Surprisal Minimisation over Goal-directed Alternatives Predicts Production Choice in Dialogue

Tom Utting[◇] **Mario Giulianelli**[◁] **Arabella Sinclair**[◇◁]
[◇]University of Aberdeen [◁]University College London
m.giulianelli@ucl.ac.uk arabella.sinclair@ucl.ac.uk

Abstract

We model utterance production as probabilistic cost-sensitive choice over contextual alternatives, using information-theoretic notions of cost. We distinguish between *goal-directed* alternatives that realise a fixed communicative intent and *goal-agnostic* alternatives defined only by contextual plausibility, allowing us to derive speaker- and listener-oriented interpretations of different cost measures. We present a procedure to generate both types of alternative sets using language models. Analysing production choices in open-ended dialogue under both deterministic and probabilistic cost minimisation, we find that surprisal minimisation relative to goal-directed alternatives provides the strongest predictive account under both analyses. By contrast, uniform information density and length-based costs exhibit weaker and less consistent predictive power across conditions. More broadly, our study suggests that alternative-conditioned optimisation with LM-generated alternatives provides a principled framework for studying speaker and listener pressures in naturalistic language production.[1]

1 Introduction

Information-theoretic and probabilistic-pragmatic models of communication provide a general framework for reasoning about utterance choice. They construe speakers as approximately rational agents that operate under resource constraints and trade off production effort against listener comprehension effort, while maintaining communicative effectiveness (Levy and Jaeger, 2006; Frank and Goodman, 2012; Franke, 2014; Giulianelli, 2022; Degen, 2023). This framing leaves open two modelling questions: (a) how production and comprehension costs should be defined, distinguished, and weighted against one another, and (b) over what set of alternative production choices this optimisation is assumed to take place.

This paper provides a principled way to formalise and compare speaker and listener costs in utterance production by making the space of alternatives explicit. Our central claim is that the interpretation of a given cost function—such as surprisal or information uniformity—depends on the set of alternative utterances with respect to which it is evaluated. When costs are evaluated relative to a goal-directed set of alternatives that all realise the same fixed communicative goal intended by the speaker, the cost function can be interpreted as a measure of speaker cost; correspondingly cost sensitivity yields a speaker-oriented explanation of utterance choice. Conversely, when costs are evaluated over a goal-agnostic set of alternatives defined solely by the shared contextual state—including conversational history, common ground, and the immediate sentential context—they give rise to a listener-oriented notion of cost sensitivity.

In our experiments, we operationalise this distinction by using large language models to generate goal-directed and goal-agnostic alternative utterances. As a case study, we identify critical choice points in production data from a conversational dialogue corpus and compute the cost of both the observed utterances and their alternatives. We evaluate multiple cost measures—surprisal, local and global information uniformity, and length—and assess which notion of cost minimisation best accounts for speakers' observed choices.

We find that probabilistic minimisation of surprisal relative to goal-directed alternatives provides the strongest predictive account of human production choices, supporting a speaker-oriented interpretation of surprisal cost sensitivity (e.g., Goodman and Lassiter, 2015; Futrell, 2024). These findings provide new evidence that surprisal functions as a production-side constraint in speaker decision-making, and they open avenues for studying production preferences in naturalistic language use using information-theoretic cost measures.

[1]Code and data can be found at https://github.com/the-context-lab/productionchoice

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39090–39109

2 Background

Speakers are thought to balance their own production costs and listeners' comprehension costs when selecting among contextually available alternatives. This section reviews prominent models of cost, with a focus on information-theoretic accounts, and examines how alternative utterances are typically handled—often only implicitly.

2.1 Production and Comprehension Costs

Production costs are the cognitive and temporal resources expended by speakers in formulating and realising an utterance. They arise from processes such as memory retrieval, advance planning of the utterance, and the impact of time pressure on formulation processes (Bard et al., 2007; Howarth and Anderson, 2007; Ivanova and Ferreira, 2019; Betz et al., 2023). Comprehension costs, by contrast, reflect the cognitive and temporal effort incurred by listeners when processing and interpreting an utterance, including efforts involved in predicting upcoming words, maintaining and updating representations in working memory, resolving meaning and references, and handling rapid turn-taking (Hadar et al., 2016; Peelle, 2017; Meyer, 2023).

Within information-theoretic models of cost, the distinction between these two categories is often blurred. Measures such as surprisal, entropy, and uniform information density, are widely used to explain speakers' production choices (Genzel and Charniak, 2002; Xu and Reitter, 2018; Giulianelli and Fernández, 2021; Giulianelli et al., 2021; Gay et al., 2026), under the idea that they capture *some* notion of processing cost. However, it is often unclear whether these measures should be interpreted as proxies for comprehension cost, production cost, or a combination of the two. Surprisal, for example, is typically motivated as a measure of comprehension difficulty (Hale, 2001; Levy, 2008), and has been shown to predict behavioural and neural indices of listener cost in eye-tracking, self-paced reading, brain imaging, and priming studies (Keller, 2004; Smith and Levy, 2013; Sinclair et al., 2022; Wilcox et al., 2023; Jumelet et al., 2024; Sinclair et al., 2026). At the same time, surprisal estimates are usually derived from language models trained to approximate *speaker* behaviour, and are sometimes interpreted as reflecting speaker costs (Goodman and Lassiter, 2015; Giulianelli et al., 2022; Yee et al., 2024; Futrell, 2024).

A similar ambiguity characterises Uniform Information Density (UID) accounts of language production. UID accounts posit that more uniform distributions of information reduce processing difficulty (Fenk and Fenk, 1980; Genzel and Charniak, 2002; Aylett and Turk, 2004; Levy and Jaeger, 2006), yet it is often left open whether this reduction should be attributed to speaker effort, listener effort, or properties of the communicative signal itself. Consequently, UID has been invoked under both speaker- and listener-oriented interpretations (Coupé et al., 2019; Meister et al., 2021; Pimentel et al., 2021). This ambiguity leaves unclear how observed linguistic behaviour should be attributed to production versus comprehension costs. In our work, we address this by modelling cost sensitivity explicitly as either speaker- or listener-oriented, evaluating utterances relative to goal-directed or goal-agnostic alternatives while preserving standard information-theoretic cost measures.

2.2 Alternatives in Models of Production

Computational models of language production differ substantially in the assumptions they make about the set of alternative realisations over which production and comprehension costs are evaluated. Many information-theoretic approaches abstract away from alternatives altogether, analysing information-theoretic properties of observed utterances or discourse without specifying the competing continuations available to the speaker (Genzel and Charniak, 2003; Giulianelli and Fernández, 2021; Giulianelli et al., 2021; Tsipidi et al., 2024, 2025). Other approaches, including classic Uniform Information Density accounts, implicitly assume competition among a small number of paraphrastic alternatives, such as syntactic variants that differ in how evenly information is distributed across an utterance (Levy and Jaeger, 2006; Jaeger, 2010). On the other hand, work in probabilistic pragmatics—including Rational Speech Act (RSA) models and rate–distortion approaches—typically studies optimisation over explicitly defined, highly restricted sets of alternative actions (Franke, 2014; Goodman and Lassiter, 2015; Futrell, 2023, 2024), partly due to practical constraints on enumerating alternatives. More recently, language models have been used to generate rich sets of contextual alternatives for investigating language comprehension and production (Hu et al., 2022, 2023; Giulianelli et al., 2023a,b, 2024, 2026; Meister et al., 2024). This development makes it possible to operationalise probabilistic pragmatic models of production as choice

models over open-ended alternative spaces. In this paper, we present methods for using language models to construct goal-directed and goal-agnostic contextual alternatives, i.e., alternatives available to the speaker and the listener, respectively.

3 Production as Cost-sensitive Choice over Contextual Alternatives

We introduce a formalisation of the language production process that enables us to distinguish between speaker- and listener-oriented costs. Our focus is on how speakers choose among alternative continuations at a given point in an utterance, but the same formalisation applies more generally to production across different choice points and granularities, including choices between words, phrases, or clauses, as well as choices spanning sentence boundaries, given appropriate cost functions.

3.1 Contextual Alternatives

Consider a classic example from Jaeger (2010), which illustrates the production choice involved in realising an English complement clause:

(1) My boss confirmed that we were absolutely crazy.

Let Σ be a non-empty set of linguistic units (typically, though not necessarily, words) and Σ^* the set of strings formed from units in Σ. We treat an utterance as a string[2] and decompose it into two substrings: the **context** $\boldsymbol{c} \in \Sigma^*$, marked in blue, and the target **continuation** $\boldsymbol{a}^\star \in \Sigma^*$, shown in green. In this example, the continuation is a complement clause, and the matrix verb *confirmed* identifies the onset of the complement clause as a decision point. We refer to such positions as **choice points**: points in the production process at which the speaker selects a continuation from a set of possible contextual alternatives.

A perfectly grammatical and communicatively equivalent alternative continuation omits the complementiser *that*:

(2) My boss confirmed we were absolutely crazy.

Choosing this continuation $\boldsymbol{a}' \in \Sigma^*$ (marked in red) over $\boldsymbol{a}^\star$ constitutes a case of syntactic reduction, which leads to less uniform information distribution across the sentence and is therefore predicted to be dispreferred under UID accounts of production (Levy and Jaeger, 2006; Jaeger, 2010).

The reduced and unreduced complement clauses are just two members of a much larger set of continuations that are grammatically and semantically licensed by the context, even if they differ in structure, meaning, or communicative goal. To characterise production choices more generally, we therefore need to consider the full space of continuations available at a given choice point.

Goal-agnostic alternatives. At a given choice point defined by context $\boldsymbol{c}$, the speaker is in principle free to continue the utterance in many different ways. We define the **goal-agnostic alternative set** $\mathcal{A}_{\boldsymbol{c}}$ as the set of all continuations that are grammatically licensed and contextually coherent given $\boldsymbol{c}$, independently of the speaker's intended communicative goal. In the example above, this includes continuations such as:

$\boldsymbol{a}_1$ = we were absolutely crazy.
$\boldsymbol{a}_2$ = that the meeting has been rescheduled.
$\boldsymbol{a}_3$ = my request for time off next week.
$\boldsymbol{a}_4$ = our participation in the next conference.
$\boldsymbol{a}_5$ = that I have her full support.

We call this alternative set goal-agnostic in that it includes continuations that would *not* result in a sentence communicatively equivalent to $\boldsymbol{c}\boldsymbol{a}^\star$ (marked in gold). Formally, we assume that the goal-agnostic alternative set $\mathcal{A}_{\boldsymbol{c}}$ is obtained by sampling N continuations from a distribution over strings conditioned on the context, i.e., a language model:

$$\mathcal{A}_{\boldsymbol{c}} \overset{\text{def}}{=} \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N\}, \quad \boldsymbol{a}_i \sim p(\cdot \mid \boldsymbol{c}). \quad (1)$$

This alternative set reflects contextually constrained uncertainty both over which communicative goal the speaker intends to realise and over how the goal may be linguistically realised. Accordingly, *we assume that this set approximates the expectations of a* ***listener***, who is uncertain about which goal the speaker will communicate.[3]

[2] A string, written in boldface, is a finite sequence of units $\boldsymbol{w} = w_1 \ldots w_n$, where units are written in normal font. The length of a string is $|\boldsymbol{w}| = n$. Concatenation of strings (and units) is denoted by juxtaposition, e.g. $\boldsymbol{w}\boldsymbol{w}'$.

[3] Note that the shared contextual state (including conversational history, common ground, and immediate sentential context) can constrain the set of plausible goals. We therefore do not assume that the listener assigns uniform probability across all conceivable goals, but rather that they maintain uncertainty over a contextually restricted set. Accordingly, some goal-agnostic alternatives may align with the speaker's intended goal, as is the case for $\boldsymbol{a}_1$ in the example above. Empirical evidence supporting this assumption is provided in App. D.2.

Goal-directed alternatives. By contrast, we define the **goal-directed alternative set** $\mathcal{A}_{\boldsymbol{c},g}$ as the set of continuations that realise a fixed communicative goal g in context $\boldsymbol{c}$. Formally, we assume that this set is obtained by sampling continuations from a distribution over strings conditioned on both the context and the communicative goal:

$$\mathcal{A}_{\boldsymbol{c},g} \stackrel{\text{def}}{=} \{\boldsymbol{a}_1^g, \ldots, \boldsymbol{a}_N^g\}, \quad \boldsymbol{a}_i^g \sim p(\cdot \mid \boldsymbol{c}, g). \quad (2)$$

Here, p is not a standard language model; we show how to approximate a goal-conditioned language model in §5.3. In our running example, g can be characterised informally as conveying that the speaker's judgement or behaviour was seriously mistaken. Examples of alternatives in $\mathcal{A}_{\boldsymbol{c},g}$ include:

$\boldsymbol{a}_1^g$ = we were absolutely crazy.
$\boldsymbol{a}_2^g$ = that we were completely irrational.
$\boldsymbol{a}_3^g$ = that our decision made no sense at all.
$\boldsymbol{a}_4^g$ = that our behaviour was seriously flawed.
$\boldsymbol{a}_5^g$ = we were wildly off the mark.

These alternatives vary in syntactic realisation, lexical-semantic choice, stylistic register, and evaluative strength, but all preserve the same underlying communicative intent. *We assume that this set reflects the production uncertainty of a* ***speaker***, who has fixed a communicative goal and is choosing among alternative realisations of that goal.

3.2 A Choice Model of Production

Having defined the space of alternatives available at a choice point, we now turn to the question of how speakers choose among them. We adopt a probabilistic, decision-theoretic perspective on production in which, in line with decision-theoretic and RSA approaches, speakers assign probability to alternative continuations in proportion to their utility.

Formally, let $\boldsymbol{a}$ denote a candidate continuation in context $\boldsymbol{c}$ given a fixed communicative goal g. The probability of the speaker producing $\boldsymbol{a}$ is:

$$P_S(\boldsymbol{a} \mid \boldsymbol{c}, g) \stackrel{\text{def}}{\propto} \exp\big(\alpha\, U(\boldsymbol{a}; \boldsymbol{c}, g)\big), \quad (3)$$

where $U(\boldsymbol{a}; \boldsymbol{c}, g)$ is a utility function and $\alpha \geq 0$ is a sensitivity parameter controlling the extent to which the speaker behaves as a utility-maximising agent (Luce, 1959).[4] As α increases, the distribution becomes increasingly peaked around higher-utility utterances, converging to deterministic utility maximisation in the limit $\alpha \to \infty$; conversely, when $\alpha = 0$, choice is uniform over alternatives.

[4]The parameter α is also commonly called an inverse-temperature or rationality parameter.

Following standard RSA formulations, we decompose utility into an effectiveness term and a cost term:

$$U(\boldsymbol{a}; \boldsymbol{c}, g) \stackrel{\text{def}}{=} E(\boldsymbol{a}, g; \boldsymbol{c}) - C(\boldsymbol{a}; \boldsymbol{c}), \quad (4)$$

where $E(\boldsymbol{a}, g; \boldsymbol{c})$ denotes the **communicative effectiveness** of utterance $\boldsymbol{a}$ for achieving goal g in context $\boldsymbol{c}$, and $C(\boldsymbol{a}; \boldsymbol{c})$ denotes the **production cost** associated with $\boldsymbol{a}$ in context $\boldsymbol{c}$. In much pragmatic work, production cost is held constant in order to isolate the role of communicative effectiveness. Here, we take the complementary perspective and isolate the role of production cost in shaping utterance choice. For goal-directed alternatives, this amounts to abstracting away from differences in effectiveness that are, by construction, absent, as all goal-directed continuations realise the communicative goal by definition. For goal-agnostic alternatives, effectiveness varies from the speaker's perspective but is uncertain from the listener's perspective, which the goal-agnostic set is intended to approximate (see §3.1). In both cases, we therefore treat effectiveness as constant across alternatives.

Formally, if communicative effectiveness is held constant across alternatives—i.e., $E(\boldsymbol{a}, g; \boldsymbol{c}) = \kappa$ for all $\boldsymbol{a}$ in the relevant alternative set—then the speaker's probabilistic production rule reduces to a softmax over cost:

$$P_S(\boldsymbol{a} \mid \boldsymbol{c}, g) \stackrel{\text{def}}{=} \frac{\exp\big(\alpha(\kappa - C(\boldsymbol{a}; \boldsymbol{c}))\big)}{\sum_{\boldsymbol{a}' \in \mathcal{A}} \exp\big(\alpha(\kappa - C(\boldsymbol{a}'; \boldsymbol{c}))\big)} = \frac{\exp\big(-\alpha C(\boldsymbol{a}; \boldsymbol{c})\big)}{\sum_{\boldsymbol{a}' \in \mathcal{A}} \exp\big(-\alpha C(\boldsymbol{a}'; \boldsymbol{c})\big)}. \quad (5)$$

Lower-cost alternatives are assigned higher production probability, with the strength of this preference controlled by the cost-sensitivity parameter α. Assuming a finite alternative set and a real-valued cost function for which a minimum exists, this probabilistic choice rule converges, as $\alpha \to \infty$, to deterministic cost minimisation over the alternative set:

$$\boldsymbol{a}^\star = \arg\min_{\boldsymbol{a} \in \mathcal{A}} C(\boldsymbol{a}; \boldsymbol{c}). \quad (6)$$

Crucially, the interpretation of this minimisation depends on the choice of alternative set $\mathcal{A}$. When $\mathcal{A} = \mathcal{A}_{\boldsymbol{c},g}$, minimisation is speaker-oriented, as the alternatives differ only in how a fixed communicative goal is realised. When $\mathcal{A} = \mathcal{A}_{\boldsymbol{c}}$, minimisation is listener-oriented, as the alternatives encode contextually constrained uncertainty about both the speaker's intended goal and its

realisation. Although the speaker makes the choice in both cases, minimisation over the goal-agnostic alternative set is listener-oriented in that it reflects the speaker's model of the listener's expectations when selecting among continuations.

4 Measures of Cost

We consider four measures of utterance cost that have featured prominently in information-theoretic and probabilistic-pragmatic accounts of language production and comprehension.

4.1 Surprisal

Surprisal quantifies how unexpected a word or sequence is given its preceding context. For an alternative $\boldsymbol{a}$ in context $\boldsymbol{c}$, surprisal is defined as the negative log probability that a language model p assigns to $\boldsymbol{a}$ given $\boldsymbol{c}$, yielding the cost:

$$C_{\text{surp}}(\boldsymbol{a}; \boldsymbol{c}) \stackrel{\text{def}}{=} -\log p(\boldsymbol{a} \mid \boldsymbol{c}). \tag{7}$$

Within surprisal-based theories of comprehension, higher surprisal corresponds to a larger update of the comprehender's probabilistic expectations over upcoming linguistic material and thus to greater processing difficulty for the listener (Hale, 2001; Levy, 2008). Speakers are therefore predicted to prefer utterances with lower surprisal.

At the same time, surprisal has also been interpreted as a speaker cost, for example in Rational Speech Act and rate–distortion models (Goodman and Lassiter, 2015; Futrell, 2024). Under the rate–distortion theory of control, surprisal can be conceptualised as reflecting an "automatic policy" (Futrell, 2024). Highly frequent sequences correspond to well-practised production routines that are executed relatively automatically and are therefore less costly to produce, whereas contextually unlikely continuations require suppression of this automatic policy (i.e., greater control) and are therefore more effortful to produce.

Moreover, surprisal estimates are most commonly derived from language models trained on corpora of production data. In this sense, notwithstanding their effectiveness at predicting behavioural signatures of comprehension effort, they are obtained from models that are directly optimised to approximate speaker behaviour.

4.2 Uniform Information Density

A related but distinct proposal is the Uniform Information Density (UID) hypothesis, according to which speakers aim to distribute information as evenly as possible across an utterance (Fenk and Fenk, 1980; Aylett and Turk, 2004; Levy and Jaeger, 2006). Avoiding sharp peaks in surprisal is argued to facilitate comprehension and to constitute a listener-oriented rational strategy given grammatical constraints. Following prior work (Collins, 2014; Jain et al., 2018; Meister et al., 2021), we operationalise UID using two metrics: local and global uniformity.

Local Uniformity. We first consider *local* uniformity, which quantifies the smoothness of the surprisal contour across adjacent units within a sequence. Given a sequence of word-level surprisals $\mathbf{s} = (s_1, \ldots, s_n)$, local UID is defined as

$$\text{UID}_{\text{loc}}(\mathbf{s}) \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{t=2}^{n} \big(s_t - s_{t-1}\big)^2. \tag{8}$$

This formulation is sensitive to fine-grained grammatical and locality-driven effects, such as the placement of optional material or function words (Levy and Jaeger, 2006). In our experiments, we compute this metric only over the continuation, since surprisal values for the context are fixed across alternatives, except for a single negligible transition at the onset of the continuation. The corresponding cost for an alternative $\boldsymbol{a}$ in context $\boldsymbol{c}$ is

$$C_{\text{UID}_{\text{loc}}}(\boldsymbol{a}; \boldsymbol{c}) \stackrel{\text{def}}{=} \text{UID}_{\text{loc}}\big(\mathbf{s}(\boldsymbol{a})\big), \tag{9}$$

where $\mathbf{s}(\boldsymbol{a})$ denotes the sequence of word-level surprisals for the continuation. Lower values indicate more uniform and thus less costly surprisal profiles.

Global Uniformity. We next consider *global* uniformity at the level of an entire sequence (e.g., a sentence or discourse). Given a sequence of word-level surprisals $\mathbf{s}$, global UID is defined as

$$\text{UID}_{\text{gl}}(\mathbf{s}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} \big(s_t - \mu\big)^2, \tag{10}$$

where $\mu \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} s_t$ denotes the mean surprisal over the n words in the sequence. This metric quantifies the extent to which individual surprisal values deviate from the overall sequence average; lower values indicate more uniform surprisal profiles and thus lower cost. Global UID has been linked to the idea of maintaining a stable average information rate over longer stretches of discourse (Genzel and Charniak, 2002; Tsipidi et al., 2024). In our

experiments, we compute this measure over the full utterance, including both the context and the continuation, and define the corresponding cost for an alternative $\boldsymbol{a}$ in context $\boldsymbol{c}$ as

$$C_{\text{UID}_{\text{gl}}}(\boldsymbol{a}; \boldsymbol{c}) \stackrel{\text{def}}{=} \text{UID}_{\text{gl}}\big(\mathbf{s}(\boldsymbol{ca})\big), \qquad (11)$$

where $\mathbf{s}(\boldsymbol{ca})$ denotes the sequence of word-level surprisals for the concatenation of $\boldsymbol{c}$ and $\boldsymbol{a}$. As with local UID, global uniformity is primarily interpreted as a listener-oriented pressure.

Length. Finally, we consider continuation length as a simple yet widely used notion of cost, where length is measured as the number of words in the continuation:

$$C_{\text{len}}(\boldsymbol{a}; \boldsymbol{c}) \stackrel{\text{def}}{=} |\boldsymbol{a}|. \qquad (12)$$

Length is typically taken as a proxy for speaker effort (Bock and Levelt, 1994; Degen et al., 2013; Bergen et al., 2016; Cohn-Gordon et al., 2019; White et al., 2020; Giulianelli, 2022), but it may also reflect pressures arising from comprehension.

5 Methods

We draw on a dialogue corpus to extract human production choices (§5.1), estimate measures of cost using language models (§5.2), and generate alternative sets following the procedure in §5.3. Finally, we align the distributions of generated and human utterances via stratified sampling (§5.4).

5.1 Dialogue Contexts and Continuations

We use the Switchboard Dialogue Act Corpus (Stolcke et al., 2000), a naturalistic spoken conversational dialogue annotated with dialogue act labels. We parse and filter the data to remove backchannels and disfluencies, which are common in spoken dialogue but are largely out of distribution for LMs. Full preprocessing details are provided in App. A.1.

We restrict the data to utterances between 10 and 30 words in length that are annotated with *statement* or *question* dialogue act tags and are preceded by an utterance from the other speaker. These criteria ensure that the selected utterances are complete, coherent sentences with sufficient semantic structure to serve as ground-truth human productions in our experiments. This resulted in 1,342 utterances, which we select from in §5.3. Each utterance is parsed and divided into a context and a continuation. We identify the root verb of each sentence as the choice point and define the context $\boldsymbol{c}$ as all material up to and including the root verb. The remaining material, from the word after the root verb to the end of the utterance, is taken to be the human continuation $\boldsymbol{a}^{\star}$. This choice is motivated by Jaeger's classic that-mentioning example (see Section 3.1 and the Limitations section for a discussion). More details in App. A.2 to A.4.

5.2 Estimating Costs

Three of the four cost measures we consider are surprisal-based (cf. §4). We estimate surprisal using GPT-2 Small (Radford et al., 2019),[5] which has been shown to align well with behavioural measures of processing effort and is widely used in psycholinguistic research (Oh and Schuler, 2023; Shain et al., 2024; Kuribayashi et al., 2024). Surprisal of utterance contexts, as required for global UID, is computed conditional on the preceding dialogue history truncated to fit the model's context window. Surprisal of continuations is computed conditional on both the utterance context and this dialogue history.

5.3 Generating Alternatives

To obtain high-quality sets of alternatives for our experiments, we use an LLM as a simulator of dialogue utterance production. All generations were produced with OpenAI's GPT-4o, a state-of-the-art model at the time of experimentation.[6]

To generate **goal-agnostic** alternatives, we sample continuations from an LLM conditioned on different amounts of dialogue history. We consider three history conditions: one in which the model is instructed to complete the sentence given only the utterance context; one in which it is provided with a single preceding utterance; and one in which it has access to the entire conversational history. This reflects the fact that human speakers may vary in the extent to which they retain or rely on prior context when evaluating costs during production. In all cases, the model is prompted to complete the sentence context via a task instruction of the form *"Your task is to complete the provided sentence"*.

To generate **goal-directed** alternatives, we instruct the LLM to produce a fixed set of unique

[5] https://huggingface.co/openai-community/gpt2

[6] We used the latest GPT-4o snapshot available via the OpenAI API on 1 August 2025. All generations were produced with the default decoding settings, i.e., a temperature of 1 and no top-k or top-p truncation. All generated continuations are available at github.com/the-context-lab/productionchoice. Full details about the generation procedure as well as additional analyses can be found in App. B and D.

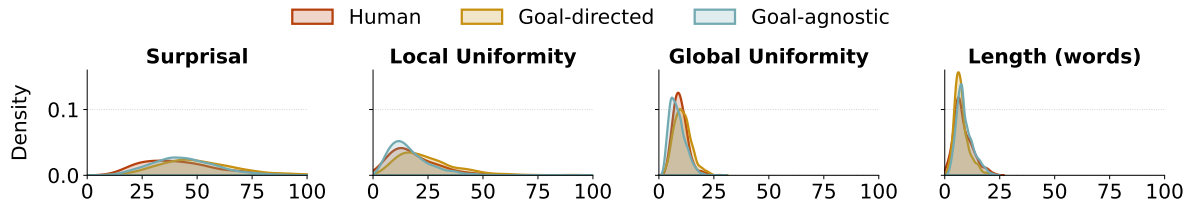

Figure 1: Global cost distribution for *human*, *goal-directed* and *goal-agnostic* alternative sets.

paraphrases of each observed human continuation, constrained to share the same initial context. We retain only those utterance contexts for which the model successfully generates at least 10 paraphrases. To ensure that goal-directed alternatives preserve the communicative intent of the observed utterance, we apply a post-hoc filtering and reclassification procedure to the entire dataset. We use an LLM-as-a-judge (Zheng et al., 2023) to classify whether alternative continuations are paraphrases of the human utterance; manual annotation of 400 sampled judgements yields an accuracy of 98.75%. Alternatives in the goal-directed sets that are not classified as paraphrases are discarded, while goal-agnostic alternatives that are classified as paraphrases of the human utterance are also treated as goal-directed, and thus belong to both sets. The proportion of these goal-matching goal-agnostic continuations is reported in Fig. 5. This procedure yields a dataset of 12,669 items (12,360 generated and 309 observed), spanning 309 contexts.

5.4 Aligning Human and Generated Distributions via Stratified Sampling

As outlined in §3.2, we model utterance choice as decision-making under probabilistic preferences within a given context. Under this view, for a cost measure to explain choice, human utterances are expected to exhibit lower cost than competing contextual alternatives. For this comparison to be meaningful, however, generated alternatives should not systematically differ from human utterances in their overall cost distributions, as such differences would otherwise confound context-specific effects.

We therefore assess whether, under each cost function, the distribution of generated continuations matches that of human utterances across contexts using independent-samples t-tests. We find no significant differences in surprisal or local uniformity, but observe small yet significant differences in length and global uniformity ($p < 0.001$).

To address this, we apply stratified sampling to the generated continuations, aligning their distributions with those of human utterances along the two affected dimensions: length and global uniformity. We first discretise human utterances into three bins per dimension and assign each utterance to a stratum defined by its $(C_{\text{len}}, C_{\text{UID}_{\text{gl}}})$ bin pair. The same bin boundaries are then applied to the generated continuations. We estimate the empirical distribution of human utterances over these strata and sample without replacement from the generated pool by selecting the largest feasible subsample whose stratum counts match the human proportions.

After stratification, differences in mean length and global UID are no longer statistically significant at the $\alpha = 0.001$ level. Figure 1 shows the resulting alignment between human and generated distributions. The final sample retains 6,335 generated utterances.[7] This adjustment ensures that subsequent analyses reflect context-specific preferences rather than global distributional differences.

6 Experiments and Results

This section evaluates whether human production choices reflect consistent preferences for continuations that have a lower cost than their contextual alternatives, under both goal-directed and goal-agnostic alternative sets. In §6.1, we analyse the rank of the observed continuation among its alternatives, testing whether it is the lowest-cost option more often than expected by chance. This corresponds to deterministic cost minimisation. In §6.2, we model production choice as probabilistic using a pairwise logistic model and directly compare the predictive strength of different cost measures.

6.1 Deterministic Cost Minimisation

To test the hypothesis that human continuations greedily minimise cost, we compute the rank of each observed continuation within both the goal-directed and goal-agnostic alternative sets. A continuation is assigned rank 1 if it minimises cost

[7] We obtain qualitatively identical results when using the full set of utterances; see App. D.6.

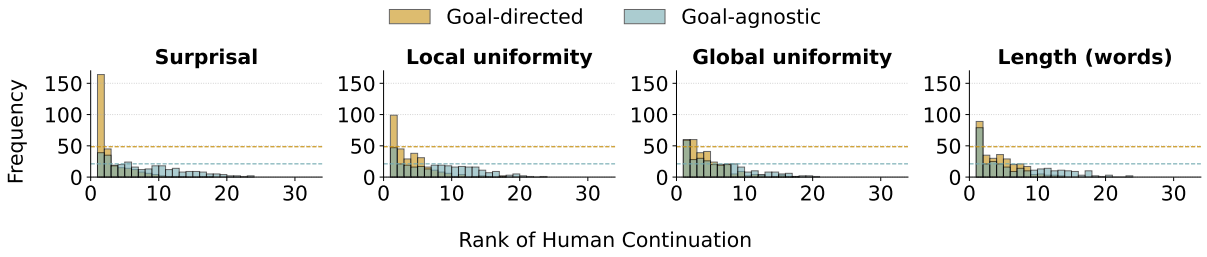

Figure 2: Ranking distributions of human continuations under different cost measures, evaluated against goal-directed and goal-agnostic alternative sets. A rank of 1 indicates that the human continuation has the lowest cost amongst available alternatives, corresponding to deterministic cost minimisation. Dashed lines indicate chance levels.

with respect to the alternative set (cf. Eq. 6). Fig. 2 shows the resulting rank distributions.

To assess whether rank-1 outcomes occur more frequently than expected by chance, we use a Poisson–binomial test. We model the rank of the human utterance on each trial as a Bernoulli outcome indicating whether the human utterance has rank 1, with trial-specific chance levels to account for differences in the size of the alternative sets. Statistical significance is assessed using a one-sided test. The full specification of the Poisson–binomial test is provided in App. C.1. Tab. 1 summarises the results, where the percentage of rank-1 outcomes quantifies the extent to which minimising a certain cost accounts for observed production choices.

Across all conditions, rank-1 outcomes occur significantly more often than expected by chance. The strongest absolute effect is observed for surprisal evaluated against goal-directed alternatives: 53.4% of human continuations minimise surprisal against this set, corresponding to a 3.24× increase over the baseline. When evaluated against goal-agnostic alternatives, the percentage drops to 15.2% (2.11×). This supports a speaker-oriented interpretation of surprisal-based cost minimisation.

Utterance length shows the largest relative increase over chance in the goal-agnostic setting (3.69×), exceeding all other cost measures. This suggests that length-based minimisation can be interpreted as a listener-oriented pressure to reduce processing effort. For local and global uniformity, rank-1 outcomes are less frequent overall but still reliably above chance. In both cases, the relative increase over chance is higher in the goal-agnostic setting, consistent with a listener-oriented interpretation in which information content reflects uncertainty over both upcoming linguistic material and the speaker's intended goal.

Overall, surprisal shows the strongest absolute effects, with over half of continuations minimising cost under goal-directed evaluation. Length shows the strongest relative effects under goal-agnostic evaluation, with uniformity falling in between. This analysis assumes that production selects the minimum-cost alternative. We next investigate whether production choices are better described as probabilistic preferences over alternatives.

Cost	Goal-directed	Goal-agnostic
Surprisal	53.4% ×3.24	15.2% ×2.11
Local uniformity	34.1% ×2.07	16.2% ×2.25
Global uniformity	24.1% ×1.46	19.3% ×2.68
Length (words)	28.6% ×1.73	26.6% ×3.69
Uniform *(baseline)*	16.5%	7.2%

Table 1: Percentage of observed rank-1 outcomes, with multiplicative increase relative to the uniform baseline (darker indicates larger deviation from the baseline). All values are significantly higher than chance under one-sided Poisson–binomial tests ($p < 10^{-5}$ or smaller).

6.2 Graded Cost Sensitivity

The rank-based analysis in §6.1 corresponds to a limiting case of the probabilistic choice model in which choice noise vanishes and the speaker deterministically selects the cost-minimising alternative (cf. §3.2). We now relax this assumption and instead model production choice as probabilistic using a using a pairwise logistic choice model.

We recast each observation as a binary comparison between the human continuation and a contextual alternative. The probability that a continuation $\boldsymbol{a}_i$ is preferred to an alternative $\boldsymbol{a}_j$ is modelled as a logistic function of their cost difference:

$$P(\boldsymbol{a}_i \succ \boldsymbol{a}_j; \boldsymbol{c}) = \sigma\big(\alpha(C(\boldsymbol{a}_j; \boldsymbol{c}) - C(\boldsymbol{a}_i; \boldsymbol{c}))\big) \tag{13}$$

This corresponds to a two-alternative reduction of the Luce-style choice rule introduced in §3.2.

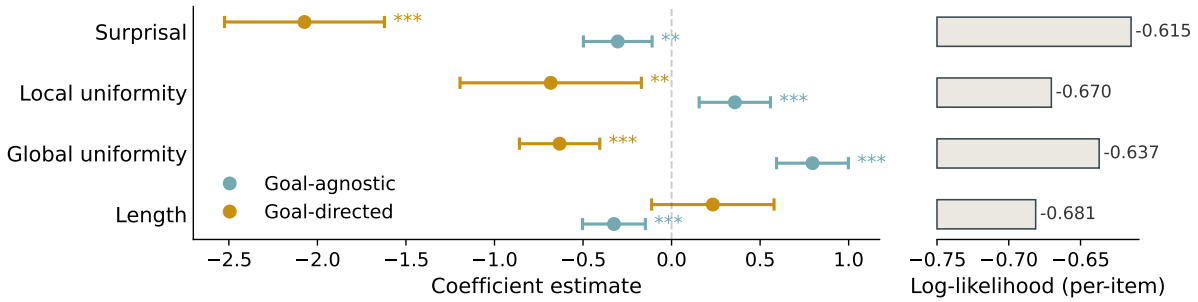

Figure 3: Logistic regression results predicting whether a continuation is selected over an alternative as a function of the difference in cost with respect to the alternative, goal condition, and their interaction. Points show coefficient estimates on the log-odds scale, horizontal bars indicate 95% confidence intervals. Asterisks indicate significance levels ($p < .05$, $p < .01$, $p < .001$). On the right panel is the corresponding per-item log-likelihood for each model.

We adopt a pairwise formulation because the goal-directed and goal-agnostic alternative sets differ both in composition and size, which makes coefficients from standard discrete-choice models, such as conditional logit, not directly comparable across the two conditions.[8] To test whether cost sensitivity differs between the goal-directed and the goal-agnostic condition, we include an interaction between cost differences and the goal condition. Standard errors are clustered at the context level to account for the non-independence induced by multiple comparisons within the same context. Full model details are provided in App. C.2.

Fig. 3 summarises the results. The surprisal-based model achieves the highest log-likelihood. Surprisal exhibits a consistent negative effect, with lower surprisal increasing the probability of selection. This effect is substantially stronger (approx. 7×) in the goal-directed condition than in the goal-agnostic condition, supporting the interpretation of surprisal as a speaker-oriented cost: among continuations that realise the same goal, speakers preferentially select those with lower surprisal.

For local and global uniformity, lower cost (i.e., a more uniform information profile) predicts choice in the goal-directed condition, but this relationship reverses in the goal-agnostic condition, where higher cost (lower uniformity) is associated with higher choice probability. In conjunction with §6.1, this suggests that although human continuations minimise uniformity-based costs at above-chance rates, this does not extend to a general probabilistic preference for higher-uniformity continuations.

Finally, we find a preference for shorter continuations in the goal-agnostic condition, but no effect in the goal-directed condition. In combination with the rank-based results, this suggests that length operates as a listener-oriented pressure, though the model itself yields the weakest overall fit.

[8]For completeness, we report conditional logit models fitted separately for goal-directed and goal-agnostic alternatives in App. D.5. These yield qualitatively similar patterns but do not permit direct cross-condition comparison.

7 Conclusion

In this paper, we argued that notions of production cost must be interpreted relative to the alternative utterances over which they are evaluated. By explicitly distinguishing between goal-directed and goal-agnostic alternative sets, we showed that the same cost measure can give rise to qualitatively different interpretations. Our empirical analyses indicate that surprisal minimisation over goal-directed alternatives provides the strongest account of production choices at the main-verb choice point. This effect is robust across both rank-based and probabilistic analyses and reflects a speaker-oriented pressure to select low-surprisal continuations among alternatives that realise the same communicative goal.

Methodologically, we introduced scalable procedures for constructing contextual alternative sets that enable probabilistic pragmatic models of production to be instantiated and evaluated at scale in open-ended settings, while providing a principled basis for comparing competing notions of cost. More broadly, our results highlight the importance of making alternative spaces explicit when interpreting information-theoretic measures of cost, and suggest that alternative-conditioned optimisation offers a fruitful framework for studying speaker and listener pressures in naturalistic language use and for reproducing them in generation systems.

Limitations

Firstly, while the theoretical and methodological approach forms a core part of our contributions, the generalisability of our empirical findings is limited to the English language dataset we choose as our case study, and the relatively small subset of samples we select. It remains for future work to assess the extent to which these results generalise to other dialogue corpora, production types, and languages.

Second, our analysis focuses on a single class of choice points, matrix verb continuations. Although this provides a controlled setting and directly compares to classic work on production choice, it limits the scope of the empirical claims. Future work could apply the same framework to a broader range of syntactically and information-theoretically determined choice points, such as dative alternations or positions of high continuation entropy, in order to assess the generality of the observed patterns.

Third, our cost metrics are computed as global aggregates over entire continuations. While we control for length by selecting relatively short human continuations and matching the distributions of generated alternatives, this approach may become problematic for longer utterances, where contributions from later parts of the string can reduce sensitivity to the region immediately following the choice-point. More refined aggregation schemes—such as weighting units by their distance from the choice point—may better capture incremental planning and more realistic production horizons.

Fourth, our analyses rely on language models both to estimate information-theoretic quantities and to generate alternative sets. These models may not fully capture human processing or behaviour and may be ill-suited for simulating alternatives in certain contexts, especially settings that are poorly represented in training data, including low-resource languages and under-represented speaker communities within a language.

Finally, our framework does not incorporate an explicit notion of communicative effectiveness. In the open-ended dialogue setting we consider, there is no clear external success signal, and well-formed utterances between competent speakers can be assumed to be broadly understood, thus making effectiveness effectively constant across alternatives. Introducing a more nuanced notion of effectiveness would require an explicit model of listener interpretation, which remains an open challenge and lies beyond the scope of the present work.

Acknowledgments

We thank the anonymous ARR reviewers for their helpful comments. We disclose the use of generative AI tools for light editing and rephrasing; the original text was our own, and we carefully reviewed all suggested edits.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56. PMID: 15298329.

E.G. Bard, A.H. Anderson, Y. Chen, H.B.M. Nicholson, C. Havard, and S. Dalzel-Job. 2007. Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language*, 57(4):616–641. Language-Vision Interaction.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9:20–1.

Simon Betz, Nataliya Bryhadyr, Olcay Türk, and Petra Wagner. 2023. Cognitive load increases spoken and gestural hesitation frequency. *Languages*, 8(1).

Kathryn Bock and Willem JM Levelt. 1994. Language production: Grammatical encoding. In *Handbook of psycholinguistics*, pages 945–984. Academic Press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 81–90.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43:651–681.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(Volume 9, 2023):519–540.

Judith Degen, Michael Franke, and Gerhard Jager. 2013. Cost-based pragmatic inference about referential expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke. 2014. Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Richard Futrell. 2023. Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, 120(39):e2220593120.

Richard Futrell. 2024. An information-theoretic account of availability effects in language production. *Topics in Cognitive Science*, 16(1):38–53.

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.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72.

Mario Giulianelli. 2022. Towards pragmatic production strategies for natural language generation tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023a. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 665–682, Online only. Association for Computational Linguistics.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2026. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Journal of Memory and Language*, 148.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023b. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

Noah D. Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics uncertainty in language and thought. In *The Handbook of Contemporary Semantic Theory*, chapter 21, pages 655–686. John Wiley & Sons, Ltd.

Britt Hadar, Joshua Skrzypek, Arthur Wingfield, and Boaz Ben-David. 2016. Working memory load affects processing time in spoken word recognition: Evidence from eye-movements. *Frontiers in Neuroscience*, 10.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Barbara Howarth and Anne H. Anderson. 2007. Introducing objects in spoken dialogue: The influence of conversational setting and cognitive load on the articulation and use of referring expressions. *Language and Cognitive Processes*, 22(2):272–296.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Preprint*, arXiv:2304.04758.

Jennifer Hu, Roger Levy, and Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Iva Ivanova and Victor S. Ferreira. 2019. The role of working memory for syntactic formulation in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(10):1791–1814.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform Information Density effects on syntactic choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, Santa Fe, New-Mexico. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

R Duncan Luce. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498, Miami, Florida, USA. 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.

Antje Meyer. 2023. Timing in conversation. *Journal of Cognition*, 6.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *Preprint*, arXiv:2410.05229.

Aron Molnar, Jaap Jumelet, Mario Giulianelli, and Arabella Sinclair. 2023. Attribution and alignment: Effects of local context repetition on utterance production and comprehension in dialogue. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 254–273, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11.

Jonathan Peelle. 2017. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39:1.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. *Preprint*, arXiv:2109.15000.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Anastasia Klimovich-Gray, Jaap Jumelet, Nika Adamian, and Agnieszka Konopka. 2026. Structural priming in humans and large language models. *Journal of Memory and Language*, 149:104713.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, and Mario Giulianelli. 2025. The harmonic structure of information contours. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31636–31659, Vienna, Austria. Association for Computational Linguistics.

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.

Julia White, Jesse Mu, and Noah D Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

Jun Sen Yee, Mario Giulianelli, and Arabella J. Sinclair. 2024. Efficiency and effectiveness in task-oriented dialogue: On construction repetition, information rate, and task success. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5562–5577, Torino, Italia. ELRA and ICCL.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Data Preparation

The target utterances and their full preceding conversations are extracted from the SWDA corpus (Stolcke et al., 2000). The target utterances are then split into a context and a continuation, and the dialogue history is stored. A turn is defined as the unit of dialogue consisting of all adjacent utterances spoken by the same speaker. This approach cuts out backchannels and disfluencies in the corpus. Table 2 displays relevant dataset statistics before and after the pre-processing steps were taken.

Dataset	Utterances	Dialogues
Before Pre-Processing	221,616	1,155
After Pre-Processing	1,342	680

Table 2: Data statistics of the corpus before and after pre-processing steps.

A.1 Dialogue Data Cleaning and Preparation

All of the text in the corpora was filtered through the use of regular expressions to remove any formatting or tags in the utterances. For example: «motorcycle noise». Additionally, regular expressions were also used to remove repeated words, interrupted words, and short disfluencies such as "um" or "uh". Additional annotations of contextual noise in the conversations were also removed. Short utterances that are backchannels are then removed, these are utterances that contain three or less words. The dataset contains several cases where adjacent utterances within a dialogue were spoken by the same speaker. This was usually because of a long pause or interruption in the dialogue. The previous step of removing backchannels also leads to more adjacent utterances by the same speaker. Adjacent utterances are therefore combined in order to keep these continuous singular thoughts together. The corpus was then re-indexed and numbered so that there were no missing or duplicated values in the identification columns.

A.2 Extracting Target Utterances

Not all of the utterances in the dataset were considered high quality enough to be used for this experiment. In order to select target utterances, the following criteria were devised: the context must be of length 3 or greater to allow for enough of the dialogue history that the model can attempt to predict the intent of the speaker, the utterance must have a verb as this is the location where the utterances are split, and the utterance must be between 10 and 30 words to guarantee that utterances will be split into contexts that are long enough to provide the language model with enough information to generate continuations without confusing the models with unnecessary or extra information. Target utterances are selected as the first sentence in a turn, which ensures that they are most directly related to the information provided by the previous turn.

A.3 Extracting Dialogue History

We are interested in the preceding dialogue history before the context of each utterance. For each target utterance we extract the preceding turn spoken by the other speaker, and the maximum number of utterances from the preceding dialogue history that can fully fit within the 1024 token limit of the GPT-2, the model used for surprisal calculation.

A.4 Choice Points: Splitting Sentences

Inspired by Jaeger's classic *that*-mentioning example, we opt for the matrix verb as our choice point. Matrix verbs mark a point at which the speaker is likely to have already developed an utterance plan and they must make a choice as to how to realise their intended completion. The fact that some meaningful partial sentence context has been established constrains the alternatives to a degree (rather than simply predicting the continuation from "the boy"), but not to an extent that the continuation is necessarily obvious (e.g., "the cat sat on the").

We thus split utterances at the root verb to create the utterance context and the human continuation. Root verbs are identified using spaCy's part-of-speech tagger and dependency parser.[9]

Other choice points would be feasible and interesting to study under the same framework, and we are excited about future work that examines this more exhaustively. Choice points could be chosen, for example, by tracking other syntactic structures. For instance, dative verbs (like "the man gave . . . ") could be completed with either a prepositional object structure ("...the book to the boy") or with a double object dative structure ("...the boy the book"). Alternatively, other more linear criteria could be used—for example, choice points could be determined by measuring continuation entropy at every position in the sentence and choosing the positions with the highest entropy.

B Generating Alternatives

Language models are used as the simulators of dialogue production in order to generate alternative continuations. The model used to generate alternatives was OpenAI's GPT-4o[10]. In order to ensure that the model generated continuations in the correct format, a one-shot prompting method was used to provide the model with a single simple demonstration of a sentence completion.

B.1 Goal-Agnostic Alternatives

The sentence continuation generation methods prompt the language model to generate continuations of the target context. This was done by providing the model with the context and various levels of context. The models' objective was defined through the use of the developer prompt, specifying to complete the provided context.

Each method of sentence continuation was provided with the same developer prompt.

(3) [Developer:] Your task is to complete the provided sentence. Complete the sentence in a natural manner, as if engaging in a phone call conversation. Only write the continuation to the sentence without any additional information or words in your response.

(4) [User:] Complete the sentence: "The cat jumped"

(5) [Assistant:] The cat jumped over the dog.

No Dialogue History. The no history condition provides the model with the context but no other information. This gives the model the most freedom to complete the sentence.

No Dialogue History User Prompt

(6) [User:] Complete the sentence: "{context}"

[9] https://spacy.io/

[10] https://platform.openai.com/docs/models/gpt-4o

Previous Utterance. The preceding history condition provides the model with the context and the directly preceding utterance.

Previous Utterance History User Prompt

(7) [User:] Given this sentence from speaker A: “{history}", Complete the sentence from Speaker B: “{context}"

All Previous Utterances. The full history conditions provide the model with all previous utterances in the dialogue, up to a maximum of 1000 tokens.

Full History User Prompt

(8) [User:] Given this phone conversation between Speaker A and Speaker B: “{history}", Complete the sentence from Speaker {SpeakerID}: “{context}"

B.2 Goal-Directed Alternatives

To generate goal-directed alternatives we prompt GPT-4o to paraphrase the full sentence, but only allow it to change the words after the context. This approach leverages the language model's internal representation of the semantic closeness of words. In order to limit the repetition of paraphrases, all of the paraphrases for a target utterance are generated at once and specified to be unique from each other. This method resulted in several occurrences where the model refused to generate the paraphrases at all, or refused to generate all 10 paraphrases. For example the model would return “I'm sorry but I don't understand." or “I can't do that." In order to keep the data balanced, it was decided to drop all continuations of the contexts that had failed to produce the 10 paraphrases. This reduced the number of contexts to 309 with 309 human continuations, and 12,360 continuations representing 3,090 continuations of each generation method, or 10 continuations per method per context. This is the final dataset upon which the calculations were performed.

Paraphrase Generation Developer Prompt

(9) [Developer]: Your task is to paraphrase the provided sentence. Paraphrase the sentence in a natural manner, as if engaging in a phone call conversation, while still keeping the sentence factually the same. Only write the paraphrase of the sentence without any additional information or words in your response. Try your best to do so even if the provided text seems nonsensical or does not have proper sentence structure.

Paraphrase Generation User Prompt

(10) [User:] Write 10 unique paraphrases of the following sentence: “{text}", the paraphrases must always start with the following words: “{context}"

Paraphrase Generation Assistant Prompt

(11) [User:] Write 10 unique paraphrases of the following sentence: “The cat jumped over the dog.", the paraphrases must always start with the following words: “The cat jumped over"

(12) [Assistant:] The cat jumped over the husky.
The cat jumped over the dog without making a sound.
The cat jumped over the dog as it wagged its tail.
The cat jumped over the dog, barely missing it.
The cat jumped over the dog and landed smoothly.
The cat jumped over the dog while it was dozing off.
The cat jumped over the dog to escape quickly.
The cat jumped over the dog in a playful manner.
The cat jumped over the dog and continued running.
The cat jumped over the dog, surprising it.

B.2.1 Paraphrase Detection

In order to determine which generations should be kept and which should be removed, an analysis of the contexts combined with the generated continuations compared to the original human sentences was performed. Following Farquhar et al. (2024), we initially made use of an MNLI model to detect paraphrases and semantic equivalence, but through manual inspection, found this approach to be less effective than GPT-4, with a 85% vs. 98% success rate on a manually-annotated sample of 400 items. We tested two Natural Language Inference

models trained on the MNLI corpus (Williams et al., 2018), Roberta Large MNLI (Liu et al., 2019) and FLAN-T5 Base MNLI (Chung et al., 2022), chosen for their high scores on the MNLI benchmark.[11] The labels produced were either entailment, neutral, or contradiction, we take entailment as goal conditioned, and merge neutral and contradiction for goal agnostic. However, upon manual inspection and annotation of a sample of 400 generations, we abandoned this approach, since there was too high a level of ambiguity in terms of the neutral label, and too high an error rate. We instead opt for prompting GPT-4o to detect whether two sentences are paraphrases, a well-defined and more simple task (especially for the relatively short sentences that we select as targets), which exploits LMs greater level of capability for semantic meaning over logical equivalency (Mirzadeh et al., 2025), as well as using a far larger and more powerful model. The following prompts were used for the paraphrase detection with GPT-4o.

Paraphrase Detection Developer Prompt

(13) [Developer:] Your task is to determine whether or not two sentences are paraphrases of each other. You are to classify the sentences into one of two labels: “yes" if the sentences are paraphrases or “no" if they are not. Do not provide any explanation for your choice, just the name of the label.

Paraphrase Detection User Prompt

(14) [User:] Classify whether these sentences are paraphrases. Sentence A: “{text}", Sentence B: “{generation}"

Paraphrase Detection Assistant Prompt

(15) [User:] Classify whether these sentences are paraphrases. Sentence A: “The cat jumped over the dog.", Sentence B: “The cat jumped over a dog."

(16) [Assistant:] Yes

Through the use of this model a set of goal-directed alternatives was created. This is the set of alternatives that are considered paraphrases by the language model.

[11] https://paperswithcode.com/sota/natural-language-inference-on-multinli

C Statistical Analyses

C.1 Poisson–binomial Test for Rank-1 Outcomes

We provide here the full specification of the test used to assess whether human utterances are ranked first more often than expected by chance.

For each trial $i \in \{1, \ldots, N\}$, let n_i denote the number of alternatives in the relevant alternative set. Let $Y_i \in \{0, 1\}$ be a Bernoulli random variable indicating whether the human utterance is ranked first under the cost measure of interest. Under the null hypothesis H_0 of random selection among alternatives, the probability of a rank-1 outcome on trial i is

$$P(Y_i = 1 \mid H_0) = \frac{1}{n_i}. \tag{14}$$

The total number of rank-1 outcomes across trials is given by the random variable

$$K = \sum_{i=1}^{N} Y_i, \tag{15}$$

which follows a Poisson–binomial distribution corresponding to the sum of independent Bernoulli variables with heterogeneous success probabilities $\{1/n_i\}_{i=1}^{N}$. Further let

$$K_{\text{obs}} = \sum_{i=1}^{N} y_i \tag{16}$$

denote the observed number of rank-1 outcomes in the data, where y_i is the realised value of Y_i. Statistical significance is assessed using a one-sided test,

$$p = P(K \geq K_{\text{obs}} \mid H_0), \tag{17}$$

which quantifies the probability of observing at least as many rank-1 outcomes as in the data under the null hypothesis of chance selection.

We employ a one-sided test because cost minimisation predicts an excess, but not a deficit, of rank-1 outcomes relative to chance.

C.2 Pairwise Logistic Choice Model

This appendix provides a detailed description of the pairwise logistic choice model used in §6.2.

Data construction. For each context, we construct binary comparisons between the human continuation and each candidate alternative. Let $\boldsymbol{a}_i$ denote the human continuation and $\boldsymbol{a}_j$ an alternative from the same context. Each pair yields a

binary outcome $y_{ij} \in 0, 1$, where $y_{ij} = 1$ indicates that the human continuation is preferred over the alternative. To obtain a balanced dataset, we include both (i, j) and (j, i).

Predictors. For each cost function $C \in \{C_{\text{surp}}, C_{\text{UID}_{\text{loc}}}, C_{\text{UID}_{\text{gl}}}, C_{\text{len}}\}$, as introduced in §4, we define a cost difference variable

$$\Delta C_{ij} = C(\boldsymbol{a}_i; \boldsymbol{c}) - C(\boldsymbol{a}_j; \boldsymbol{c}), \tag{18}$$

which captures the relative cost of the human continuation compared to the alternative. Cost differences are standardised to have mean zero and unit variance prior to estimation. We also include a binary indicator $\text{GD}_j \in \{0, 1\}$, where $\text{GD}_j = 1$ if the alternative $\boldsymbol{a}_j$ is drawn from the goal-directed set and $\text{GD}_j = 0$ if it is drawn from the goal-agnostic set.

Model specification. We model the probability that the human continuation is preferred using a logistic regression:

$$\log \frac{P(y_{ij} = 1)}{1 - P(y_{ij} = 1)} = \beta_0 + \beta_1 \, \Delta C_{ij} + \beta_2 \, \text{GD}_j + \beta_3 \, (\Delta C_{ij} \times \text{GD}_j) \tag{19}$$

The coefficient β_1 captures cost sensitivity in the goal-agnostic condition, while β_3 captures the change in cost sensitivity in the goal-directed condition. The total cost sensitivity under goal-directed alternatives is therefore given by $\beta_1 + \beta_3$. As mentioned in §6.2, this specification corresponds to the two-alternative restriction of the softmax production rule. In this interpretation, the slope on ΔC_{ij} estimates the negative cost-sensitivity parameter $-\alpha$, such that more negative coefficients indicate stronger sensitivity to a given cost.

Estimation. Models are estimated via maximum likelihood. Standard errors are clustered at the context level to account for the dependence induced by multiple pairwise comparisons within the same context. For each cost function, we fit a separate model and report coefficient estimates, confidence intervals, and per-observation log-likelihoods.

Interpretation. A negative coefficient on ΔC_{ij} indicates that lower-cost continuations are more likely to be selected. The interaction term allows us to test whether this effect differs between goal-directed and goal-agnostic alternatives within a single unified model.

D Additional Results

D.1 Lexical Overlap between Context and Continuation

We evaluate lexical overlap between context and continuations across human and LM-generated alternatives, similar to Molnar et al. (2023). The results of these comparisons can be found in Fig. 4.

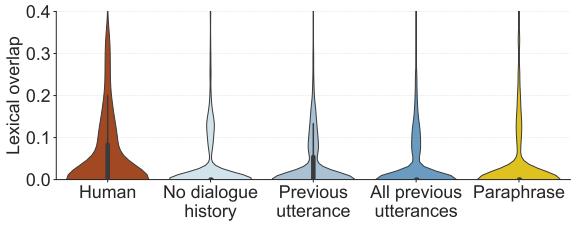

Figure 4: Distribution of lexical overlap between context and continuations by continuation type.

D.2 Goal Predictability from Context

Goal-agnostic alternative sets are intended to capture a listener's uncertainty over both the speaker's intended goal and its realisation. In practice, however, contextual constraints often make some goals more predictable than others. As a result, even without conditioning explicitly on the goal, a listener may assign non-zero probability to continuations that align with the speaker's intended goal.

We examine this in our data by identifying goal-agnostic continuations that are also contained in the goal-directed set. We find a small but non-zero overlap, which increases with additional context: with longer context windows, a larger proportion of goal-agnostic continuations match the intended goal. Fig. 5 breaks down these proportions.

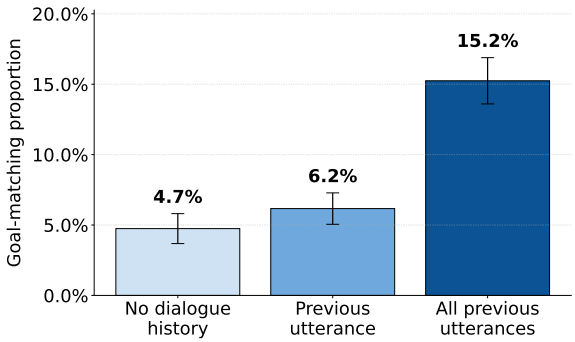

Figure 5: Proportion of goal-agnostic alternatives that match the goal of the observed utterance, across dialogue history conditions.

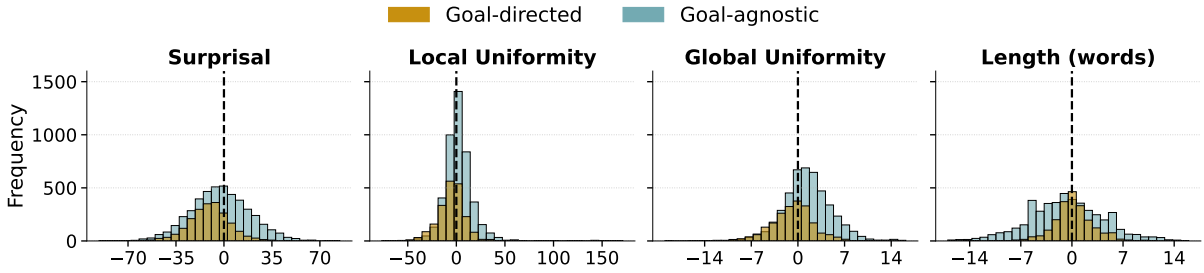

Figure 6: Pairwise cost differences between human continuations and alternatives, computed as the cost of the human continuation minus the cost of the alternative. Negative values indicate that the human continuation has lower cost than the alternative (i.e., is preferred under the cost), while positive values indicate that the alternative has lower cost.

D.3 Cost Distribution Differences by Context

We additionally examine per-context differences in cost between human continuations and sampled goal-directed and goal-agnostic alternatives. Under the model in §3, speakers are expected to prefer continuations with lower cost than competing alternatives. To operationalise this, we sample a single alternative per context and compute its cost difference relative to the human continuation. The resulting difference distributions are shown in Fig. 6.

One-sided t-tests (Tab. 3) show that surprisal exhibits a strong effect under both the goal-directed and the goal-agnostic conditions, with a markedly larger effect in the goal-directed case. Local uniformity and global uniformity show a significant effect only in the goal-directed comparison, with differences in the opposite direction, and thus non-significant under the one-sided test, in the goal-agnostic condition. The opposite pattern holds for length. Overall, these results confirm that surprisal most clearly distinguishes human continuations from goal-directed alternatives.

	Goal-directed		Goal-agnostic	
Cost difference	t	p-value	t	p-value
Surprisal	-32.48	$< 10^{-8}$	-8.23	$< 10^{-8}$
Local uniformity	-13.37	$< 10^{-8}$	10.57	1.00
Global uniformity	-12.11	$< 10^{-8}$	26.48	1.00
Length (words)	2.99	1.00	-10.72	$< 10^{-8}$

Table 3: Results of one-sided t-tests comparing observed productions to alternative continuations, testing whether mean differences are smaller than zero.

D.4 Pairwise Logistic Choice Model

Tab. 4 contains supporting details of coefficients and confidence intervals for Fig. 3. The pairwise logistic choice model and the interpretation of the coefficients is presented in App. C.2.

D.5 Conditional Logit Model of Graded Cost Sensitivity

In addition to our analyses in the main body of the paper, we also model utterance choice as probabilistic using a conditional logit model. For each context $\boldsymbol{c}_i$ in the dataset, let $\mathcal{A}_i$ denote the corresponding alternative set—either the goal-agnostic set $\mathcal{A}_{\boldsymbol{c}_i}$ or the goal-directed set $\mathcal{A}_{\boldsymbol{c}_i,g}$ as defined in §3.1. Let Y_i be a categorical random variable over $\mathcal{A}_i$ representing the production choice in context $\boldsymbol{c}_i$. Each alternative $\boldsymbol{a} \in \mathcal{A}_{\boldsymbol{c}_i}$ is associated with a cost $C \in \{C_{\text{surp}}, C_{\text{UID}_{\text{loc}}}, C_{\text{UID}_{\text{gl}}}, C_{\text{len}}\}$, as introduced in §4. We model the probability of selecting continuation $\boldsymbol{a}$ from $\mathcal{A}_i$ as

$$P(Y_i = \boldsymbol{a} \mid \mathcal{A}_{\boldsymbol{c}_i}) = \frac{\exp(-\alpha\, C(\boldsymbol{a}; \boldsymbol{c}_i))}{\sum_{\boldsymbol{a}' \in \mathcal{A}_{\boldsymbol{c}_i}} \exp(-\alpha\, C(\boldsymbol{a}'; \boldsymbol{c}_i))}, \tag{20}$$

where α is a scalar coefficient corresponding to the sensitivity parameter of the choice model in §3.2.

We fit conditional logit models with a single cost measure at a time, standardising each cost across the dataset, and compare them using four metrics: (1) the estimated cost-sensitivity coefficient α; (2) the average per-item log-likelihood ℓ; (3) the expected probability that the model assigns to the lowest-cost alternative, $P(\text{rank} = 1)$, which describes how confidently the model selects the top alternative (in other words, when $P(\text{rank} = 1) = 1$, the model selects the top alternative deterministically); and (4) the probability that the cost-minimising alternative is preferred over the second-best, $P(\text{best vs. 2nd})$. This last metric is comparable across goal-directed and goal-agnostic conditions, as it is insensitive to alternative set size. Tab. 5 summarises these results.

Within goal-directed alternative sets, surprisal is by far the strongest predictor of choice, with a

	Goal-agnostic		Goal-directed		Interaction		
Cost	β_1	CI	$\beta_1 + \beta_3$	CI	β_3	CI	**Per-item LL**
Surprisal	$-0.304^{\dagger}$	$[-0.498, -0.110]$	-2.073	$[-2.525, -1.622]$	-1.769	$[-2.210, -1.328]$	-0.615
Local uniformity	0.357	[0.156, 0.559]	$-0.683^{\dagger}$	$[-1.195, -0.170]$	-1.040	$[-1.532, -0.548]$	-0.670
Global uniformity	0.796	[0.593, 0.999]	-0.632	$[-0.859, -0.405]$	-1.428	$[-1.678, -1.179]$	-0.637
Length	-0.325	$[-0.503, -0.148]$	0.233	[−0.113, 0.579]	0.558	[0.250, 0.867]	-0.681

Table 4: Pairwise logistic regression estimates per cost measure. Results are significant at $p < 0.001$, daggers (†) indicate coefficients with $p < 0.01$, gray indicates non-significance.

	Goal-directed alternatives				Goal-agnostic alternatives			
Cost	α	$\ell(\uparrow)$	$P(\text{rank}=1)$	$P(\text{best vs 2nd})$	α	$\ell(\uparrow)$	$P(\text{rank}=1)$	$P(\text{best vs 2nd})$
Surprisal	**1.918**	**-0.187**	**0.464**	**0.701**	$0.253^{\dagger}$	-0.173	0.090	0.518
Local uniformity	0.825	-0.235	0.245	0.557	-0.359	-0.172	0.111	0.549
Global uniformity	0.748	-0.241	0.223	0.552	**-0.840**	**-0.164**	**0.169**	**0.584**
Length (words)	$-0.335^{\ddagger}$	-0.250	0.171	0.526	0.339	-0.172	0.098	0.521
Uniform *(baseline)*	–	-2.018	0.142	0.500	–	-2.727	0.067	0.500

Table 5: Conditional logit results for production choice in goal-directed and goal-agnostic alternative sets, reporting the estimated cost-sensitivity coefficient α, the average per-item log-likelihood ℓ, the expected probability $P(\text{rank}=1)$ that the human continuation is ranked first under the corresponding cost measure, and the probability $P(\text{best vs 2nd})$ that the cost-minimising alternative is preferred over the second-best alternative. Daggers (†) indicate coefficients with $p < 0.01$; double daggers (‡) indicate coefficients with $p < 0.05$; all other coefficients are significant at $p < 0.001$.

large positive cost-sensitivity coefficient, the highest log-likelihood, and a $P(\text{rank}=1)$ substantially higher than chance. This model also achieves the highest $P(\text{best vs. 2nd})$ across the board, confirming speaker-oriented surprisal minimisation as the account that best explains production choice.

Uniformity-based costs also show an asymmetry across alternative sets. In goal-directed contexts, both local and global uniformity significantly predict choice. In contrast, within goal-agnostic sets both measures have significant negative coefficients, indicating that greater uniformity reduces the likelihood of selection when evaluated against listener-available alternatives. Utterance length shows the weakest effect within goal-directed alternatives, where it has a negative coefficient, and remains comparatively weak relative to uniformity costs in goal-agnostic sets.

D.6 Results without Stratified Sampling

In this section, we report the results obtained using the full set of generated utterances, without applying the stratified sampling procedure described in §5.4. The results, shown in Figs. 7 and 8 and Tab. 6 and 7, are qualitatively aligned to those reported in the main text. One numerical difference is in the log-likelihood of the global uniformity model, which is larger in the results without stratified sampling. This is likely due to the global distributional differences that we remove through stratification.

Cost	**Goal-directed**	**Goal-agnostic**
Surprisal	47.6% ×5.12	10.7% ×3.24
Local uniformity	22.1% ×2.38	12.4% ×3.76
Global uniformity	13.0% ×1.40	13.4% ×4.06
Length (words)	22.8% ×2.45	22.5% ×6.82
Uniform *(baseline)*	9.3%	3.3%

Table 6: *Results without Stratified Sampling*. Percentage of observed rank-1 outcomes, with multiplicative increase relative to the uniform baseline. All values exceed the uniform baseline under one-sided Poisson–binomial tests ($p < 10^{-8}$ or smaller), except for global uniformity against goal-directed alternatives ($p = 0.018$), which does not show a robust effect.

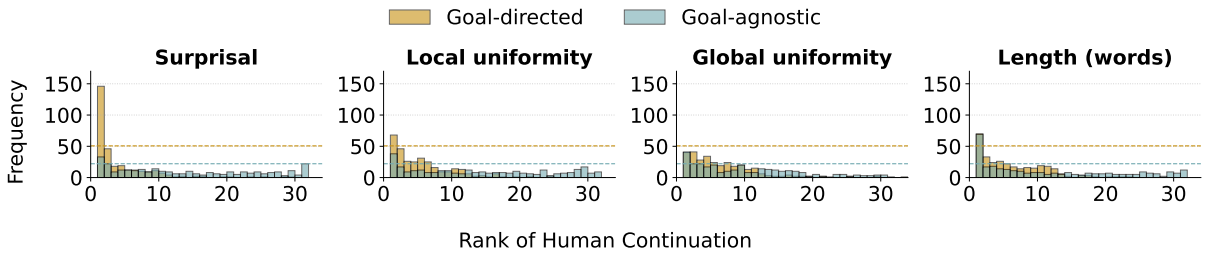

Figure 7: *Results without Stratified Sampling*. Ranking distributions of observed human continuations under different cost measures, evaluated against goal-directed and goal-agnostic alternative sets. A rank of 1 indicates that the human continuation has the lowest cost among the available alternatives, corresponding to deterministic cost minimisation.

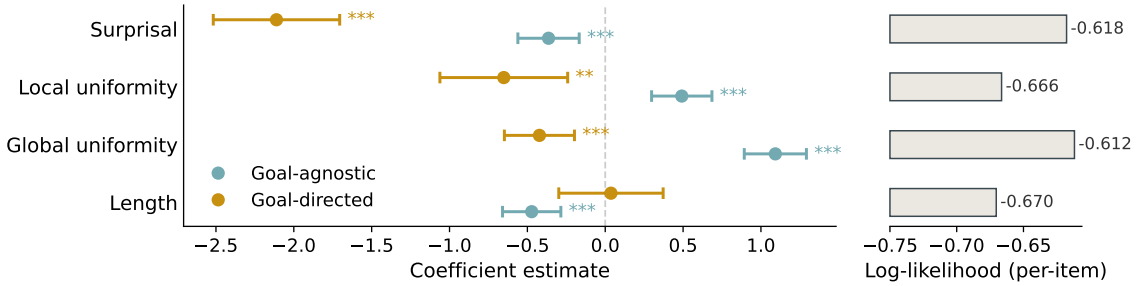

Figure 8: *Results without Stratified Sampling*. Logistic regression results predicting whether a continuation is selected over an alternative as a function of the difference in cost with respect to the alternative, goal condition, and their interaction. Points show coefficient estimates on the log-odds scale, horizontal bars indicate 95% confidence intervals. Asterisks indicate significance levels ($p < .05$, $p < .01$, $p < .001$). On the right panel is the corresponding per-item log-likelihood for each model.

	Goal-agnostic		Goal-directed		Interaction		
Cost	β_1	CI	$\beta_1 + \beta_3$	CI	β_3	CI	**Per-item LL**
Surprisal	-0.364	$[-0.561, -0.167]$	-2.111	$[-2.517, -1.705]$	-1.747	$[-2.147, -1.347]$	-0.618
Local uniformity	0.492	$[\ 0.298,\ 0.685]$	$-0.651^{\dagger}$	$[-1.062, -0.241]$	-1.143	$[-1.531, -0.756]$	-0.666
Global uniformity	1.093	$[\ 0.893,\ 1.292]$	-0.423	$[-0.648, -0.198]$	-1.515	$[-1.753, -1.278]$	-0.612
Length	-0.472	$[-0.660, -0.285]$	0.037	$[-0.298,\ 0.372]$	0.509	$[\ 0.219,\ 0.799]$	-0.670

Table 7: *Results without Stratified Sampling*. Pairwise logistic regression estimates per cost measure. Results are significant at $p < 0.001$, daggers (†) indicate coefficients with $p < 0.01$, gray indicates non-significance.