual permeability of three ontological categories, namely *human*, *animal*, and *machine*, by masking these terms and examining the substitutions proposed by a contemporary language model. By analysing what a probabilistic model deems plausible within given linguistic contexts, this study captures not only the stability of referential expectations but also their points of breakdown. These misalignments gesture toward areas of semantic permeability or categorical ambiguity. Taken collectively, these "errors" reveal broader patterns of generic difference both within science fiction, and between the genre and the broader fiction landscape, thereby exposing genre-specific estrangement effects that operate beneath the level of explicit narration.

Through this comparative framework, our study operationalises the notion of estrangement through measurable shifts in prediction probabilities generated by a masked language model. Specifically, while it may seem self-evident that science fiction, by virtue of its genre-specific discourse, engenders increased conceptual permeability across three core ontological categories of *human*, *animal*, and *machine*, we ask whether this assumption can be substantiated at scale, and, if so, what our analytical pipeline can reveal about the mechanisms by which these ontological boundaries are explored at the microstructural level of language.

4 Materials and methods

4.1 Corpus Selection

To enable a controlled comparison of conceptual permeability across science fiction and general fiction, two corpora were selected for their viability as bounded discursive environments within which the linguistic negotiation of categorical boundaries could be meaningfully modelled. The science fiction corpus comprises 336 published works drawn from the Gollancz SF Masterworks series (1818–2019), encompassing both standalone novels and individually extracted stories from anthologies, with over 90% of the material concentrated between 1910 and 2000. The general fiction corpus is derived from the HathiTrust NovelTM dataset [17], from which we randomly sampled 700 Anglophone works published between 1910 and 2000.

4.2 Contextualised Model

We employed *roberta-base* [13] to generate predictions for masked tokens across all sentences. RoBERTa (A Robustly Optimised BERT Pretraining Approach) is a refined implementation of BERT [12], a bidirectional transformer designed to learn the probabilistic relationships between words by predicting masked tokens in context; RoBERTa is conceptualised in this study as an instrument for detecting estrangement through which to detect, replicate, and reflect the distributional conventionalities of linguistic production as sedimented through pragmatic histories and examine the structural inertia of language: its anthropocentric defaults, its resistance to nonhuman agency, and its syntactic regulation of who or what may occupy grammatically legitimate positions of action.

4.3 Experimental Procedure

Our experimental procedure consists of three core stages, illustrated in Figure 1.[1]

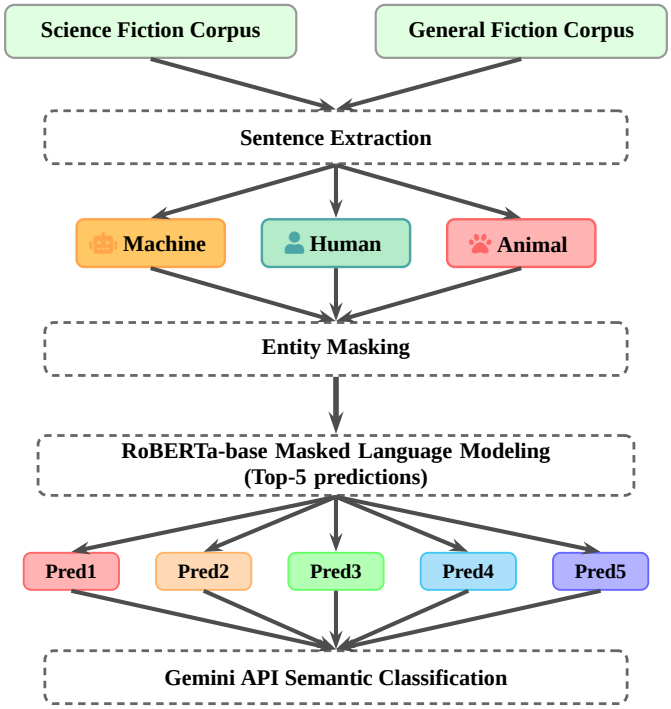

Figure 1: Cross-genre conceptual permeability detection model

4.3.1 Sentence Extraction and Masked Language Modelling

A case-insensitive lexical query extracts all sentences containing the target terms *human/humans*, *human being/human beings*, *animal/animals*, and *machine/machines* from both corpora. This procedure yields 11,709 sentences from the science fiction corpus and 8972 sentences from general fiction corpus. In each sentence, the target lexical item is replaced with a `[MASK]` token to prepare the input for MLM. Some of these sentences contain more than one target category; for instance, a sentence might mention both a *human* and a *machine*, or even include all three. In such cases, the sentence is processed iteratively: in each pass, only one term is masked, while all others remain intact. Thus, if a sentence contains two target terms, the model is run twice on the same sentence: once with the first term masked and the second visible, and once with the reverse configuration. As a result, the total number of masked sentences exceeds the number of original sentences. During prediction extraction, the model returns its top 5 candidate predictions, for each sentence, along with associated token-level probabilities.

4.3.2 Semantic Classification via Gemini

The model's predictions are subjected to semantic classification using Google's Gemini [19]. Each predicted token is classified within its full sentential context via a prompt comprising the masked sentence and the model's prediction. The classification framework employs a taxonomy centred on three core ontological categories: *Human*, *Animal*, and *Machine*. Each core category is supplemented by adjacent figures of *otherness*, including *Other-Human*, *Other-Animal*, and *Other-Machine*, as well as further categories such as *Other-Hybrid* and *Other-Ambiguous*, the latter of

[1] Development used Jupyter Notebooks launched via the OnDemand environment [18], with classification via the Gemini API. All code for the pipeline is available at `https://github.com/yuxliuu89/semantic-category-disruption-mlm`.

which accounts for liminal or indeterminate referents. Additional categories are generated ad hoc by the Gemini model in response to the semantic type of each MLM prediction and its contextual specificity. All predictions are processed in batches of 200, with quality control mechanisms in place to ensure a classification coverage rate exceeding 98%. To mitigate category fragmentation and overlap, Gemini's post-classification fusion is employed to consolidate semantically adjacent labels both within and across ontological domains.

4.4 Metrics of Analysis

The interpretive framework is structured around three related metrics:

4.4.1 Retention Rate

This metric assesses the extent to which the model's predicted substitutes for a masked word remain within the same semantic category as the original term. For example, if the masked word *human* is replaced by terms such as *person*, *man*, or *child*, the substitution is considered category-preserving. Retention rate thus serves as an index of category stability, reflecting how consistently the model maintains the semantic identity of key terms during prediction.

4.4.2 Replacement Rate

This metric quantifies how often the model's predictions transgress the semantic boundaries delineating the categories of *human*, *animal*, and *machine*. Specifically, the replacement rate indexes instances in which a masked lexical item originally classified within one category is predicted to belong to another, such as when a term denoting a *machine* elicits top-ranked predictions associated with *human*, *animal*, or *deity*. These cross-category substitutions are treated as markers of conceptual permeability. A higher replacement rate thus signals greater semantic fluidity and ontological instability, whereas lower rates suggest the reinforcement of categorical boundaries. Moreover, by attending to the directional asymmetries in replacement patterns (e.g., *machine* $\rightarrow$ *human* versus *human* $\rightarrow$ *machine*), this metric allows for an analysis of how specific genres modulate the movement of meaning across categories of being, and whether, and in which direction, they foster slippage, containment, or reification in the linguistic construction of subjecthood.

4.4.3 Entropy

Entropy is used here to quantify the degree of uncertainty in the model's predictions. As a measure of probability dispersion, entropy reflects how the model navigates semantic constraints embedded within different literary contexts. Low entropy values correspond to highly concentrated distributions, where the model assigns disproportionately high probability to a single lexical candidate, suggesting strong contextual anchoring and reduced semantic ambiguity. Conversely, high entropy signals a more evenly distributed probability mass across multiple candidates, indicating a looser context that invites several plausible predictions. While the previous metrics evaluated the model's predictions by assessing whether each masked token was retained within or replaced across its original category, entropy was used to capture the probabilistic dispersion across the top five predictions collectively, thereby reflecting how uncertainty is distributed within the model's semantic space.

4.4.4 Aggregation Schemes

In quantifying the retention of a masked token within its original semantic category and its replacement into alternative categories, two ways of aggregating RoBERTa's top-5 predictions for each masked position were considered. Method 1 treats the five highest-probability predictions for each masked position as equally weighted candidates, interpreting their presence as an indication of semantic plausibility. For example, a prediction for *machine* with a probability of 0.02 is counted in

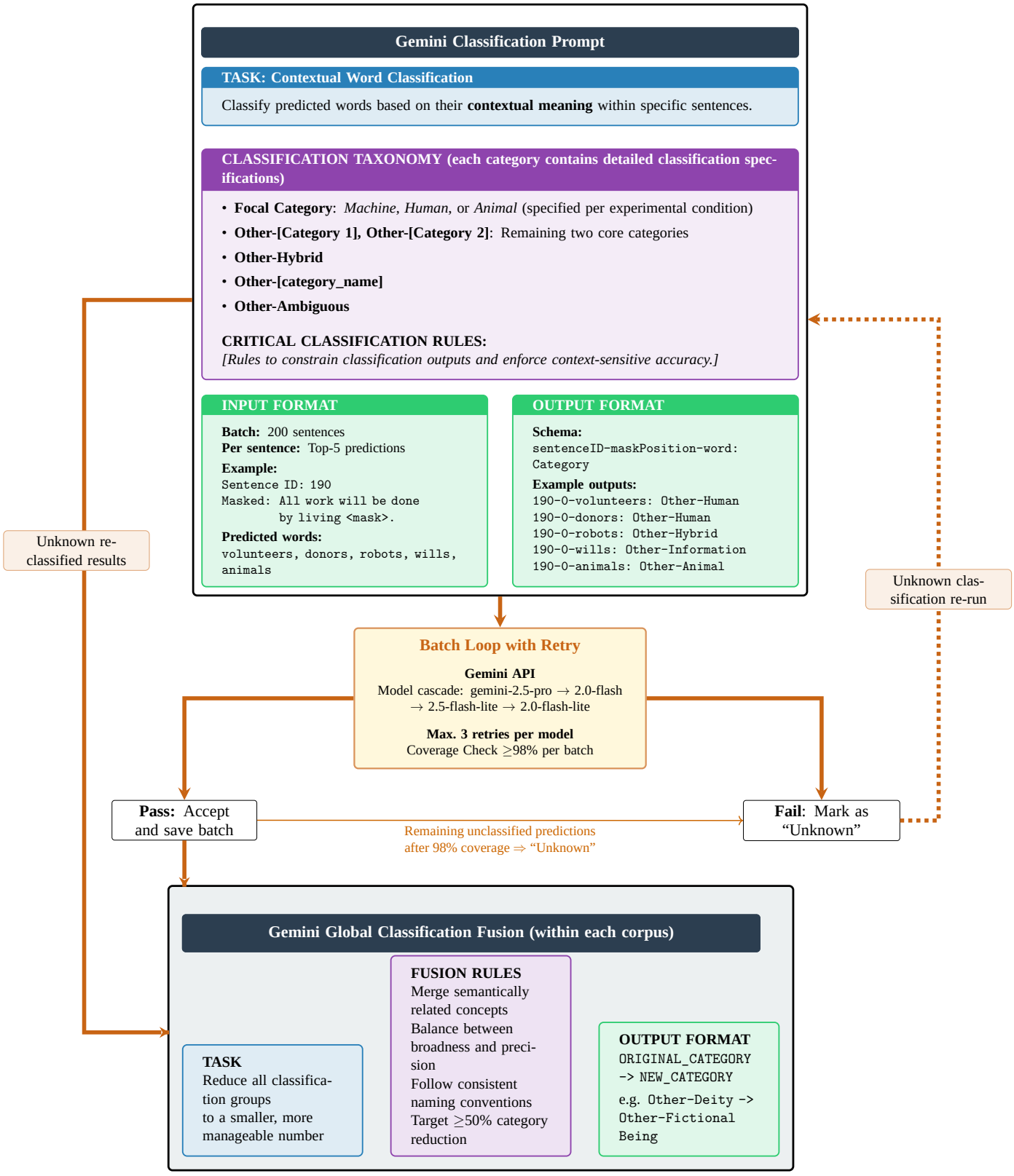

Figure 2: Three-stage Gemini-based classification workflow

the same way as one with a probability of 0.35. Under this method, retention is defined as:

$$\text{Retention} = \frac{\text{Number of source-category predictions}}{\text{Total number of predictions}},$$

with analogous calculations applied to each replacement category.

By contrast, Method 2 weights candidates according to their probabilities, summing the total probability mass assigned to each category within a sentence and then averaging across all masked positions. Retention is thus calculated as:

$$\text{Retention} = \frac{1}{n} \sum_{\substack{\text{source} \\ \text{category}}} P_{(x)},$$

where $P_{(x)}$ denotes the probability assigned by the model to a token belonging to the source category and *n* represents the number of masked positions.

However, Method 1 captures weak but non-trivial cross-category substitutions, amplifying marginal predictions that may index latent semantic permeability. By granting uniform status to all top-5 predictions irrespective of their associated probabilities, Method 1 lowers the threshold for registering both category-consistent and category-divergent outcomes, ensuring that incursions across semantic boundaries are not dismissed prematurely. Given that the analytical aim is to detect the semantic permeability of ontological categories rather than to replicate the model's internal probability structure, and in view of the high consistency observed between the retention and replacement patterns generated by Method 1 and Method 2 (full results for Method 2 are provided in the appendix; see Figure 6), the use of an equal-weight approach that retains marginal signals under these conditions remains analytically justified.

4.4.5 Statistical Tests

For retention, we binarised the model's top-5 predictions, whether they remain within the same semantic category or not, for each masked sentence, and expressed this as a proportion (0–1). To test whether retention differs between corpora, we computed the observed mean difference in the binarised per-sentence retention between Gollancz SF and NovelTM, and evaluated its significance using a 10,000-iteration permutation test that randomly shuffled corpus labels. We also obtained 95% confidence intervals for the mean difference by bootstrap resampling sentences within each corpus. This model-free approach ensures that significance does not depend on distributional assumptions and treats each sentence as an independent observation. For Entropy analysis, we used a two-way analysis of variance (ANOVA) test with category and corpus as the main variables and their interaction.

5 Results

5.1 Conceptual Retention and Semantic Dispersion Across Genres

The computational findings presented in this subsection delineate a spectrum of conceptual stability and semantic dispersion across the categories *human*, *animal*, and *machine*, as measured respectively by retention rates and entropy values in Gollancz SF and NovelTM (Figure 3).

Among the three conceptual categories, *machine* terms demonstrate the lowest rate of conceptual self-retention, indicating a marked semantic volatility. In the science fiction corpus, only 25.8% of masked *machine* tokens are predicted by RoBERTa as belonging to the same conceptual category, compared to 30.2% in the general fiction corpus. This decline was statistically significant (Δ = -−0.044, 95% CI [−-0.063, −-0.035], $p < 0.001$) in our permutation test, suggesting that science fiction narratives induce a destabilisation of the conceptual coherence of the *machine* category. The *animal* category showed a similar pattern. In general fiction, 36.5% of masked *animal*

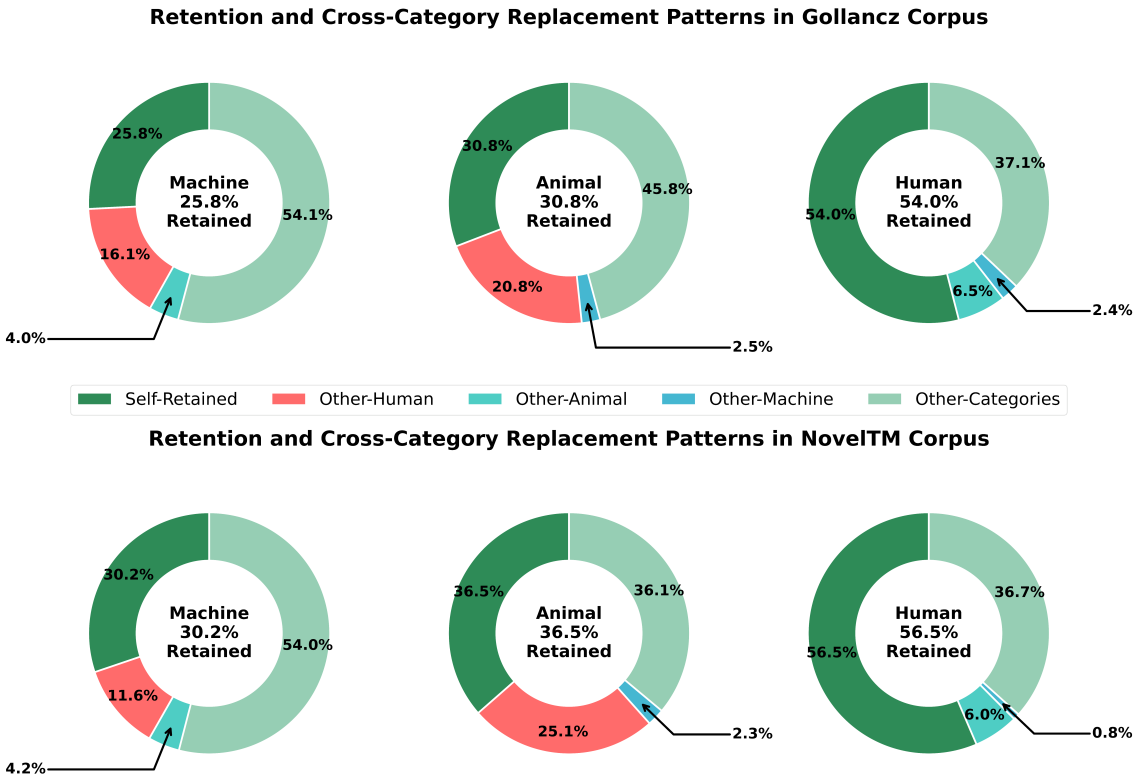

Figure 3: Cross-Entity Replacement Patterns in Gollancz SF and NovelTM Corpora

tokens are predicted within the same category, a figure that drops to 30.8% in science fiction (Δ = -−0.056, 95% CI [−-0.076, −-0.045], p < 0.001).

In contrast to the marked instability in the two categories above, the *human* category exhibits a notable semantic resilience across genres. Retention rates remain comparatively high, with 54.0% of masked *human* tokens in science fiction predicted as belonging to the same conceptual category, compared to 56.5% in general fiction. This modest decline did not reach statistical significance (Δ = -−0.021, 95% CI [−-0.045, −-0.004], p = 0.1). This relative consistency suggests that the *human* category maintains a high degree of semantic stability across both corpora, exhibiting limited permeability of conceptual boundaries when compared to other categories.

The entropy analysis substantiates and amplifies the findings drawn from retention rates by measuring the degree of predictive uncertainty associated with each conceptual category, namely the dispersion of plausible substitutes generated when the model is deprived of the original lexical item. Entropy analysis (Figure 4) shows a statistically significant corpus-by-category interaction ($F_{(2,20652)} = 15.7, p < .001$, η^2=0.0015), indicating that overall entropy differs between corpora, but that the pattern of entropy variation across categories (*human*, *animal*, *machine*) changes as a function of corpus type. In particular, masked *machine* tokens in science fiction exhibiting the highest mean entropy, exceeding that of both *animal* and *human* referents. [2]

This elevated entropy in the science fiction corpus indicates that the *machine* category elicits the widest semantic field, thereby marking it as the most conceptually unstable within that genre. The lower entropy observed in general fiction corpus, however, suggests that *machine* remains more semantically constrained outside the science-fictional context. When compared across the two corpora, only *machine*-related tokens register statistically significant entropy increase, from

[2] The low η^2 reflects the inherent variability of sentence-level data rather than a weak effect. In linguistic datasets, small η^2 values are expected and still capture reliable, generalisable differences across conditions.

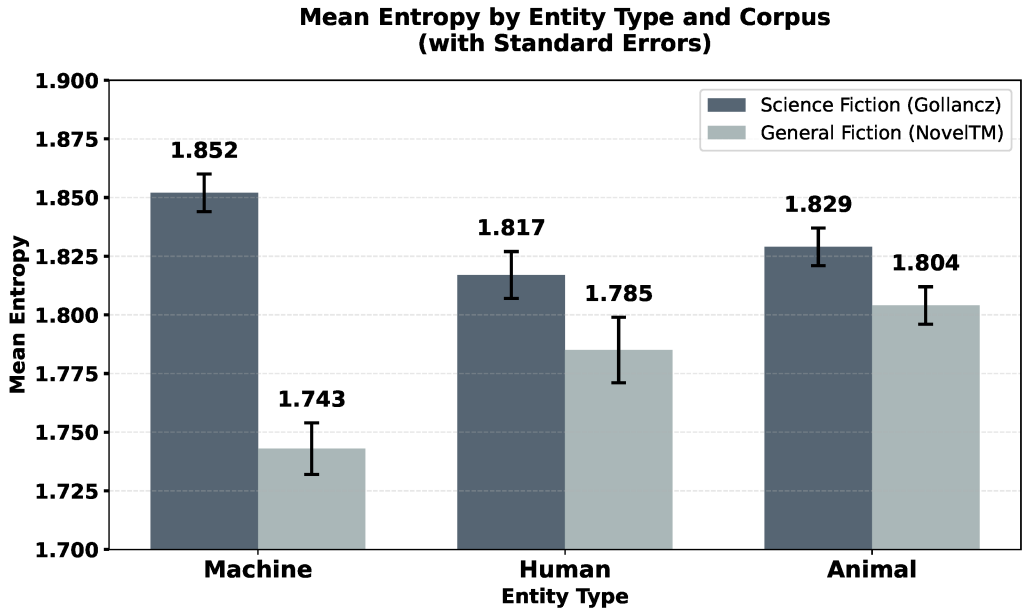

Figure 4: Mean Entropy by Entity Type and Corpus (with Standard Errors)

$H = 1.743$ in NovelTM to $H = 1.852$ in Gollancz SF ($t = 8.93, p < .001$, Bonferroni post-hoc test). By contrast, the changes in entropy of either *human* or *animal* concepts did not reach statistical significance ($t = 1.87$ and $t = 2.04$, respectively). While this pattern broadly accords with existing expectations about science fiction's treatment of technology, it offers a statistical means of showing how that semantic pressure is enacted on machine referents in language, expanding their substitutional latitude and, by extension, destabilising their categorical fixity.

5.2 Semantic Replacement and Conceptual permeability Across Genres

Building upon the findings of the preceding subsection, this subsection examines how conceptual destabilisation manifests through replacement rate analysis (Figure 5).

Comparison of replacement rates reveals a pronounced asymmetry in directional substitution across genres. In Gollancz SF, 16.1% of masked *machine* tokens were replaced with *human*-related terms, while only 2.4% of *human* tokens were substituted with *machine*-related ones. This directional imbalance is echoed in the NovelTM sample corpus, albeit at lower magnitudes, 11.6% and 0.8%, respectively, indicating that the anthropomorphisation of *machines* is a widespread feature of fiction more generally, yet notably accentuated in Gollancz SF, where *machine*-to-*human* substitutions increase by approximately 38.8%.

The permeability of boundaries between *human* and *animal* categories also exhibits a marked asymmetry. In Gollancz SF, 6.5% of *human* tokens were replaced with *animal*-related terms, while 20.8% of *animal* tokens were replaced by *human* terms. NovelTM yields similar directional asymmetry: 6.0% of *human* tokens were replaced with *animal* references, while 25.1% of *animal* tokens were supplanted by *human* terms. This imbalance suggests that, in the NovelTM corpus, sentences containing *animal* tokens more frequently align with the linguistic context in which RoBERTa expects *human* referents to occur, revealing a more frequent tendency toward anthropomorphism. The Gollancz SF corpus, by contrast, exhibits this pattern less often. In the opposite direction, *human* tokens in the Gollancz corpus are more often situated in contexts that resemble those in which RoBERTa anticipates *animal* referents, indicating a more frequent inclination toward zoomorphism. Such an inversion in narrative perspective indicates a partial decentring of the human subject in Gollancz SF, a gesture towards post-anthropocentric thinking that, while not

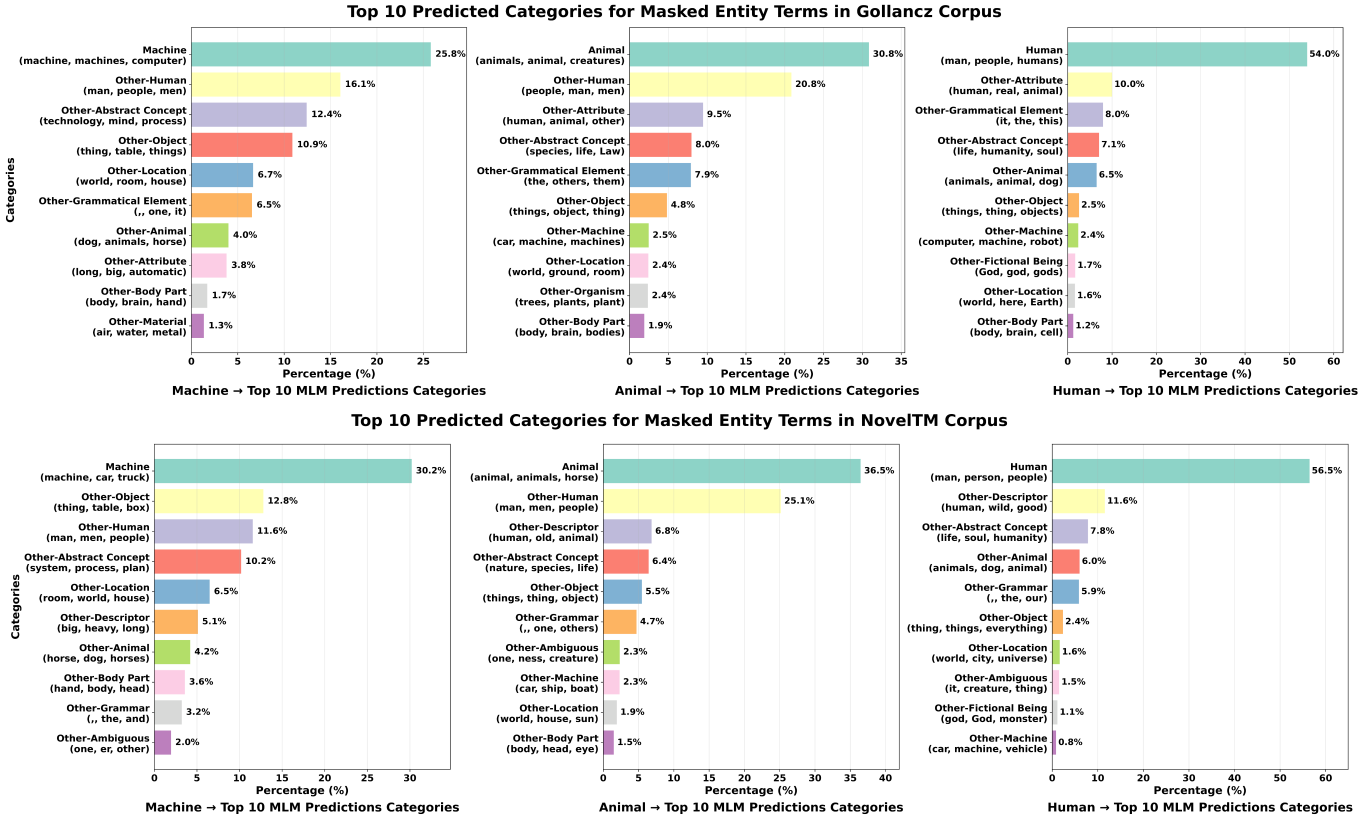

Figure 5: Top 10 Predicted Categories for Masked Entity Terms in Gollancz SF and NovelTM Corpora

overriding the broader human-oriented defaults of language, nonetheless differentiates the genre in its narrative strategies.

The boundary between *animal* and *machine* categories also exhibits signs of permeability, albeit less markedly than in other category pairings. Despite the relatively low overall rates, the bidirectional substitutions observed in both corpora indicate a nuanced semantic overlap between zoological and mechanical domains. In Gollancz SF corpus, 2.5% of masked *animal* tokens were replaced with *machine*-related terms, while 4.0% of *machine* tokens were substituted with *animal*-related terms. The NovelTM corpus displays a comparable pattern, with a slightly lower replacement rate in the animal-to-machine direction (2.3%) and a slightly higher rate in the machine-to-animal direction (4.2%) compared to the Gollancz SF corpus. The directionality of these replacements reveals contrasting narrative tendencies across genres. Gollancz SF more frequently situates animals within machinic narratives, attributing mechanical qualities to animals; NovelTM, by contrast, more frequently describes machines through animal-related language, imbuing them with animacy.

6 Discussion

Across measures of retention, replacement, and entropy, Gollancz SF demonstrates a heightened degree of conceptual permeability and a higher incidence of semantic category disruption than NovelTM. These patterns indicate that the syntactic and semantic contexts surrounding the categories *human*, *animal*, and *machine* are less constrained by the entrenched linguistic norms RoBERTa has internalised from standard English usage than those found in NovelTM. Silverstein has argued that these linguistic patterns are structured by a referential hierarchy which maps to degrees of perceived animacy [20]. In this hierarchy, the sequence proceeds down the scale from first- and second-person pronouns at the top, to third-person pronouns, then to proper names, human common nouns, other animates, and finally to inanimates [20]. Our results reflect a partial adherence to this hierarchy: in both corpora, retention rates follow the order *human* > *animal* > *machine*, consistent with the predicted ranking of animacy and agency. However, the patterns of cross-category replacement reveal reorganisitions of this ordering, with the degree of deviation differing across the two corpora.

This reorganisation is more complex and fine-grained than the statistical overviews in our results section can convey. In the following discussion we show how RoBERTa's "wrong" predictions offer the key to more fine-grained insights. These can be navigated by adopting what Bamman et al. have termed *classification-assisted close reading* [21] — an approach that mobilises machine classification not as an end in itself, but as a supporting method of textual triage, directing critical attention to sites of conceptual instability. By classifying the more than one hundred thousand outputs generated by RoBERTa's MLM through Gemini, the analysis enables a targeted investigation of predictions. Such an approach enables us to trace a persistent pattern of conceptual permeability across the categories *human*, *animal*, and *machine*: a phenomenon that resonates with theoretical accounts of posthuman technoculture, in which signification is no longer anchored to a stable referent but is instead distributed across a networked semantic field [22].

When *machine* terms are replaced by human referents, the lexical contents of these substitutions reveal distinct genre-specific tendencies. In Gollancz SF, the five most frequent replacements for masked *machine* tokens are *man*, *people*, *woman*, *human*, and *guy*, each encompassing both singular and plural forms. These predictions indicate a gendered and age-coded reconfiguration of machinic entities. A notable feature is RoBERTa's frequent prediction of *human* itself, rather than gender-signifying nouns such as *man* or *woman*. This pattern suggests that, compared with NovelTM, Gollancz SF more often assimilates *machines* to the superordinate *human* category without specifying a finer-grained social role. In contrast, while the NovelTM corpus also registers human-subcategorical replacements such as *man* and *woman*, its broader distribution gravitates

toward functionally defined or service-oriented role terms including *soldier*, *driver*, and *pilot*. The divergence between the two corpora thus signals distinct conceptions of linguistic agency: Gollancz SF reimagines the *machine* as a bearer of subjectivity grounded in *human* social and affective registers, whereas NovelTM constructs it as an extension of *human* labour and function.

That subjectivity, however, is not monolithic. The model reveals that the humanisation of the *machine* across the Gollancz SF corpus also disrupts stereotypical interactional norms, which are surfaced through the different sub-categories of *human* entities predicted by RoBERTa for *machine*. For example, in John O'Neill's *Land Under England* (1935), the model substitutes *sons* for *machines* in the sentence "They might as well have tried to marry her to one of their `[MASK]`" [23]. In this instance, a mechanical referent is reconstituted as a gendered, kinship-bound subject, male offspring eligible for marriage, thereby reinscribing familial and heteronormative logics into a passage that, in its original estranging design, sought to unsettle such anthropocentric orderings. This act of reconstitution is animated by the human epistemologies sedimented in RoBERTa's linguistic substrate and, in cognitive terms, enacts what Epley, Waytz, and Cacioppo term Elicited Agent Knowledge [24], whereby the unfamiliar is assimilated through the most readily available *human* categories, reaffirming the very anthropocentric order the text itself sought to estrange.

By contrast, when *human* is the masked token, the model's substitutions occasionally suggest mechanical terms, but at much lower rates. This suggests that the mechanised *human* is a markedly less prevalent trope in RoBERTa's training data, and may register as cognitively dissonant for the reader as well as statistically improbable for the model. This dynamic is particularly evident in RoBERTa's treatment of the phrase *human machine* across several Gollancz SF texts. The metaphor of the *human machine* has a long and fraught intellectual lineage, rooted in early modern mechanistic philosophy [25; 26]. In H. G. Wells's *The War of the Worlds*, the first instance of this linguistic formulation in the corpus appears in the sentence "I began to compare the things to `[MASK]` machines, to ask myself for the first time in my life how an ironclad or a steam engine would seem to an intelligent lower animal". RoBERTa ranks *mechanical machines* among its top predictions [27]. The same pattern recurs in Karel Čapek's *R.U.R. / War with the Newts*, where "The `[MASK]` machine, Miss Glory, was terribly imperfect" prompts substitutions such as *sewing* or *washing*, collapsing complex subjectivity into the register of domestic appliance [28]. In highlighting how Gollancz SF challenges the statistical regularities of the model's semantic space, we can see the appeal of the *human machine*, both as concept and linguistic formulation, for science fiction authors.

When *animal* is masked, the substitutions that Gemini groups under the category *Fictional being* exhibit a genre-sensitive split across the two corpora. In NovelTM, the most common replacements are *god* and *angel*, alongside a smaller presence of monstrous or diabolical figures such as *monster* and *devil*. In Gollancz SF, by contrast, the leading substitute is *monster*, followed closely by *god*, with *angel*, *alien*, and *ghost* also prominent, which steers the field first toward abjected monstrosity and only thereafter toward the divine. An instance in which *animal* is replaced by *God* appears in Herbert's *Dune* [29], where the masked sentence "Humans must never submit to `[MASK]`" yields *God* as the top ranked prediction. This outcome reframes the *animal* slot within a sacral register, binds it to a semantic field of obedience and reverence, and imagines a hierarchy above the human that requires submission. This animal-to-fictional-being substitution pattern differs from the other two source categories. Across both Gollancz SF and NovelTM, when *Human* tokens are replaced by predictions within the *Fictional Being* category, they are most commonly refigured as divine beings such as *god* or *angel*, which suggests an upward, aspirational verticality, even though spectral or monstrous readings remain available at the margins. *Machine* tokens, by contrast, are shaped by genre: in NovelTM the distribution centres on theistic titles while non-theistic figures persist, whereas in science fiction it widens across a more heterogeneous mythic and monstrous field that includes *deity*, *spirit*, *demon*, *angel*, and *alien*. Across all three core

categories, substitutions into *Fictional Being* occur more frequently and span a wider lexical spectrum in Gollancz SF than in NovelTM, both in overall incidence and relative to category scope. This reflects a diversification of permeability, as science fiction positions *human*, *animal*, and *machine* within linguistic settings that resonate with a broader set of category contexts internalised by RoBERTa.

Building on the bifurcated imaginary that the category of *Fictional being* establishes across the three domains, our MLM predictions disclose a higher ontological register above the *human* axis, inhabited by figures of the divine and the godlike. Set against a longer intellectual history, this vertical inflection recalls Lovejoy's *Great Chain of Being*, which conceives existence as a hierarchical continuum extending from *God* at the apex, through the angelic orders, to rational humanity, then to animal life, vegetal life, and finally to inanimate nature [30]. Although *machines* do not occupy a canonical rung in this classical *Great Chain of Being*, modern discourse repeatedly threads them into the same vertical schema, either by analogising organisms to mechanisms or by treating devices as agents that unsettle the human's median rank [25; 26; 31; 32]. The following example makes this vertical ordering legible at the level of linguistic form. In Walter M. Miller Jr.'s *Conditionally Human*, "Anthropos feared making quasi-human too intelligent, lest sentimentalists proclaim them really human" [33]. When prompted with "Anthropos feared making quasi-`[MASK]` too intelligent, lest sentimentalists proclaim them really human", RoBERTa predicts *bots*, which positions quasi-bots beneath the human. Conversely, when completing "Anthropos feared making quasi-human too intelligent, lest sentimentalists proclaim them really `[MASK]`", the model returns *gods*, tracing a conceptual arc of *quasi-human* $\rightarrow$ *god*. The resulting topology, *machine* $<$ *human* $<$ *gods*, implicitly restores a vertical scale of being and positions the *human* as both a referential anchor and a threatened middle term. This sentence does more than expose latent hierarchies; it constructs an unstable hybrid figure through the prefix *quasi-*, gestures towards the fluidity of ontological borders, and simultaneously reproduces the logic of tiered status it appears to challenge.

In this regard, the topology reconstructed by RoBERTa stands at odds with Paul Gilroy's account of *planetary humanism*, which calls for an expansive and non-hierarchical humanism attentive to shared vulnerability, dignity, and conviviality beyond racialised and imperial boundaries [34]. Yet it is precisely through modelling these substitutions with MLM that the linguistic infrastructures naturalising hierarchy become visible, even where science fiction aspires to dismantle them.

7 Conclusion

Instruments, whether conceptual or mechanical, shape the kinds of work we can do and the questions we can ask. This study has demonstrated how masked language models, when used as interpretive instruments, can reveal latent semantic dynamics in literary texts, enabling a computational engagement with ontological instability and conceptual permeability. Our findings on the intermingling of *human*, *machine*, and *animal* categories resonate with Rosi Braidotti's conceptualisation of the posthuman subject as the convergence of *zoē* (the life of all living beings), *bíos* (the life of humans organised in society), and technological agency [35], suggesting that literary language itself encodes such entangled formations. More broadly, we observe the opportunity for this methodological pipeline for the computational study of literary analysis, reception studies, and conceptual history [15]. The lexical extraction step, in particular, can be readily adapted to explore alternative conceptual binaries, allowing for flexible extensions of the framework.

Recent debates have raised a further question about the role of interpretation itself. Scholars have argued that we may be entering the era of the death of the reader: a moment when machines read for us, summarising without surprise and extracting without encounter [36]. Our findings underline what human readers have always known — so obvious it has rarely needed saying — that the power of reading lies not in retrieval but in challenge. Reading tests a point of view, pushes

against a world-view, and tugs at the seams of our categories. Science fiction's estrangements carry force only when they meet a mind that can be unsettled. A model can detect deviation; it cannot be deviated from. If literature is to do its work, a human reader must be in the loop.

This recognition does not stand in opposition to computational methods, but a chance to clarify and coordinate their interpretive function alongside that of traditional reader-led analysis. We would contend that our findings show how MLMs can serve as interpretive partners, illuminating sites of textual frisson and literary surprise, based on the level of textual deviation from what the model has determined is statistically likely. The power of this method is the way it employs classification to assist the intractability of close reading across large-scale corpora, connecting insights at the level of the sentence to meso-scale patterns [37] and macro-scale shifts across (and between) large corpora, from the peculiarities of a particular author, to markers of genres and sub-genres, to longitudinal temporal shifts. As such, we believe that this is a fruitful area of inquiry, as well as an intellectual process that explores the permeability of the disciplinary interface between the humanities and machine learning.

Acknowledgments

We gratefully acknowledge support from Riksbankens Jubileumsfond, *Change is Key!* (Grant No. M21-0021; award to Haim Dubossarsky), and from the QMUL–CSC PhD Scholarship program (China Scholarship Council; Grant No. 202408890010; award to Yuxuan Liu).

References

[1] Suvin, Darko. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. New Haven, CT: Yale University Press, 1979.

[2] Shklovsky, Viktor. "Art as technique". In: *Literary Theory: An Anthology*, ed. by Julie Rivkin and Michael Ryan. 3rd ed. Chichester, UK: Wiley-Blackwell, 2017, pp. 8–14.

[3] Braidotti, Rosi. *The Posthuman*. Cambridge, UK: Polity Press, 2013.

[4] Haraway, Donna J. *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York, NY: Routledge, 1991.

[5] Wolfe, Cary. *What Is Posthumanism?* Minneapolis, MN: University of Minnesota Press, 2010.

[6] Firth, J. R. *Papers in Linguistics 1934–1951*. London: Oxford University Press, 1957.

[7] Harris, Zellig. "Distributional structure". In: *Word 10*, no. 2–3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

[8] Turney, Peter D. and Pantel, Patrick. "From frequency to meaning: Vector space models of semantics". In: *Journal of Artificial Intelligence Research 37* (2010), pp. 141–188. DOI: 10.1613/jair.2934.

[9] Baroni, Marco, Dinu, Georgiana, and Kruszewski, Germán. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. Baltimore, MD: Association for Computational Linguistics, 2014, pp. 238–247. DOI: 10.3115/v1/P14-1023.

[10] Underwood, Ted. "Mapping the latent spaces of culture". Humanities Commons. 2021. DOI: 10.17613/faaa-1r21.

[11] Klein, Lauren, Martin, Michael, Brock, Andrew, Antoniak, Maria, Walsh, Melanie, Johnson, Jessica Marie, Tilton, Lauren, and Mimno, David. “Provocations from the humanities for generative AI research”. arXiv preprint. 2025. arXiv: 2502.19190 [cs.CL].

[12] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Minneapolis, MN: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[13] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. “RoBERTa: A robustly optimized BERT pretraining approach”. arXiv preprint. 2019. DOI: 10.48550/arXiv.1907.11692. arXiv: 1907.11692.

[14] Ardanuy, Mariona Coll, Nanni, Federico, Beelen, Kaspar, Hosseini, Kasra, Ahnert, Ruth, Lawrence, Jon, McDonough, Katherine, Tolfo, Giorgia, Wilson, Daniel C. S., and McGillivray, Barbara. “Living machines: A study of atypical animacy”. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Barcelona, Spain (online): International Committee on Computational Linguistics, 2020, pp. 4534–4545.

[15] Wilson, Daniel C. S., Ardanuy, Mariona Coll, Beelen, Kaspar, McGillivray, Barbara, and Ahnert, Ruth. “The living machine: A computational approach to the nineteenth-century language of technology”. In: *Technology and Culture 64*, no. 3 (2023), pp. 875–902. DOI: 10.1353/tech.2023.a903976.

[16] Hosseini, Kasra, Beelen, Kaspar, Colavizza, Giovanni, and Ardanuy, Mariona Coll. “Neural language models for nineteenth-century English”. In: *Journal of Open Humanities Data 7*, no. 0 (2021), p. 22. DOI: 10.5334/johd.48.

[17] Underwood, Ted, Kimutis, Peter, and Witte, Jessica. “NovelTM datasets for English-language fiction, 1700–2009”. Journal of Cultural Analytics. 2020. DOI: 10.22148/001c.13147.

[18] Hudak, Dave, Johnson, Doug, Chalker, Alan, Nicklas, Jeremy, Franz, Eric, Dockendorf, Trey, and McMichael, Brian L. “Open OnDemand: A web-based client portal for HPC centers”. In: *Journal of Open Source Software 3*, no. 25 (2018), p. 622. DOI: 10.21105/joss.00622.

[19] DeepMind, Google. “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities”. arXiv preprint. 2025. DOI: 10.48550/arXiv.2507.06261. arXiv: 2507.06261.

[20] Silverstein, Michael. “Hierarchy of features and ergativity”. In: *Grammatical Categories in Australian Languages*, ed. by R. M. W. Dixon. Reprinted in *Features and Projections*, edited by Pieter Muysken and Henk van Riemsdijk, Dordrecht: Foris, 1986. Canberra: Australian Institute of Aboriginal Studies, 1976, pp. 112–171. DOI: 10.5281/zenodo.4688088.

[21] Bamman, David, Chang, Kent K., Lucy, Li, and Zhou, Naitian. “On classification with large language models in cultural analytics”. In: *Proceedings of the Computational Humanities Research Conference (CHR 2024)*. Vol. 3834. CEUR Workshop Proceedings. Aarhus, Denmark: CEUR-WS, 2024, pp. 494–527.

[22] Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, IL: University of Chicago Press, 1999.

[23] O’Neill, Joseph. *Land Under England*. SF Masterworks. London: Gollancz, 2018. ISBN: 9781473224063.

[24] Epley, Nicholas, Waytz, Adam, and Cacioppo, John T. “On seeing human: A three-factor theory of anthropomorphism”. In: *Psychological Review 114*, no. 4 (2007), pp. 864–886. DOI: 10.1037/0033-295X.114.4.864.

[25] Descartes, René. *The Philosophical Writings of Descartes*. Trans. by John Cottingham, Robert Stoothoff, and Dugald Murdoch. 2 vols; cites *Discourse on the Method*, vol. I, pt. V. Cambridge: Cambridge University Press, 1985.

[26] La Mettrie, Julien Offray de. *Machine Man and Other Writings*, ed. by Ann Thomson. Trans. by Ann Thomson. Cambridge Texts in the History of Philosophy. Cambridge: Cambridge University Press, 1996. DOI: 10.1017/CBO9781139166713.

[27] Wells, H. G. *The War of the Worlds*, ed. by Martin A. Danahay. Peterborough, ON: Broadview Press, 2003.

[28] Čapek, Karel. *R.U.R. / War with the Newts*. London: Gollancz, 2011. ISBN: 9780575099456.

[29] Herbert, Frank. *Dune*. SF Masterworks. London: Gollancz, 2007. ISBN: 9780575081505.

[30] Lovejoy, Arthur O. *The Great Chain of Being: A Study of the History of an Idea*. Cambridge, MA: Harvard University Press, 1936.

[31] Leibniz, G. W. “Monadology”. In: *Philosophical Essays*, ed. by Roger Ariew and Daniel Garber. Trans. by Roger Ariew and Daniel Garber. Cited as §§64–69. Indianapolis, IN: Hackett Publishing Company, 1989.

[32] Riskin, Jessica. *The Restless Clock: A History of the Centuries-Long Argument over What Makes Living Things Tick*. Chicago, IL: University of Chicago Press, 2016.

[33] Jr., Walter M. Miller. *Conditionally Human*. New York, NY: Ballantine Books, 1962.

[34] Gilroy, Paul. “Postcolonial melancholia”. In: *The New Social Theory Reader*, ed. by Steven Seidman and Jeffrey C. Alexander. 2nd ed. London: Routledge, 2008, pp. 427–434.

[35] Braidotti, Rosi. *Posthuman Knowledge*. Cambridge, UK: Polity Press, 2019.

[36] Baron, Naomi S. *Reader Bot: What Happens When AI Reads and Why It Matters*. Stanford, CA: Stanford University Press, 2026. ISBN: 9781503643949.

[37] Ahnert, Ruth, Ahnert, Sebastian E., Coleman, Catherine Nicole, and Weingart, Scott B. *The Network Turn: Changing Perspectives in the Humanities*. Elements in Publishing and Book Culture. Cambridge: Cambridge University Press, 2021. DOI: 10.1017/9781108866804.

A Supplementary Figures

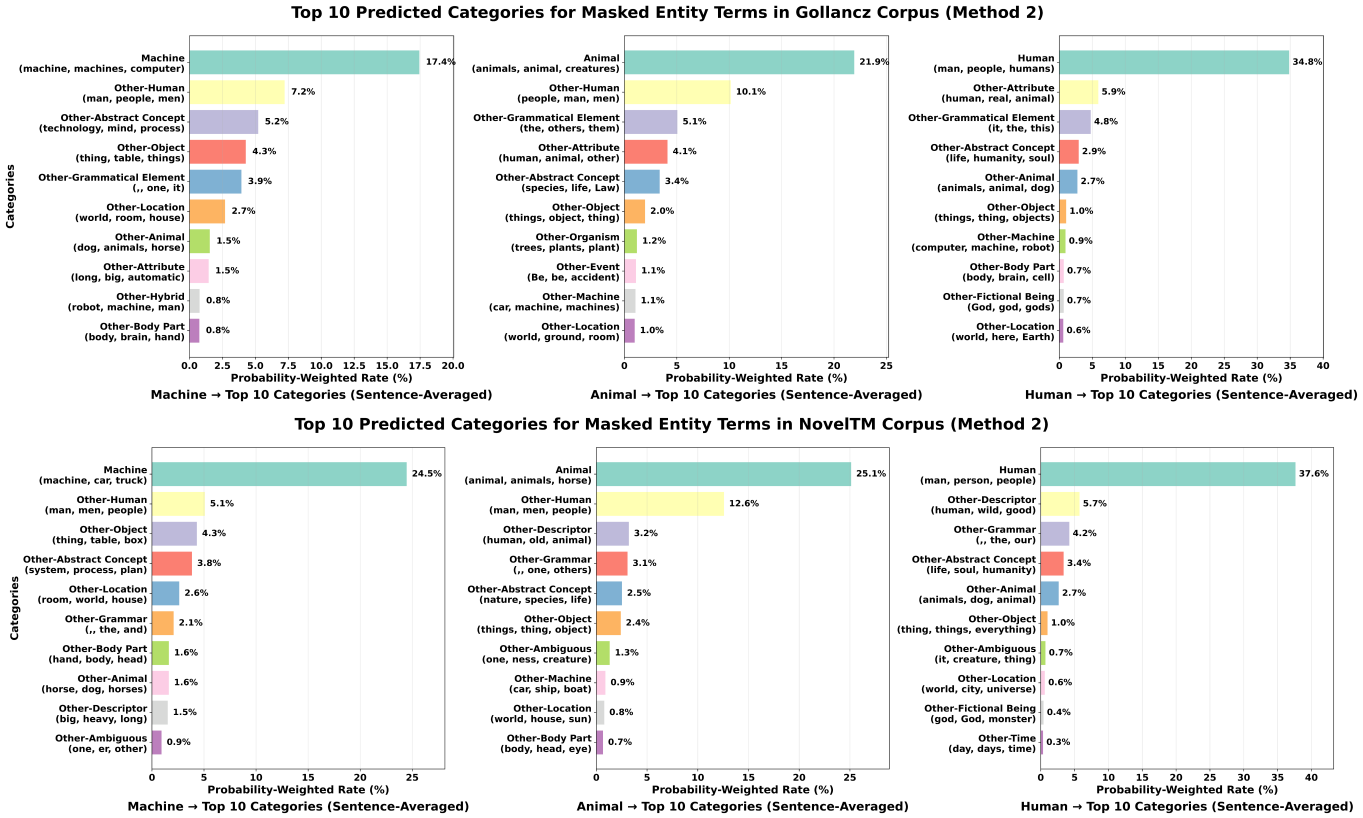

Figure 6: Probability-Weighted Top 10 Predicted Categories for Masked Entity Terms in Gollancz SF and NovelTM Corpora