Expect the Unexpected? Testing the Surprisal of Salient Entities

Jessica Lin and **Amir Zeldes**
Department of Linguistics
Georgetown University
{yl1290, amir.zeldes}@georgetown.edu

Abstract

Previous work examining the Uniform Information Density (UID) hypothesis has shown that while information as measured by surprisal metrics is distributed more or less evenly across documents overall, local discrepancies can arise due to functional pressures corresponding to syntactic and discourse structural constraints. However, work thus far has largely disregarded the relative salience of discourse participants. We fill this gap by studying how overall salience of entities in discourse relates to surprisal using 70K manually annotated mentions across 16 genres of English and a novel minimal-pair prompting method. Our results show that globally salient entities exhibit significantly higher surprisal than non-salient ones, even controlling for position, length, and nesting confounds. Moreover, salient entities systematically reduce surprisal for surrounding content when used as prompts, enhancing document-level predictability. This effect varies by genre, appearing strongest in topic-coherent texts and weakest in conversational contexts. Our findings refine the UID competing pressures framework by identifying global entity salience as a mechanism shaping information distribution in discourse.

1 Introduction

A large body of work in discourse processing has examined what makes certain entities central to a text. From a discourse perspective, texts invariably have central participants (entities the discourse is fundamentally "about") and peripheral ones that play supporting roles (Givón 1983, Chafe 1994, 182–185). This notion of centrality operates at the document level, differing from local attentional salience studied at the utterance-level e.g. in Centering Theory (Grosz et al., 1995). While local salience determines which entities are most prominent at any given moment through mechanisms such as grammatical role and recency, global salience captures which entities are most important to the discourse as a whole. Little is known about how this global notion of salience interacts with information distribution across extended discourse.

Information-theoretical work suggests that speakers tend to structure text such that information flow is approximately even (Fenk and Fenk, 1980; Jaeger and Levy, 2006). This Uniform Information Density (UID) hypothesis posits that surprisal (negative log-probability) of new words should remain relatively constant. However, perfect uniformity is hardly achieved because multiple competing pressures simultaneously shape how speakers and writers build discourse (Clark et al., 2023; Pimentel et al., 2021; Tsipidi et al., 2024). Recent work has identified syntactic constraints (Clark et al., 2022), phonological factors (Pimentel et al., 2021), and discourse structures (Tsipidi et al., 2024) as pressures that create departures from uniformity. This raises the question: does discourse-level referential structure also exert pressure on information distribution? On the one hand, salient entities may be high in informational content, and therefore more surprising. On the other hand, they may be central to the topics of the documents they appear in, making document content more predictable when they appear. Rather than predicting a global shift in average information density, we hypothesize that global salience gives rise to localized non-uniformities in discourse. In particular, salient entities tend to be followed by stretches of reduced surprisal, creating troughs in surprisal contours that reflect increased expectations about upcoming content.

To investigate this relationship, we leverage a dataset of over 70,000 entity mentions across 16 genres of spoken and written English, in which entities are annotated for global salience based on summary-worthiness (Lin and Zeldes, 2025): entities appearing consistently across multiple independent summaries of a document are treated as globally salient. We examine how this measure

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30618–30629

relates to surprisal computed using language model probabilities, addressing three research questions:

1. Do globally salient entities themselves exhibit different surprisal profiles than non-salient entities in naturally occurring discourse?

2. Do globally salient entities systematically reduce surprisal for surrounding discourse content compared to non-salient entities?

3. Does the relationship between global salience and surprisal vary systematically across genres?

For RQ1, we find significant differences in raw corpus data where multiple pressures operate simultaneously. For RQ2, we develop a minimal-pair paradigm controlling for confounds to demonstrate that globally salient entities increase predictability for document contents. For RQ3, we find stronger effects in more topic-coherent genres (i.e. ones focuses on single topic, such as academic articles) than in topic-shifting ones (e.g. conversation).

Our results provide evidence, to our knowledge for the first time, that globally salient entities are more surprising and systematically reduce surprisal for upcoming content. For UID, our results support a view in which uniformity is a tendency shaped by competing constraints, with global salience contributing a referential mechanism that interacts with predictability across discourse.

2 Related Work

2.1 UID and Competing Pressures

The Uniform Information Density (UID) hypothesis proposes that speakers tend to distribute information so that surprisal remains approximately even across a discourse (Fenk and Fenk, 1980; Jaeger and Levy, 2006). A large body of work supports this tendency, showing that speakers avoid sequences of extremely high or low surprisal when alternative formulations are available. For example, speakers use optional complementizers ("that") more often before less predictable clauses to smooth information flow (Jaeger and Levy, 2006).

However, many studies have shown systematic departures from uniformity driven by independent linguistic pressures. Pimentel et al. (2021) report a robust cross-linguistic pattern in which word-initial segments carry higher surprisal than word-final segments, reflecting phonotactic and information structural pressures to front-load disambiguatory information even when this produces locally uneven information distribution. At the syntactic level, Clark et al. (2023) demonstrate that real word orders are more uniform than most counterfactual reorderings, but that only linguistically implausible grammars achieve perfectly uniform profiles, indicating that grammatical constraints limit the degree of uniformity that can be achieved. For longer texts, Tsipidi et al. (2024) show that surprisal varies systematically across a document and that hierarchical discourse structure predicts these non-uniform contours. Together, these findings motivate viewing UID as a tendency that interacts with other linguistic and discourse pressures to shape information distribution.

2.2 Discourse Salience and Surprisal

The notion of salience in discourse processing operates at multiple levels. Local salience, studied extensively in Centering Theory, determines which entities are most prominent at any given moment through mechanisms such as grammatical role, recency, and pronominalization (Grosz et al., 1995). By contrast, global salience captures document-level importance: which entities the discourse is fundamentally "about" (Givón, 1983; Grosz and Sidner, 1986). These views are complementary rather than competing, addressing different aspects of salience.

Previous work on whether salient entities themselves are more or less predictable has yielded mixed results. Some studies suggest that marked or salient forms carry higher surprisal (Rácz, 2013; Zarcone et al., 2016), while others show that discourse-prominent entities become more predictable through repeated mention and establish topicality (Givón, 1983; Arnold, 2010). Multiple factors influence entity predictability, including grammatical role (Grosz et al., 1995), recency (Arnold, 2025), and referential form choice (Arnold, 2010; Rohde and Kehler, 2014), making it difficult to isolate salience effects in natural contexts.

No previous study has examined the surprisal of globally salient entities, which have been operationalized through summary-worthiness in previous work (Lin and Zeldes, 2025). We hypothesize that globally salient entities are more surprising locally, but reduce surprisal across extended discourse for

other document contents, creating systematic departures from uniform information distribution, analogous to phonotactic or syntactic pressures identified in prior work.

2.3 Operationalizing Global Entity Salience

While there are many kinds and definitions of salience (see von Heusinger and Schumacher 2019; Boswijk and Coler 2020), work in Computational Linguistics has focused on the identification of the most prominent, notable or memorable entities in a document, which can be identified by a number of methods, including direct annotation via human intuition (Dojchinovski et al., 2016), extraction from user click-stream data (Gamon et al., 2013), use of hyperlinks and categories in Wikipedia data (Wu et al., 2020) and automatic alignment with document summaries (Dunietz and Gillick, 2014). The latter paradigm in particular has recently been extended to the extraction of gradient salience ratings based on multiple summaries of a document (Lin and Zeldes, 2025), based on the idea that if an entity is salient it will be difficult to write a summary without mentioning it, and that the number of summaries mentioning an entity can represent its relative salience. Below we use this definition and data to collect and compare surprisal and salience scores.

3 Methods

3.1 Data

The dataset we use for this paper, **GUM**-based **S**ummary **A**ligned **G**raded **E**ntities (GUM-SAGE, Lin and Zeldes 2025), is built on version 11 of the GUM corpus (Zeldes, 2017), an open-access, multilayer resource for English spanning over 250K tokens across 16 genres. GUM features a variety of annotations such as Universal Dependencies parses (de Marneffe et al., 2021), entity types, Wikification (Lin and Zeldes, 2021), discourse parses (Liu et al., 2024), and crucially for our purposes, coreference resolution (Zhu et al., 2021; Zeldes, 2022).

Each document in GUM is accompanied by 5 summaries following strict guidelines (Liu and Zeldes, 2023), which are either all human-written (in the dev/test sets) or one human and 4 model-generated ones (training set). In GUM-SAGE, entities mentioned in the documents were aligned to the summaries to derive salience scores corresponding to **summary worthiness**, based on the idea that summaries will tend to include the most salient entities. A score of 5 is assigned to entities present in all summaries, and a score of 1 to entities appearing in only one, with 0 representing totally non-salient entities (mentioned in no summary, about 84.5% of entities in the dataset).

Since the data contains manual coreference annotations, we are able to rate all mentions[1] of each entity, regardless of how it is mentioned (e.g. proper name, common noun or pronoun). The data contains just over 70K mentions of over 31K unique entities. See Table 1 for an overview of the data.

3.2 Surprisal

We measured entity surprisal using different experimental scenarios to systematically control for potential confounding factors (see below), but in all cases, token-level surprisal scores[2] were computed by processing each document with a sliding window of 1,024 tokens using a distilled version of GPT-2 language model (See Appendix A for implementation details). For a multi-word mention such as *a full romantic evening*, mention-level surprisal scores were computed by averaging token-level surprisals, and entity-level surprisal scores for the cluster of all mentions of a single entity in a document were obtained by averaging mention-level scores across all mentions of the entity, based on the gold standard coreference annotations available in the corpus.

Since surprisal can be affected by document length (the longer the document, the more context for later mentions), and document lengths vary across genres, we apply standard scaling to raw mean surprisal scores, resulting in **mean surprisal z-scores** (i.e. an entity's mentions' mean surprisal score can be z standard deviations above or below the mean of its document). We use this method throughout the paper, and when surprisal scores are discussed they should be understood in terms of negative and positive standard deviations away from the mean scores for mentions or entities in

[1] We use *mention* to refer to any textual span annotated as referring to an entity in GUM, including proper names, common nouns, and pronouns. GUM permits nested mentions, so phrases such as (1) [the tips of the noses of the aardvarks] contain multiple markables (e.g., the aardvarks, the noses of the aardvarks, the tips of the noses of the aardvarks), each treated as a distinct entity mention.

[2] Token-level surprisal is the standard unit in UID research, since autoregressive language models define probabilities incrementally over tokens. Higher-level units (e.g., mentions or sentences) necessarily require aggregation from token-level estimates (Jaeger and Levy, 2006; Meister et al., 2021; Pimentel et al., 2021; Clark et al., 2022)

Genre	Docs	Tokens	Modality	Mentions	Entities	% Top1	% Top2	% Top3	% Salient
Academic writing	18	17,169	Written	5,055	3,001	1.90	3.35	5.00	15.65
Biographies	20	18,213	Written	5,772	3,052	2.32	3.71	5.57	16.02
Vlog	15	16,864	Spoken	4,499	1,256	1.63	3.44	5.61	14.83
Conversations	15	17,932	Spoken	4,531	1,166	1.12	2.49	4.64	16.24
Courtroom transcripts	9	11,148	Spoken	2,938	1,053	1.80	3.46	5.82	14.68
Essays	9	10,842	Written	3,061	1,672	0.93	1.74	2.89	10.88
Fiction	19	17,511	Written	4,977	2,018	2.51	3.55	4.96	13.53
Forum (reddit)	18	16,364	Written	4,544	1,958	2.10	2.71	3.95	10.23
How-to guides	19	17,081	Written	4,468	2,011	2.17	2.85	4.22	14.73
Interviews	19	18,196	Spoken	5,216	2,293	2.98	3.49	4.21	6.68
Letters	12	9,989	Written	2,848	1,325	3.34	5.22	7.10	19.62
News stories	24	17,186	Written	5,023	2,305	3.80	5.03	7.40	15.56
Podcasts	10	11,986	Spoken	3,059	1,163	1.46	2.62	4.37	11.37
Political speeches	15	16,720	Spoken	4,847	2,297	1.84	2.71	3.46	7.79
Textbooks	15	16,693	Written	4,719	2,687	1.55	2.59	3.31	8.27
Travel guides	18	16,515	Written	4,471	2,559	1.09	1.45	2.54	10.63
Total GUM	**255**	**250,409**	-	**70,028**	**31,816**	**2.06**	**3.18**	**4.71**	**13.21**

Table 1: Overview of GUM-SAGE genres with document counts, token counts, modality, mention and entity counts, and proportion of salient entities by salience level (Top1: score=5, Top2: score $\geqslant$ 4, Top3: score $\geqslant$ 3). Percentages are calculated over all entities per genre.

the document, as appropriate. This procedure allows us to normalize across potentially dissimilar documents, which are inevitably rather different in a dataset spanning a broad range of spoken and written genres, and also makes intuitive sense if we want to contrast the salient and non-salient entities in each document as possibly positive or negative deviations from an average baseline.

4 Experiments

We present three experiments examining how global salience relates to surprisal. Experiment 1 (Section 4.1) measures entity-level surprisal in natural contexts, establishing a baseline that reveals how multiple pressures operate simultaneously in discourse. Experiments 2 and 3 (Section 4.2 and 4.3) focus on a complementary hypothesis, that salient entities will reduce the surprisal of their documents' contents more when used as prompts. Here we propose a novel minimal-pair paradigm to control for position, length, and nesting confounds, allowing us to measure whether globally salient entities reduce surprisal regardless of where they appear.

Since we aim to examine whether salience-predictability relationships generalize across diverse discourse contexts, all experiments use the GUM-SAGE dataset described in Section 3, which provides gradient salience scores across 16 genres of spoken and written English.

4.1 Natural context analysis

We first test whether globally salient entities exhibit different surprisal profiles than non-salient entities by measuring average token-level surprisal per mention of each entity and correlating this with salience scores.

Figure 1 shows mean surprisal changes relative to entities with salience score 0. Salient entities (scores 1–5) show significantly higher surprisal than non-salient entities ($t = -2.15$, $p = 0.031$), though the effect size is small ($\sim$0.1 standard deviations), and varies seemingly at random with salience score. This demonstrates that discourse-level salience (or 'summary-worthiness') exerts a measurable but almost negligible influence on information distribution even in naturally occurring text, though salience score matters less.

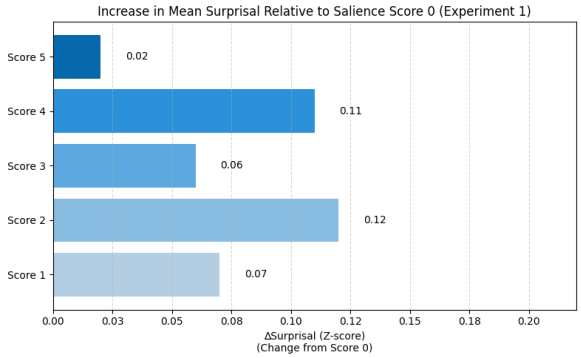


Figure 1: Change in mean surprisal for salience scores 1–5 relative to score 0 in Experiment 1.

To understand why the effect is so weak, we consider several confounding factors that complicate the relationship between salience and surprisal, using a linear regression model shown in Table 2.

Position effects. Position shows a positive effect ($\beta = 0.0002$, $p < 0.05$, ΔAIC=2.68), indicating

Predictor	Coef.	P-value	95% CI	ΔAIC
Intercept	0.8205	***	[0.764, 0.877]	–
Position	0.0002	*	[0.000, 0.000]	2.68
Salience score	-0.0471	***	[-0.070, -0.025]	14.86
Mention length	-0.0535	***	[-0.059, -0.048]	352.11
Is Nesting	-0.8493	***	[-0.903, -0.796]	886.27

Table 2: Results from an OLS regression model `Surprisal ~ Salience score + Mention length + Is Nested + Position`, with predictors ordered by ΔAIC from single-term deletions (least to most informative): larger ΔAIC indicates a larger degradation in model fit when that predictor is removed. Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

higher surprisal for mentions later in documents, perhaps due to less explicit descriptions.

Length and function words. Mention length substantially reduces surprisal ($\beta = -0.054$, $p < 0.001$, ΔAIC=352.11). Longer mentions are more predictable on average because averaging token-level surprisal disproportionately weights highly predictable function words. For example, out of context, (1) has lower mean surprisal than (2) despite containing the improbable noun *aardvarks*, driven by repeated predictable tokens such as *the* and *of*.

(1) [The tips of the noses of the aardvarks]$_{5.23}$

(2) [People]$_{8.12}$

Nesting. Nesting shows the strongest effect ($\beta = -0.849$, $p < 0.001$, ΔAIC=886.27): mentions that nest other mentions exhibit lower surprisal. This occurs because internal context makes the head noun more predictable. For example, comparing [the aardvarks] with [the noses of the aardvarks], the latter has lower average surprisal because "noses of" provides syntactic context that constrains what follows, whereas transitioning from external content to an NP is less predictable.

With these confounds incorporated, salience becomes a much more significant predictor, confirming that discourse salience influences information distribution substantially, with +1 salience score comparable to an extra word in length ($\sim$-0.05 coefficient for both). However, to establish a two way connection between salience and surprisal, we would also like to test whether salient entities are higher in content in the sense that they can make surrounding text less surprising, to which we turn next.

4.2 Surprisal reduction from salient entities

While measuring the surprisal of entity mentions in running texts is straightforward, measuring how informative an entity is for its text is more complicated: firstly, because some of the text in a document can precede that entity, and secondly, because of the confounds identified above: non-salient entities nesting salient ones may appear to reduce surprisal like the salient entities they contain, and longer mentions will provide more context, creating an unfair comparison. In order to test whether the content of salient entities reduces surprisal more than non-salient entities in a controllable way, we therefore develop two minimal-pair paradigms. The main idea of both is to test identical sentences paired with either salient or non-salient entity mention prompts.

In the first minimal-pair design, for each entity in a document we extract its first non-pronominal mention and place it as a prompt followed by a colon. We then measure the mean surprisal of each sentence in the document when placed after this prompt. For example, in a document about discrimination which mentions that psychologists have studied the phenomenon, we create pairs like:

(3) **Discrimination:** *The prevalence of discrimination across racial groups in contemporary America.*

(4) **Psychologists:** *The prevalence of discrimination across racial groups in contemporary America.*

Here, "Discrimination" (salient, score = 5) and "Psychologists" (non-salient, score = 0) from the same document are both followed by identical sentence content. We then compute surprisal only for the sentence tokens (excluding the prompt and colon), then average across all sentences in the document to obtain each entity's discourse-level effect on facilitating the predictability of its document's contents.

Because sentence lengths vary, for cross-document comparison, we standardize sentence-

level scores into z-scores using document-level means (μ_{doc}) and standard deviations (σ_{doc}) of entity-level averages (the mean of sentence surprisal scores $\overline{Surp}(sentence)$):

$$z = \frac{(\overline{Surp}(sentence) - \mu doc)}{\sigma doc} \quad (1)$$

This formulation naturally controls for position effects since we test each entity against all sentences in its document. If globally salient entities establish discourse-wide predictability, then on average they should reduce mean surprisal more than non-salient entities when used as prompts. However, we cannot control for length or nesting using this method, issues we address in Section 4.3.

Figures 2 and 3 show results for this setup. Figure 2 presents mean surprisal decreases relative to score 0 entities. All salience levels show negative values, indicating that salient entities reduce sentence surprisal when used as prompts. The pattern is nearly monotonic: higher salience corresponds to lower average z-scored surprisal, with sentences following non-salient entity prompts remaining close to baseline (0.0171 in Experiment 2; 0.0369 in Experiment 3), while those following highly salient entities (score = 5) show substantially lower values (-0.3481 in Experiment 2; -0.4710 in Experiment 3).

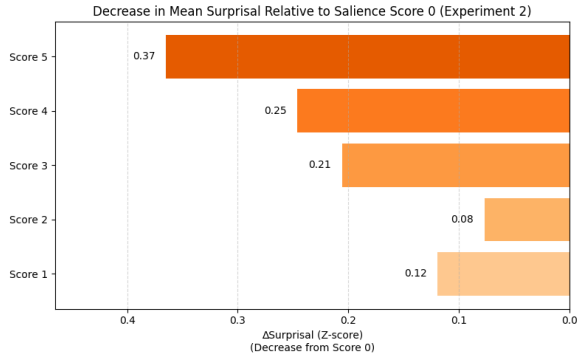


Figure 2: Decrease in mean surprisal relative to salience score 0 in Experiment 2. Magnitude of surprisal reduction largely grows with salience, indicating that more salient entities facilitate sentence-level predictability.

Figure 3 shows consistent cross-genre patterns: salient entities yield lower mean surprisal than non-salient ones across all 16 genres (orange consistently below green). Maximum differences appear in formal interviews, scholarly writing (`academic` and `textbook`), informative texts (`voyage`, i.e. travel guides from wikivoyage, and how-to guides from `wikihow`), with `speech` and `vlogs` much less pronounced. This universality across diverse discourse types suggests global salience robustly shapes predictability regardless of modality or register.

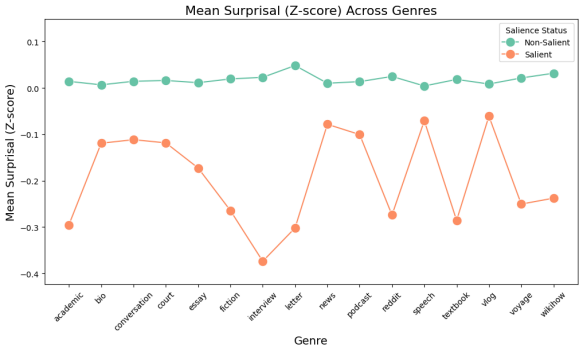


Figure 3: Mean surprisal scores across genres for Experiment 2 (salience score>0 in orange, score=0 in green).

However, entity length and nesting remain confounds. Function word sequences like "of the" within multi-word mentions are highly predictable, potentially inflating the apparent predictability of longer entities. Additionally, non-salient entities can nest salient ones, diluting distinctions. Experiment 3 addresses these issues through head-noun isolation, at the cost of less accurate representations of entity phrases.

4.3 Paired Head Noun Design

To eliminate length and nesting confounds, Experiment 3 extracts syntactic head nouns from entities' first non-pronominal mentions using GUM's gold standard UD syntax tree annotations. The head token for each entity mention is assumed to be the token whose dependency parent token is outside of the span of the mention. Representing mentions using their head tokens eliminates common sub-sequences (e.g.,"of the"), standardizing prompt length to single tokens while removing predictable function words.[3] We apply the same minimal-pair design, measuring sentence surprisal following head-noun prompts.

Figure 4 shows stronger, steeper surprisal reduction as salience increases compared to Experiment 2. The minor misordering of scores 1 and 2 persists but is negligible. Genre patterns remain consistent (Figure 5), with `interviews` leading, followed by `fiction`, `speeches`, and courtroom transcripts. `Reddit` and `vlogs` are borderline.

[3] WordPiece tokenization may still create multiple tokens, but we expect impacts to be modest compared to phrase-level effects.

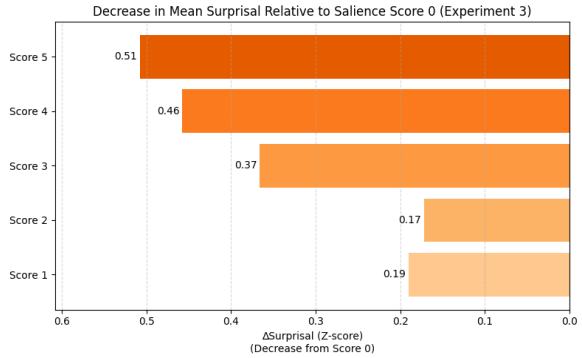


Figure 4: Decrease in mean surprisal relative to salience score 0 in Experiment 3.

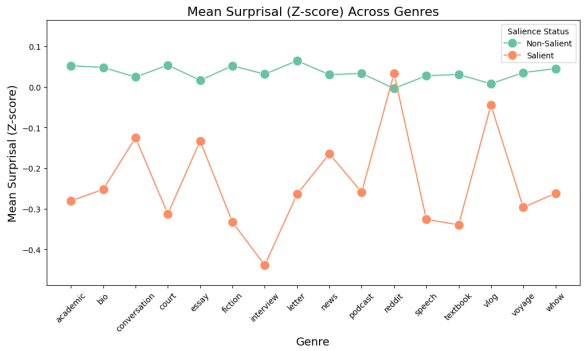


Figure 5: Mean surprisal scores across genres for Experiment 3 (salience > 0 in orange, salience=0 in green).

Compared to Experiment 2, this stricter design controls for both length and nesting. While head-noun reduction sacrifices semantic completeness, the persistence of robust salience-predictability relationships across both experiments demonstrates that global salience does interact with discourse predictability. These effects do not merely correlate with surface features like length or syntactic complexity, the potential confounds mentioned above. Instead, they reveal a discourse-structural pressure: globally salient entities increase expectations for upcoming content, yielding localized troughs in surprisal in the sentences that follow. These effects align with a view of UID in which systematic non-uniformities arise at specific points in discourse due to identifiable pressures analogous to phonotactic and syntactic pressures.

To verify that these patterns are not model-specific, we replicate all experiments using GPT2-SMALL (124M parameters), seeing similar trends; full figures and analyses are provided in Appendix B.

5 Genre Analysis

The results from Experiment 3 suggest that globally salient entities reduce surprisal, but whether this effect persists across genres remains an open question. If global salience functions as a discourse-structural pressure competing with UID, its strength could vary with discourse structure itself. Genres differ systematically in discourse structure, topic continuity, and interactivity (especially for multi-party conversations). Expository writing (academic articles, textbooks) maintains tight topic focus and hierarchical organization, while conversational genres show frequent topic shifts (multi-party dialogue, forum discussions), lowering predictability. We therefore hypothesize that the salience effect should be strongest in topic-coherent genres and weakest in topic-shifting ones.

5.1 Quantitative analysis

Figure 6 shows mean surprisal z-scores for salient (orange) and non-salient (blue) entities across 16 genres using data from Experiment 3, sorted by surprisal difference. Error bars show 95% confidence intervals adjusted for multiple testing, with asterisks for significant differences ($p < .05$, $p < .01$, $p < .001$), and non-significant genres shaded gray.

In all genres but `reddit`, salient entities reduce surprisal more predictable than non-salient ones, and in most genres significantly so. This pattern is especially robust in `academic`, `fiction`, and `bio`, where differences are large and error bars do not overlap. These genres exhibit structured or goal-directed discourse supporting repeated, predictable use of salient referents (e.g. protagonists, research topics or methods).

In contrast, the surprisal gap is smaller and non-significant in `conversation`, `vlog`, and inverted for `reddit`. These informal, topic-shifting genres show greater surprisal variability and overlapping confidence intervals. Salience aligns less closely with predictability here, suggesting genre moderates the salience-predictability relationship. Results for `essay` and `podcast` are also not significant, though these have substantially less data in the corpus (see Table 1).

These patterns reveal how competing pressures interact. In some genres, UID pressure toward evenness is systematically counteracted by discourse-structural pressure favoring predictability zones around central entities. In other more topic-shifting genres, topic instability may prevent salient enti-

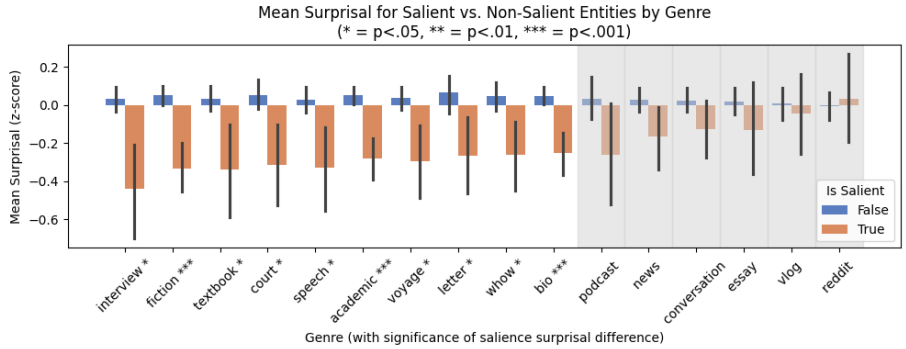


Figure 6: Mean surprisal (z-score) for salient vs. non-salient entities across genres. Bars show mean surprisal for salient (orange) and non-salient (blue) entities, with genres sorted by surprisal difference. Error bars represent 95% confidence intervals. Asterisks indicate significant differences based on adjusted t-tests (* $p < .05$, ** $p < .01$, *** $p < .001$). Genres shaded gray show no significant difference between salient and non-salient entities.

ties from creating sustained predictability, allowing UID pressure to dominate more. The balance between these pressures can vary with discourse context and coherence structure.

5.2 Qualitative Analysis

Figure 7 illustrates the patterns above through contrasting examples. In an academic article on discrimination (Figure 7a), salient entities like "discrimination" and "the United States" yield lower surprisal across sentences. These entities align with the discourse topic, constraining expectations and making upcoming content like "racial groups" or "contemporary America" more predictable than with non-salient "psychologists" or "Add Health" (the name of a study, which is less salient).

Conversely, in a conversation between "Sabrina" and her mom about staying somewhere overnight and other topics (Figure 7b), salient entities like "last night" or "Spend the night" do not reliably reduce surprisal across the document. The sentence "What'd you do, Sabrina?", with low semantic overlap with the rest of the text, yields nearly identical surprisal curves for non-salient entity prompts like "some other weekend". This reflects looser topic structure: frequent topic shifts mean even salient content offers little predictive help.

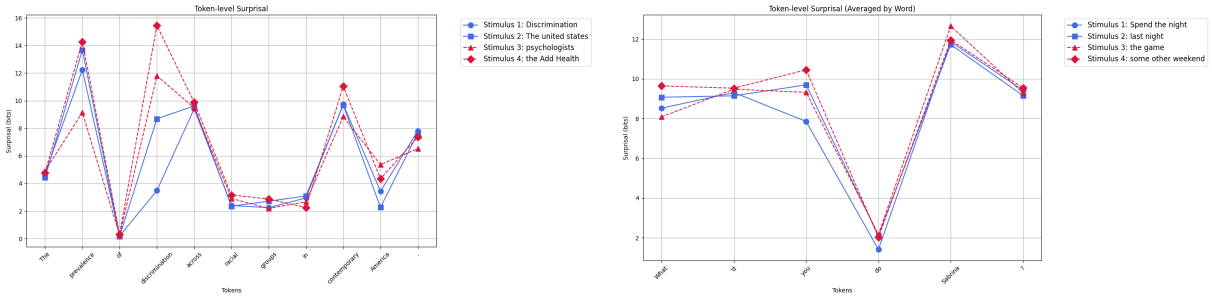
These examples highlight why predictive utility of salient entities varies by genre. In structured expository texts with strong coherence, global salience creates robust predictability zones (systematic departures from UID driven by referential structure). In conversational settings with weak coherence, lexical cohesion is lower and the facilitatory effect diminishes, allowing information to distribute more evenly as UID would predict.

6 Conclusion

We examined whether global discourse salience, operationalized as summary-worthiness, functions as a systematic pressure on information distribution in discourse. Three experiments reveal nuanced interactions between salience and surprisal. Experiment 1 showed that globally salient entities exhibit significantly higher surprisal than non-salient ones in natural contexts. A linear model revealed that while multiple pressures (position, mention length, nesting) operate simultaneously, salience remains a significant predictor of surprisal even alongside these confounding factors. Experiments 2 and 3 used a novel minimal-pair paradigm to isolate salience effects by controlling confounds. Results show that globally salient entities systematically reduce surprisal for surrounding discourse content when used as prompts. This effect is robust across most genres but varies predictably with discourse structure: stronger in topic-coherent texts, weaker in topic-shifting conversational contexts.

These findings establish that global salience introduces localized non-uniformities in discourse. Like syntactic and phonotactic pressures identified in prior work (Clark et al., 2023; Pimentel et al., 2021), the discourse prominence of entities competes with the tendency toward uniformity, with salient entities enhancing predictability of surrounding content. This effect is sensitive to discourse structure: salience-driven predictability is stronger in topic-coherent genres than in conversational ones where topic shifts dominate.

For UID theory, our results show where and why non-uniformities arise: salient entities increase expectations for subsequent discourse, producing



(a) Surprisal contour for a minimal pair in an academic excerpt.

(b) Surprisal contour for a minimal pair in a conversation.

Figure 7: Surprisal contour in different genres. Blue lines represent the surprisal contour of the sentence followed by salient entities; Red lines represent the surprisal contour of the sentence followed by non salient entities. Each stimulus is distinguished by different markers.

localized reductions in surprisal, while leaving document-level averages unchanged and preserving overall information density. For discourse theory, these findings validate that summary-worthiness captures document-central entities from an information theoretic perspective: entities independently selected for summaries constrain discourse-wide predictability. This relationship between summarization and predictability opens questions about how discourse centrality shapes both what speakers express and how listeners process information across extended discourse.

Although this paper focused exclusively on entity mentions, future work may be able to tell us more about the ways in which surprisal relates to salience at large, and current work on annotating summary-based salience for propositions (Zeldes et al., 2026) should allow for follow up studies using methods analogous to the ones proposed in this study. A further avenue of possible research relates to the extent to which our findings in textual data could have parallels in multimodal inputs, where work on UID is now beginning to disentangle pressures on uniformity when language is grounded in visual perception (Gay et al., 2026). We leave these areas to future research and hope that our data and methodology will be fruitful for these and other research questions.

Limitations

This study is limited to English, a high-resource language with abundant training data for the language model used to compute surprisal. It remains an open question whether the observed patterns involving salience and surprisal hold in lower-resource languages, where language models may have weaker estimates of discourse structure and referential continuity, and morphosyntactic constraints differ. In addition, while we operationalized predictability using token-level surprisal estimates, other measures such as entropy or mutual information may capture different dimensions of uncertainty in discourse processing. Future work could explore how these alternative metrics relate to referential salience across languages and model architectures. The operationalization of salience using summaries is also vulnerable to subjectivity in the summarization process itself, though we expect that the use of multiple summaries helps to mitigate the effects of any single outlier summaries. Additionally, we do not separately analyze how global salience interacts with other referential factors such as definiteness, NP type, or grammatical role. While these factors influence local mention-level predictability (as discussed in Section 2.2), our focus is on document-level salience effects that operate across all mention types. Future work should examine how global and local salience factors interact to shape information distribution at both mention and discourse levels. Finally, although we employed standard practices from previous studies to conform to and ensure comparability with previous work, we acknowledge that different tokenization strategies, especially in terms of word-piece tokenization, may also influence the results, a limitation we leave for future studies on surprisal using LM probability estimates in general.

References

Jennifer E Arnold. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4(4):187–203.

Jennifer E Arnold. 2025. Why does recency guide pronoun comprehension? It's not just topicality, atten-

tion, or predictability. *Discourse Processes*, pages 1–23.

Vincent Boswijk and Matt Coler. 2020. What is salience? *Open Linguistics*, 6(1):713–722.

Wallace Chafe. 1994. *Discourse, Consciousness, and Time. The Flow and Displacementof Conscious Experiencein Speaking and Writing*. University of Chicago Press, Chicago & London.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Thomas Hikaru Clark, Ethan Gotlieb Wilcox, Edward Gibson, and Roger P. Levy. 2022. Evidence for availability effects on speaker choice in the Russian comparative alternation. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, pages 3044–3050.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomáš Vitvar, and Harald Sack. 2016. Crowdsourced corpus with entity salience annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3307–3311.

Jesse Dunietz and Dan Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209.

August Fenk and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis – Konstanz im sprachlichen Informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013. Identifying salient entities in web pages. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2375–2380.

Matteo Gay, Coleman Haley, Mario Giulianelli, and Edoardo Ponti. 2026. Is information density uniform when utterances are grounded on perception and discourse? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3825–3853, Rabat, Morocco. Association for Computational Linguistics.

T. Givón. 1983. *Topic Continuity in Discourse*. John Benjamins.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

T. Florian Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for wikification in 12 genres. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2025. GUM-SAGE: A novel dataset and approach for graded entity salience prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 438–455, Vienna, Austria. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. GUMSum: Multi-genre data and evaluation for English abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9315–9327, Toronto, Canada. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

Hannah Rohde and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Péter Rácz. 2013. *Salience in Sociolinguistics. A Quantitative Approach*. De Gruyter Mouton, Berlin, Boston.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

Klaus von Heusinger and Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127.

Chuan Wu, Evangelos Kanoulas, Maarten de Rijke, and Wei Lu. 2020. WN-Salience: A corpus of news articles with entity salience annotations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2095–2102.

Alessandra Zarcone, Marten Van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, 7(844).

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Amir Zeldes, Katherine Conhaim, and Lauren Levine. 2026. Not worth mentioning? A pilot study on salient proposition annotation. ArXiv preprint 2603.27358.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A Experimental Setup and Hyperparameters

Model. We computed token-level surprisal using DISTILGPT-2, a distilled version of GPT-2 small. DistilGPT-2 contains 82M parameters and has been shown to preserve the core distributional properties of its larger counterpart (Sanh et al., 2019) while being computationally efficient.

Implementation. We used the Hugging Face `Transformers` library. Documents were processed with a sliding window of 1,024 tokens to maintain consistent context length. For each token, surprisal was calculated as the negative log-probability:

$$\text{surprisal}(w_i) = -\log P(w_i|w_1, \ldots, w_{i-1}).$$

The model was run in fp32 precision with all parameters at their default pretrained values without fine-tuning.

Computation. Inference was performed on a single NVIDIA A100 GPU with batch size of 16. Tokenization used automatic padding and truncation to handle variable-length inputs within each batch.

Aggregation. Token-level surprisals were averaged to obtain mention-level scores, and mention-level scores were averaged to obtain entity-level scores, as described in Section 3.2.

B Replication with GPT2-Small

To assess whether the observed effects in Section 4 depend on the choice of language model, we replicate all experiments using GPT2-SMALL (124M parameters). The results closely mirror those obtained with DISTILGPT-2 (82M parameters). In Experiment 1 (Figure 8), salient entities again exhibit slightly higher surprisal than non-salient ones. In Experiments 2 and 3 (Figures 9- 10), we observe the same monotonic relationship between salience and surprisal reduction, with higher salience levels consistently yielding larger decreases (e.g., up to 0.66 in Experiment 3). The relative ordering across salience levels and the overall effect sizes remain comparable, indicating that our findings are not artifacts of a specific model architecture.

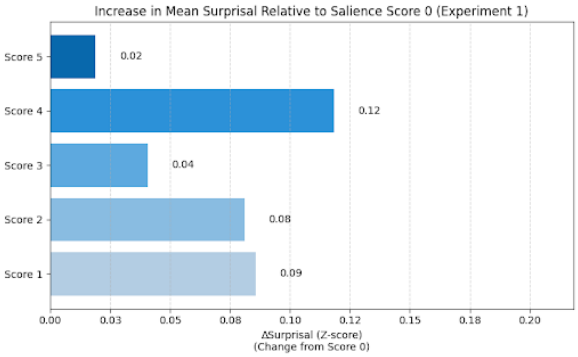


Figure 8: Change in mean surprisal for salience scores 1–5 relative to score 0 in Experiment 1 using GPT2-SMALL.

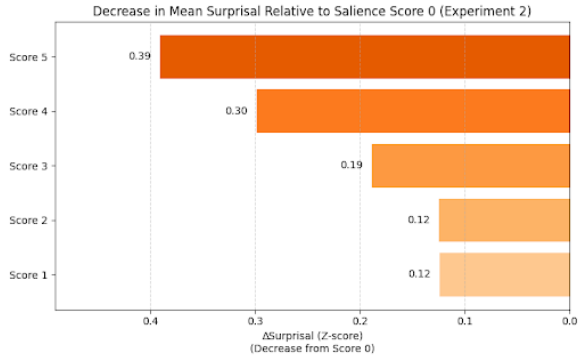


Figure 9: Change in mean surprisal for salience scores 1–5 relative to score 0 in Experiment 2 using GPT2-SMALL.

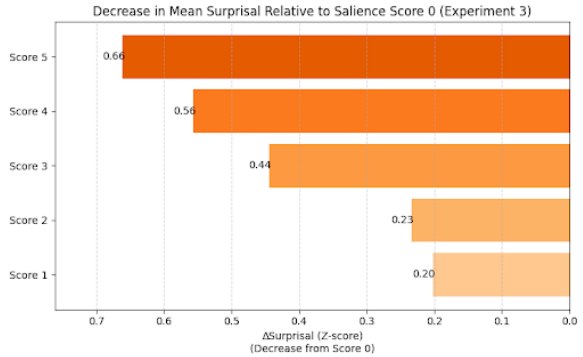


Figure 10: Change in mean surprisal for salience scores 1–5 relative to score 0 in Experiment 3 using GPT2-SMALL.