Hedging Our Bets on LLMs

Challenges in Calibrating the Language of Uncertainty

Nikola Datkova

A THESIS

in

Linguistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Master of Arts

2025

Supervisor of Thesis
Charles Yang
Professor, Director of Undergraduate Studies in Cognitive Science

Graduate Group Chair
Meredith Tamminga
Associate Professor, Graduate Chair of Linguistics

ACKNOWLEDGEMENTS

There is a timeless scene in *The Imitation Game*, where Alan Turing, a woefully underappreciated mathematician and a long-overlooked war hero, is interrogated about his role at Bletchley Park and asked a question that happened to have shaped much of his tragically short career: "Can machines think?" He dismisses that machines think as humans do but concludes with a thought-provoking line: "Just because something thinks differently from you, does that mean it's not thinking?"

For those of us with neurospicy brains, it has been abundantly clear for much of our lives that not all minds think alike. It's no surprise that quite a few of us grow so fascinated by the human language, a medium—or a gift, if you will—that helps us connect despite our differences. An eighth wonder of nature that is unique to each of us yet unites all of humanity. A fine art carrying an infinite spoken and unspoken code of which we can only name fragments, no matter how close we listen.

And yet, we find ourselves living an age when machines, of which no one until recently expected anything more than cold computation, use this medium to talk to us. Now, isn't it sort of incredible how something that has never interacted with anything from our world *except* language has built such a robust idea of what the world's like? On the other hand, do machines use genuine human language or just an eerily similar pretence? Can they be considered linguistically competent?

The emergence of capable LLMs is making us reconsider fundamental questions about language, mind, and the future of everyday interaction. Throughout my time at Penn, with a brief rotation to Cambridge, I've been fortunate enough to get to dissect these questions from numerous unique angles in the company of the biggest names in their fields. It has been an intellectually fulfilling time, in which I gathered some answers about language and minds—human and artificial—but more so questions to carry forward.

For that, I have personally to thank a number of folks:

First and foremost, to Julie and Charles for always believing the best in me even when I didn't deserve it. To Julie also for always having an astute question up her sleeve and for sticking around even when it wasn't always syntax that I was up to. To Charles, amongst other things, for his always encouraging comments such as "I suggest you submit the draft, you know, *expediently*" or "Were you thinking of finishing this by May or August?" They worked, clearly.

To everyone I crossed paths with during my time in the department. To Dan, Héctor, Giang, Annika, Giachi, and AJ, for the friendships we shared and the lessons we learnt together along the way. To Florian, Marlyse, and Martin, for supporting my academic as well as entrepreneurial endeavours for no good reason other than, you know, on principle. To Gareth, Don, Colin, and Amy for never turning down good fun. To Meredith for keeping an open mind. To Rolf for his captivating stories. To all other members of this small yet vivid community, past and present, with whom I grabbed coffee or shared a chat, however brief. It all counted.

To the professors I had an opportunity to work closely with in my time in Cambridge, namely Professor Napoleon Katsos for always inserting interesting takes on games and pragmatics, Professor Dora Alexopoulou for prompting me to ask questions about parts of language I wouldn't get to ask otherwise, and Professor John Williams for offering clear guidance as I was trying to learn the means of experimental research.

To Professor Chris Callison-Burch from Penn Engineering for painting an honest yet hopeful picture of technology. To Professor Mark Yatskar for not being easily impressed. Also an anti-thank-you to my computer science study buddy Jan for getting me hooked on Altoids. This study could've been completed on about 70 fewer grams of added sugars without you.

To Mihály for inspiring creativity in research and notadvising on my projects (including but not limited to this one). To Andis and Maria, for saying yes to things. To Harry for his

thoughtful opinions and thorough feedback (LaTeX-ed, naturally) on literally any and every rabbit hole I was ready to go down, in and outside of computer science. To Gavin, well, for always having an opinion.

To Euan, for tirelessly reminding me of where home is.

To Will, for making life less ordinary.

To Marek and Rebeka, for making life more ordinary.

To everyone with and from whom I learnt essential (not only) survival skills in school. To Matúš for paying attention. To Juraj, for making me read pop-ups. University would've been so much more difficult without you.

To my family for supporting me the best they knew how. To the Kramer family for cheering for me all along.

If anyone finds themselves of the opinion that they have been wrongfully left out, please submit your complaint to nowe.moore@gmail.com, and I'll respond within the statutory period.

All mistakes, in this work and beyond, remain my own.

ABSTRACT

We live an age where systems employing large language models (LLMs) are increasingly becoming a greater part of society. In order for this technology to help us make appropriate everyday (and less everyday) choices, we need to be able to rely on it to convey accurate information. The present study examines the capability of LLMs to produce well-calibrated language with regard to expressing uncertainty in decision-making situations. To do so, the study first carries out an experiment where human subjects evaluate distinct linguistic hedging strategies to establish the order of dominance for low-, medium-, and high-stake contexts. The Bradley-Terry Model is used to analyse the rates, and a novel statistically significant hierarchy is identified. In the second part of the study, LLM-generated language is analysed for the previously selected hedging strategies. No adjustments with regard to the human perception hierarchy were identified, but LLMs demonstrated distinct behaviours which are discussed in the context of LLM alignment. The discovered patterns may influence how humans interpret the information conveyed by LLM-based agents, with potential consequences for trust, clarity, and decision-making in future societies.

Keywords: hedging, pragmatic competence of LLMs, human-computer interaction, computer-human interaction

CONTENTS

I INTRODUCTION There are countless fascinating things about the human language, ranging from the scope of possibilities humans have evolved to express thought through the implications the linguistic choices we wind up making have on the meaning they convey. Specifically linguistic choices encoding uncertainty are somewhat of an otherworldly phenomenon because their purpose is not to communicate insight but rather the lack thereof. They describe informational voids that are crucial to most—if not all—decisions we make, from minor to life-changing choices, yet the ways they tweak language to convey these voids remain poorly understood.

The fact that we do not exactly understand how effectively distinct linguistic strategies mediate uncertainty in itself results in an incomplete picture of how humans may affect each other's decision-making via language. However, as we transition towards more hybrid societies, the need for clarity grows even further. We increasingly rely on systems powered by large language models (LLMs) to inform our decisions across domains, but we do not have the means to measure how well-calibrated these systems are to our internal models of risk. In other words, we currently have no clear way of knowing whether, when LLMs express uncertainty, they do so in ways we can intuitively understand.

The goal of this study is twofold. The first aim is to examine how linguistic hedging strategies impact human evaluation of reality: How are we using language to determine the extent of informational void? Are we responsive to the manner in which uncertainty is communicated (rather than its mere presence)? The second aim is to shed light on how these preferences reflect on what LLMs have been able to pick up on from us. Does their use of individual hedging strategies correspond to how humans interpret and respond to them? Through this project, the author aims to clarify these relationships in order to support more responsible use of LLMs in decision-making contexts.

II LITERATURE **Why and When We Speak Uncertain.** To tackle the problem of modelling uncertainty in human language, we first need to acknowledge that our judgements about truth and reality are a matter of degree. The notion of interpretive flexibility stems from the concept of so-called Fuzzy Logic (Zadeh, 1965). This theory proposes that while in abstract domains, such as maths, the difference between right and wrong may be clear-cut (e.g. it is definitely true that the number seven belongs in the set of primes), in everyday social contexts, judgments about whether something is true or belongs to a category are often socially ambiguous (e.g. most humans could more readily accept a raven as a bird compared to a penguin) (Vlasyan, 2019; Zadeh, 1965).

In order to convey various kinds of social ambiguities linguistically, natural languages implement hedging (Lakoff, 1973). The linguistic choices involved in hedging extend beyond syntax and grammatical dependencies. For one, the ways we hedge vary not only based on languages (e.g. Johansen, 2020) based on sociolinguistic factors such as gender (e.g. Du, 2021). Likewise, hedging serves a number of pragmatic functions such as conveying politeness (e.g. Liu, 2020) or avoiding conflict (e.g. Vlasyan, 2019). While acknowledging these broader functions, this study narrows its focus to how hedging shapes the epistemic value of statements. Still, recognising the above dimensions provides crucial context for interpreting the results of this study.

How We Hedge. In one of the most influential works on hedging, Prince et al. divides hedges into two main categories based on whether or not they change the truth value of the statement (1982). Approximators—which can be further divided into Adaptors (1-a) and Rounders (1-b)—change the meaning of the proposition, while Shields—which can be further subdivided into Plausibility Shields (1-c) and Attribution Shields (1-d)—do not directly change the value as much as they express the speaker's attitude. While the semantic distinctions are relatively clear, these hedges span across many part-of-speech categories and their syntactic realisation is inconsistent and often unpredictable, as shortly discussed.

(1) a. *The guests will come* ***somewhat*** *soon.*
b. *The guests will come* ***around*** *now.*
c. *The guests will* ***presumably*** *come soon.*
d. *The guests will come soon,* ***they say****.*

The problem with morpho-syntactic analysis of hedging is that "[t]here is no limit to the linguistic expressions that can be considered as hedges" (Clemen, 1997, cited in Kaltenböck et al., 2010, p. 23). Still, in *New Approaches to Hedging* (2010), Fraser proposes a revisited categorisation of hedging, where he outlines 19 mostly syntactically motivated constructions that may serve this purpose. Although this analysis is more grammatically comprehensive and provides greater syntactic nuance, it comes at the cost of specificity: it conflates polyfunctional constructions (e.g. negation, progressive tense, agentless passives (2)) with more prototypical hedging forms (e.g. modal verbs, epistemic verbs, and modal adverbs (3)), limiting the theory's applicability in quantitative analysis.

(2) a. ***Didn't*** *Harry leave?*
b. *I* ***am hoping*** *you will come.*
c. *Many of the troops* ***were*** *injured.*

(3) a. *John* ***might*** *leave now.*
b. *It* ***seems*** *that...*
c. *I can* ***possibly*** *do that.*

Measuring the Impact of Hedging. Conveying uncertainty is an essential pragmatic skill—not merely for signalling limits in our knowledge but also for supporting informed decision-making. Yet we do not quite understand how uncertainty travels between speakers—so much that, in some high-stakes contexts, we even chose to develop standardised number-to-word scales to help us communicate without misunderstandings (Dhami and Mandel, 2020). Although not linguistically informative, this intervention highlights a somewhat unsurprising yet important fact about the perception of hedging: it shifts with the stakes at play.

Efforts have been made to understand the impact of hedging on speech. For example, Naugle (2011) compares native and non-native speakers' perception of hedging on legal discourse and shows that both groups' impression of the strength of arguments changes—although to different extents—with frequency. These findings crucially suggest that frequency should be controlled for when investigating hedging, but they tell little about the role of hedging type.

Other studies propose that type matters, too. Vass (2015) argues based on corpus data that different types of hedges, as proposed by Prince et al. (1982), are found more appropriate in some contexts than others but does not account for the hedges' pragmatic function (i.e. expressing uncertainty, politeness, avoiding conflict, or other). Teigen and Brun (2003) show that the sentiment of the hedge impacts how uncertainty is perceived, where positive hedges (such as *likely* or *probably*) result in more optimistic predictions than negative hedges (such as *unlikely* or *uncertain*), but turns a blind eye to differences between hedging classes. More comprehensive research on context-dependent interpretations of various hedging strategies is missing.

Artificial Uncertainty. Now we've lived in a world where we can afford to be imprecise because we have been allowed to assume that our comprehenders have encountered the world in ways similar to our own—but this common ground is changing with technological progress. It is particularly relevant to preface this section by stating that making an artificial intelligence (AI) aware of her own uncertainty is considered a hard alignment problem: getting an AI to say `I don't know` if and only if she does, in fact, not know has even led to the formation of an entire branch of computer science inquiries (see research on *abstention*, e.g. Wen et al. (2025)).

That pulling uncertainty out of LLMs is like pulling teeth, however, does not mean that they cannot hedge per se.

AI does not have access to experience grounded in reality. She learns about our world by overhearing conversations we, humans, have about our experience of the world and relies on our sharing for building her own understanding of reality. Likewise, she does not (yet) simply run into an unknown concept or an unclear situation—unless we share our own uncertainty about it. (Which, by the way, does not mean she will not *generalise* based on the data provided, beyond the data provided (Wei et al., 2022).) That said, while spontaneous recognition of informational voids can be difficult, reporting on uncertainty is well within LLMs' abilities (Lin et al., 2022). This begs the question, are LLMs good at conveying uncertainty at least when *asked* to do so?

AI-Human Calibration. For a while now has the scientific community been interested in how what AI is thinking projects on the human understanding. For example, we know that humans are sceptical of AI authorship (Jia et al., 2024)—but we also know that this bias is mostly social and does not reflect the quality of the generated content in itself (Parshakov et al., 2025). Social factors such as ethos perception may change very quickly—or very slowly—depending on progress in the field as well as on exposure (remember people used to be afraid of trains once). Even as social attitudes evolve, questions on artificial cognition—though not permanent reference points either—offer insight into evaluating and guiding AI development.

A good illustrative case of such research related to the purpose of this project is that of Steyvers et al. (2025), which investigated how humans interpret AI's *confidence*. The results of this work describe a so-called Calibration Gap which stands for the difference between the models' internal confidence and the human perception of it. The research shows that humans—when unaware of the authorship—tend to heavily overestimate the models' internal confidence in this statement. These findings call for a deeper inquiry into how and why human users misread LLMs' expressions of informational voids. A more robust understanding of the nuance which impacts our perception of what LLMs are trying to tell us is so far missing.

III RESEARCH DIRECTIONS & QUESTIONS

The purpose of this project is to better understand how LLMs navigate uncertainty and how well their linguistic choices reflect the expectations of human comprehenders. To investigate the cause, the present study asks the following questions: (i) What, if any, differences are there in the weight human comprehenders assign to common hedging strategies in English? (ii) How accurately does LLM-generated output in response to uncertainty-inducing prompts reflect human perceptions of hedging? The first question will be answered through a quantitative experimental study with human subjects in Part I, which will provide a basis for answering the latter question through a quantitative analysis of LLM-generated language executed in Part II.

IV METHODOLOGY-PART I

Materials & Experimental Set-Up. To tackle the above questions, the present study first seeks to understand how various types of hedges impact human speakers' perception of reality. More specifically, it specifically investigates how hedging affects the perceived need for action. Given that the latter part of this study consists of a quantitative analysis of LLM output, this study selected four prototypical (rather than polyfunctional) hedging strategies to examine (based on the morpho-syntax-inspired analysis proposed in Fraser 2010). The final list of selected strategies, alongside representative examples, is outlined in Table 1. Table 2 shows a non-exhaustive list of lexemes in the selected categories and highlights those used in this study.

CATEGORY NAME	CODE	EXAMPLE
modal verbs	MOV	*might, could, should, would, can*
modal adverbs, modal adjectives	MOA	*possible, likely, generally, usually, seemingly*
quantifiers, rounders, approximators	QRA	*kind of, some, somewhat, around, in part*
epistemic phrases, mental verbs	EMV	*think, believe, suppose, assume, guess*

Table 1: The final set of hedging strategies used in the experiment, along with illustrative examples.

A relatively randomly selected representative was injected in a context sentence and juxtaposed with another randomly selected representative of another category in the same context. The author says "relatively randomly" because the grammatical structure of each sentence imposed some limits on what lexemes were feasible for a particular context. The survey included all combinations of hedging strategies for each of the 15 prepared contexts. The order in which the hedges were presented was likewise randomised. Altogether, the study comprised $15 \times \binom{4}{2}$ unique context pairings, resulting in 15×12 experimental items (as each context appears in two possible orders).

MODAL VERBS					
can	could	might	may	ought	should
would					
MODAL ADVERBS, MODAL ADJECTIVES					
possible/y	probable/y	likely	seeming/ly	theoretical/ly	ideal/ly
apparent/ly	presumable/y	potential/ly	virtual/ly	arguable/ly	typical/ly
general/ly	occasional/ly	frequent/ly	evident/ly	supposed/ly	usual/ly
random/ly	rare/ly	comparative/ly	relative/ly	approximate/ly	logical/ly
certain/ly	conceivably/ly	understandable/ly	inevitable/ly	undeniable/ly	
QUANTIFIERS, ROUNDERS, APPROXIMATORS					
some	somewhat	kind of	sort of	a bit	a few
several	around	about	approximately	more or less	nearly
almost	roughly	close to	in part	a number of	part of
many	much	certain	various	a couple of	a handful of
pretty much					
EPISTEMIC PHRASES, MENTAL VERBS					
think	believe	suppose	assume	guess	figure
suspect	estimate	reckon	imagine	consider	feel

Table 2: This table presents a non-exhaustive list of lexemes belonging to the selected hedging categories. The list also serves as a heat map, with lighter colours indicating lower frequency of the given lexeme in the experimental items.

Now each of the pairings were presented to the participants in a side-by-side setting with one multiple-choice question underneath, accompanied by a question prompting participants to select the more fitting speaker. The question of "Which of the above would make you more inclined to take action immediately?" was designed to give insight into the ACTIONABILITY of the examined hedges. See a full example experimental item in Figure 1.

Speaker A	Speaker B
Dinner is ready and the guests could arrive soon. We should set up the table.	Dinner is ready and the guests will likely arrive soon. We should set up the table.

Which of the above would make you more inclined to take action immediately?

○ Speaker A ○ Speaker B

Figure 1: Example experimental item.

As previously highlighted, the need for action obviously changes with the perceived seriousness of the situation (e.g. *we* ***may*** *have to take the bin out* hits different than *we* ***may*** *have to give him emergency treatment*, even though the uncertainty indicator appears equivalent). Therefore, this study has gone the extra mile and not only tested hedges in 15 diverse contexts but more specifically made sure to collect data for 5 low-, 5 medium-, and 5 high-stake contexts. The definitions of "low", "medium", and "high" stake are outlined in Table 3. Each context prompt consists of one hedged predicate and one action proposal. All 15 concrete prompts alongside their respective types are listed in Table 4. The final list of all 60 experimental items can be found in Table 23.

Type	Definition
low	non-urgent situations that should be addressed eventually but do not require immediate attention or action
medium	issues that may interfere with daily routines or tasks and could escalate if not managed in a timely manner
high	critical situations that pose an immediate risk to safety, well-being, or the functioning of essential systems, often affecting multiple individuals

Table 3: Definitions of low-, medium-, and high-impact contexts for the purposes of this study.

Type	Prompt
low	Dinner is ready, **and the guests arrive soon**. We need to set up the table.
low	**The milk in the fridge expires quickly**. We need to use it soon.
low	**Heavy rain starts during the picnic**. We should pack a tent.
low	**The gym gets crowded after work**. We should leave to get there earlier.
low	**The printer is running out of ink**. We need to order more.
med	**The stock will provide good returns next year**. We should consider buying it.
med	**The heating system is broken**. We should get it serviced before winter fully arrives.
med	**The test results indicate an issue**. We need to schedule follow-up tests.
med	**Our best candidate is considering another offer**. We should make our offer soon.
med	**Insurance covers only part of the procedure**. We need to budget for the remainder.
high	**This alarm system is malfunctioning during power outages**. We need backup protocols.
high	**The symptoms suggest an allergic reaction**. We should give emergency treatment now.
high	**The virus spreads through direct contact**. Quarantine protocols must begin immediately.
high	**The witness statement contradicts key evidence**. We must reassess our legal strategy.
high	**The merger will lead to major lay-offs**. We need to start job-hunting.

Table 4: Experimental items by the type of context. Parts in which the hedges get injected are highlighted in bold.

To verify that the author of this study does not merely label contexts as low-, medium-, or high-stake based on their private opinion, this study also utilised one of the filler questions to back up the validity of the author's assessments. This filler listed a randomly selected context (no hedge) and added that the call for action would not be fulfilled. The participants were asked to assess, on a scale of 0 to 10, to what extent the consequence from the inaction would impact their "comfort, safety, or routine". The results of this screening are outlined in Table 5 (for individual contexts) and Table 6 (for the average of all five in a group).

The increasing scores of low-, medium-, and high-stake contexts confirm the author's prior assumptions.[1] To show the statistical significance of these results, this study performed a Kruskal-Wallis test (Kruskal and Wallis, 1952; Choi et al., 2003) using the `stats` module of Python's `SciPy` library. This test considers raw values of two or more groups (in this case, assessments between 0 and 10 of three different context groups) and computes the probability that the values from each group come from the same distribution. In other words, it determines whether or not one group dominates over another and, if so, with what probability. Crucially, this test is non-parametric and does not assume normal distribution.

For the values of the three context pools, the $p-$value that they were drawn from the same distribution is < 0.001. The H-statistic, a value that measures the degree of difference between the average ranks of your different groups, was 151.498. Somewhat interestingly, each context pool showed decreasing standard deviation, meaning that the evaluations varied less as stakes increased. Likewise, the absolute skewness increased with stakes, with the original negative values suggesting that the distribution curve leaned towards the right end of the scale with a more pronounced left tail. All data is reported in Table 6.

PROMPT	AVG
low-1	5.913
low-2	5.236
low-3	7.776
low-4	5.221
low-5	4.141

(a) low

PROMPT	AVG
med-1	4.570
med-2	8.100
med-3	7.658
med-4	5.670
med-5	6.886

(b) med

PROMPT	AVG
high-1	7.370
high-2	7.929
high-3	7.817
high-4	7.616
high-5	8.090

(c) high

Table 5: Average perceived impact of inaction on participants' comfort, safety, or routine across three stake levels.

STAKE	AVG	STD	MED	SKEW
low	5.609	2.832	6	-0.483
med	6.525	2.351	7	-0.812
high	7.744	1.805	8	-0.995

Table 6: This table shows the average rating for a whole group of contexts (pre-labelled as low, medium, high). Note that these values are not equivalent to the unweighted averages of values in Table 5 because the randomised distribution of questions and subsequent filtering meant that some contexts resulted in a different number of responses than others.

The finalised survey consisted of a brief informed consent and a total of 7 questions: a randomly selected item out of the low-stake, medium-stake, and high-stake pools each, two fillers, and

[1] The author is also fully aware that 5.609/10 is quite a high score to rate a context as "low-impact", but they would point the reader's attention to the fact that this same sample would mind a soon-to-be-spoiled milk in their fridge at 5.236/10. To the participants' defence, they were only presented with a singular context for this assessment and had no additional points of reference, by design. Therefore, it can be reasonably assumed that the rising average is enough evidence for the author's initial subjective hypothesis to hold.

two attention checks. The prompt from each of the above pools as well as the specific hedges compared were randomised algorithmically. The survey was constructed using Qualtrics XM due to the availability of advanced randomisation features.

Participants. This study recorded 1,362 unique responses to the above survey. All participants were recruited using the Prolific platform, the pre-screening included the requirement that English was the participants' first language. The vast majority of participants reside in the United Kingdom (76.74%) and the rest of the participants (23.36%) reside in the United States, see Figure 2. The average age of participants was 43.32 ($\sigma = 14.03$). 40.79% of the sample identified as female and 58.69% as male, with data for the remaining 0.52% unavailable. Sex distribution of the sample is shown in Figure 3.

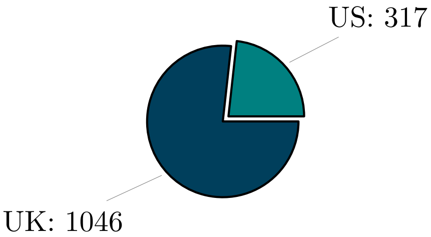

Figure 2: Place of residence distribution.

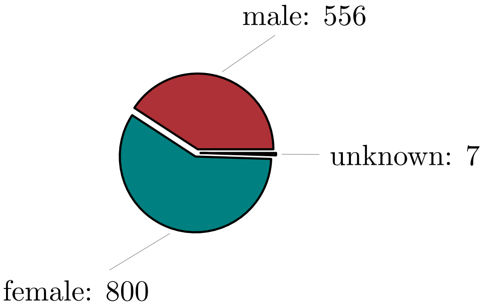

Figure 3: Sex distribution of participants.

Analysis Tools. Recall that the intention of this section is to investigate how different hedging categories compare with one another. An attentive reader may have noticed, however, that all experimental items (as described above) only juxtapose two specific hedges, never all four categories investigated. This is because this study chooses an approach that has long helped researchers avoid overwhelming participants with long surveys: instead, it is possible to record numerous pairwise comparisons and apply the BRADLEY-TERRY MODEL (BRADLEY and TERRY, 1952) to extract the desired results.

The Bradley-Terry Model (BTM) is a statistical tool that applies conditional probability to understand how any two items in a set relate to one another in terms of dominance. More specifically, according to the BTM, any pair of items i, j drawn from a sample S can be defined by a specific relation shown in Equation 1.[2] A non-maths enthusiast reader should not despair (yet): the below relation very intuitively displays that the probability of i ranking higher than j (on a particular axis) equals to the observed number of times i dominates over j, p_i, over the overall number of observed games (in this case, comparison judgements), $p_i + p_j$.

$$p(i \succ j) = \frac{p_i}{p_i + p_j} \tag{1}$$

The property of BTM particularly useful to this analysis is that it can be aggregated. In other words, BTM can not only help us *compare* any two elements $i, j \in S$ but also *order* any list of elements $i, ..., j \in S$. The generalised formulation of BTM is called the PLACKETT-LUCE MODEL (LUCE, 1979; PLACKETT, 1975) and is described by Equation 2. Once again, while this equation may initially appear somewhat monstrous, it essentially formalises the idea that the probability of an element x_i being better than x_j depends on

[2] The funny "greater than"-like sign in Equation 1 reads "succeeds" and it is a common notation in ranking and decision theories to denote that one element is ranked or preferred over another.

$$p(x_1 \succ x_2 \succ \ldots \succ x_n) = \frac{p_{x_1}}{p_{x_1} + p_{x_2} + \ldots + p_{x_n}} \times \frac{p_{x_2}}{p_{x_2} + p_{x_3} + \ldots + p_{x_n}} \times \ldots \times \frac{p_{x_n}}{p_{x_n}} \quad (2)$$

Note that according to this model, an example ordering x_1, x_2 would not have the same probability as, for instance, x_2, x_1 even though it is a multiplication, because the denominator for the term x_n only includes games not processed in one of the previous terms. This means that the permutation of items in S with the greatest computed probability represents the most likely order of elements in S. An efficient and fast algorithmic inference of this model was designed and described relatively recently in Maystre and Grossglauser (2015).

V ANALYSIS & RESULTS-PART I. Of the 1,362 unique participants who completed the study, 1,321 passed both attention checks (96.989%). After excluding the noise, this study tallied up the number of times hedge type x was selected as preferred to hedge type y by the participants and modelled the numbers in game matrices separately (Matrices 7a, 7b, 7c) by context as well as altogether (Matrix 7d). Any value `[i][j]` means that the column label `i` lost against the row label `j`. In each context, the participants evaluated 1,321 games, totalling 3,963 games altogether. Every pair played an average of 44.03 games per prompt and 220.17 games per context.

	MOV	MOA	QRA	EMV
MOV	-	119	65	134
MOA	107	-	105	166
QRA	141	115	-	141
EMV	80	80	68	-

(a) Low

	MOV	MOA	QRA	EMV
MOV	-	119	121	85
MOA	95	-	119	100
QRA	126	98	-	110
EMV	125	107	116	-

(b) Medium

	MOV	MOA	QRA	EMV
MOV	-	106	91	134
MOA	128	-	124	147
QRA	127	102	-	143
EMV	78	57	84	-

(c) High

	MOV	MOA	QRA	EMV
MOV	-	334	277	353
MOA	330	-	348	413
QRA	394	315	-	394
EMV	283	244	268	-

(d) Cross-Context

Table 7: Pairwise win matrices for different context strengths.

All observed data were tested against the null hypothesis. The null hypothesis, in this case, states that in any given game, both players have an equal, uniformly distributed chance to win. In other words, any two hedging strategies compared are mutually interchangeable without any impact on the actionability of the presented context. Each pair was subject to the two-tailed Binomial Test to see how likely it would be to obtain the observed balance under the null hypothesis. This analysis was performed using Python's `SciPy` library, where the `binomtest` function performed the test according to the formula in Equation 3.

Now the author will let the non-maths enthusiast reader despair a little but also clarify that the below function merely iterates through every i in the set $\mathcal{I}$, representing instances lesser than the observed value, and tallies up the probabilities that i victories occur under the null hypothesis p_0. These probabilities consist of a product of the number of ways to select i victories out of n games, $\binom{n}{i}$, the probability of i victories under the null hypothesis, p_0^i, and the probability of $n - i$ losses under the null hypothesis, where $(1 - p_0)$ is the probability of not winning. In this case, the probabilities of winning and losing are equivalent, $p_0 = (1 - p_0) = 0.5$. The final number represents the probability that one player wins *at least* the observed number of times given p_0. Every possible pair (disregarding order) was subject to the above analysis. The results are represented in Table 8.

$$p = \sum_{i \in \mathcal{I}} \binom{n}{i} \times p_0^i \times (1 - p_0)^{n-i} \tag{3}$$

In addition, this study also fits the values from Table 7d onto the maximum likelihood estimate model using the iterative Luce Spectral Ranking algorithm (Maystre and Grossglauser, 2015), the purpose of which has been outlined in the previous section. Recall that this algorithm returns a list of estimated model parameters, where each parameter (in this case, a particular hedging strategy) receives a value that estimates how it relates to all other parameters. This estimate was computed using Python's `choix` library, which converts numbers to log-space for analysis. Table 9 outlines this order relation for parameters described by matrix Table 7d.

Relation	p-Value	Sig
MOV-MOA	0.907	F
MOV-QRA	7.161×10^{-6}	T
MOV-EMV	0.006	T
MOA-QRA	0.214	F
MOA-EMV	4.397×10^{-11}	T
QRA-EMV	1.359×10^{-11}	T

Table 8: Probability that the observed values were generated under the null hypothesis (i.e. that the hedges are freely interchangeable and any hedge has a 50% probability to beat any other hedge).

Strategy	Score
QRA	0.160
MOA	0.144
MOV	-0.021
EMV	-0.283

Table 9: Maximum likelihood estimate of model parameters computed using the Iterative Luce Spectral Ranking (I-LSR). The difference between any two values represents the probability of a human preferring the higher-ranked value over the lower-ranked value in log-space values.

Table 9 outlines the estimated overall order of the model's parameters (MOV, MOA, QRA, EMV). Table 8 shows the statistical significance values of the results between each pair. Even though not all data in Table 8 is statistically significant, combining the outcomes from both tables, it is still possible to propose a model where any bracketed part has a statistically insignificant difference and may be freely interchangeable on the measured axis. Note that even though QRA ranked over MOA in Table 9, the difference between these two is statistically insignificant as shown in Table 8. However, the difference between QRA and MOV is statistically significant and thus the freely interchangeable group {QRA, MOA} can be placed above {MOV}, as shown in Equation 4

$$\{\mathrm{QRA}, \mathrm{MOA}\} \succ \{\mathrm{MOV}\} \succ \{\mathrm{EMV}\} \tag{4}$$

Furthermore, given that `choix` outputs log-space score values, values in Table 9 can be converted from the log-space values to more easily interpretable probabilities using the formula in Equation 5. This formula states that for any items i, j in a predefined set S, the probability of i preceding j in the maximum likelihood order of items in S is defined by the exponential function outlined in Equation 5. The outcome of this transformation is outlined in Table 10. Note that for any $i, j \in S$, $p(i \succ j) + p(j \succ i) = 1$.

$$p(i \succ j) = \frac{e^i}{e^i + e^j} \tag{5}$$

	MOV	MOA	QRA	EMV
MOV	-	50.404%	45.494%	60.889%
MOA	49.596%	-	54.105%	60.504%
QRA	54.506%	45.895%	-	56.511%
EMV	39.111%	39.496%	43.489%	-

Table 10: Log-space BTM scores converted into decimal probabilities.

It is also relevant to point out that the win rates were not consistent across contexts. For example, even though EMV ranks significantly below QRA in the final ranking in Table 9, it in fact beats QRA in one of the contexts in raw data, see Table 7b. Recall that the raw data is already an aggregation of numbers from across five different contexts. By looking at the separated data, there is no obvious pattern to why the outcomes of data in Table 7b do not correspond to the final order. Fortunately, the BTM accounts for this data as well. The final order in Table 9 displays the final scores *with regard to* the fact that a sizeable chunk of the games between QRA and EMV, QRA lost. This factor is reflected in a decreased difference between these two players. Even still, QRA beats EMV in other contexts harder than it gets beaten in Table 7b, which is why it tops the rank in Table 9.

VI METHODOLOGY-PART II All the way up until now, this paper has explored how humans understand hedging, but recall that a significant part of the motivation for this study comes from the need to understand the interaction between humans and machines. The second question of this study asks how human impressions of hedging reflect on LLM-generated output. The expectation would be that if LLMs pick up on the preferences for how humans interpret caution in language, LLMs would reflect similar shifts in the use of hedging strategies across contexts of varying consequence.

Spontaneous Hedging Production. The first step to measure whether any adjustments take place is to obtain a selection of appropriate prompts. Now while prompts in the previous part were carefully hand-crafted to fit a specific purpose, the idea behind the prompts in this latter part is to stimulate the spontaneous production of hedging instead. To collect relevant information, this study maintained a few invariants to the prompts engineered for this part: (a) put the model in the role of an informator, (b) specify the source of uncertainty, and (c) imply that a possible decision or action needs to be taken. All the above invariants directly relate to the motivation of LLM-based agents informing human decisions with imperfect information.

Clearly, the role of an informator is a wide term that must be carefully defined if this study is to evaluate LLM performance with any degree of rigour. This study attempted to contain the variety of scenarios by always including an equal ratio of casual to professional to service interaction contexts, in which this study attempts to simulate real-world conditions under which LLMs might eventually be integrated to assist with decisions under uncertainty. 90 prompts across low-, medium-, and high-stake contexts were selected based on the definitions in Table 3. All used prompts are listed in Tables 20, 21, 22.

It was essential that none of the uncertainties identified in the prompts relate to objectively known facts. This is because the study sought to isolate the model's reasoning under ambiguity, rather than its ability to retrieve facts. In other words, ambiguity not grounded in objectively known facts rightly limits the LLMs' ability to verify uncertainty by browsing the Internet or databases. Predictive (modelling futures with incomplete information) and interpretative (multiple plausible interpretations) sources of ambiguity were preferred.

Models. This study examined four of the most widely used LLMs: (a) `GPT-4o` (OpenAI), (b) `Claude 3.7 Sonnet` (Anthropic), (c) `Gemini 2.0 Flash` (Google DeepMind), and

(d) `Llama 3.3` (Meta). The answers to the pre-defined prompts were collected through calls to the providers' APIs in approximately the same time, with the context window cleared after each prompt. Any information about token counts used throughout the study was obtained through each model's own tokenisers (GPT-4o, Claude Sonnet 3.7, Gemini 2.0, Llama 3.3).

VII ANALYSIS & RESULTS-PART II All data collected from the models were recorded in a plain text file. The token counts were obtained from a copy of the original data before any preprocessing took place and are shown in Figure 6. Preprocessing consisted of turning all data lowercase and stripping it of extra spaces and punctuation using Python's `regex` library. The data was compared to the admittedly non-exhaustive yet appropriately extensive list of lexical items of each examined category outlined in Table 2. Relevant hedge counts were normalised (per 100 tokens) and plotted per context, as shown in Figure 4. It is worth pointing out that while token counts across low-, medium-, and high-stake contexts reliably increase (Figure 4), (normalised) hedge counts decrease (Figure 6).

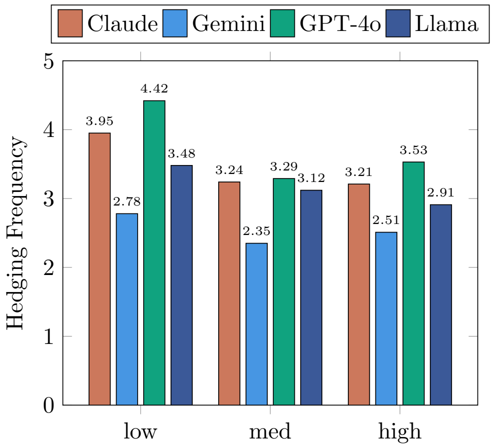

Figure 4: Number of hedging tokens normalised per 100 tokens.

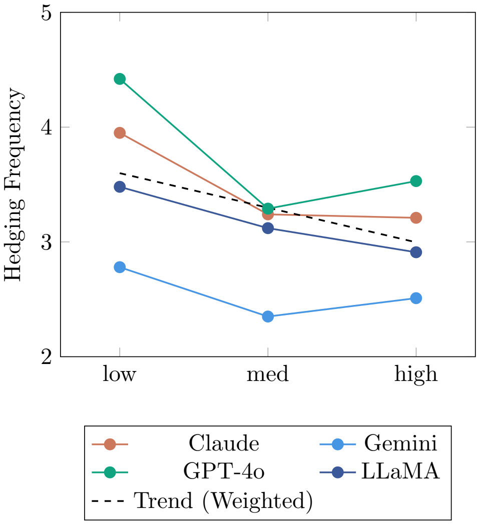

Figure 5: Hedging frequency across stake levels with weighted trendline.

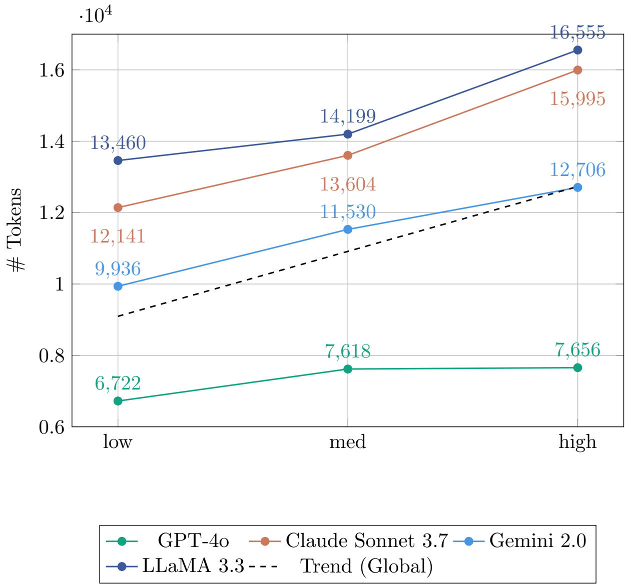

Figure 6: Number of tokens generated across raising context stakes.

A more detailed analysis of the hedges was performed. Each category was tallied up separately and plotted on a separate graph for each model, see Figure 7. While this study observed no dramatic variation between values across context types in this experiment, it did record a global preference for the generation of modal adverbials and adjectives (MOA) and modal verbs (MOV) over any other examined hedging type. The observed global production of quantifiers, rounders, and approximators (QRA) as well as epistemic and mental verbs (EMV) was negligible.

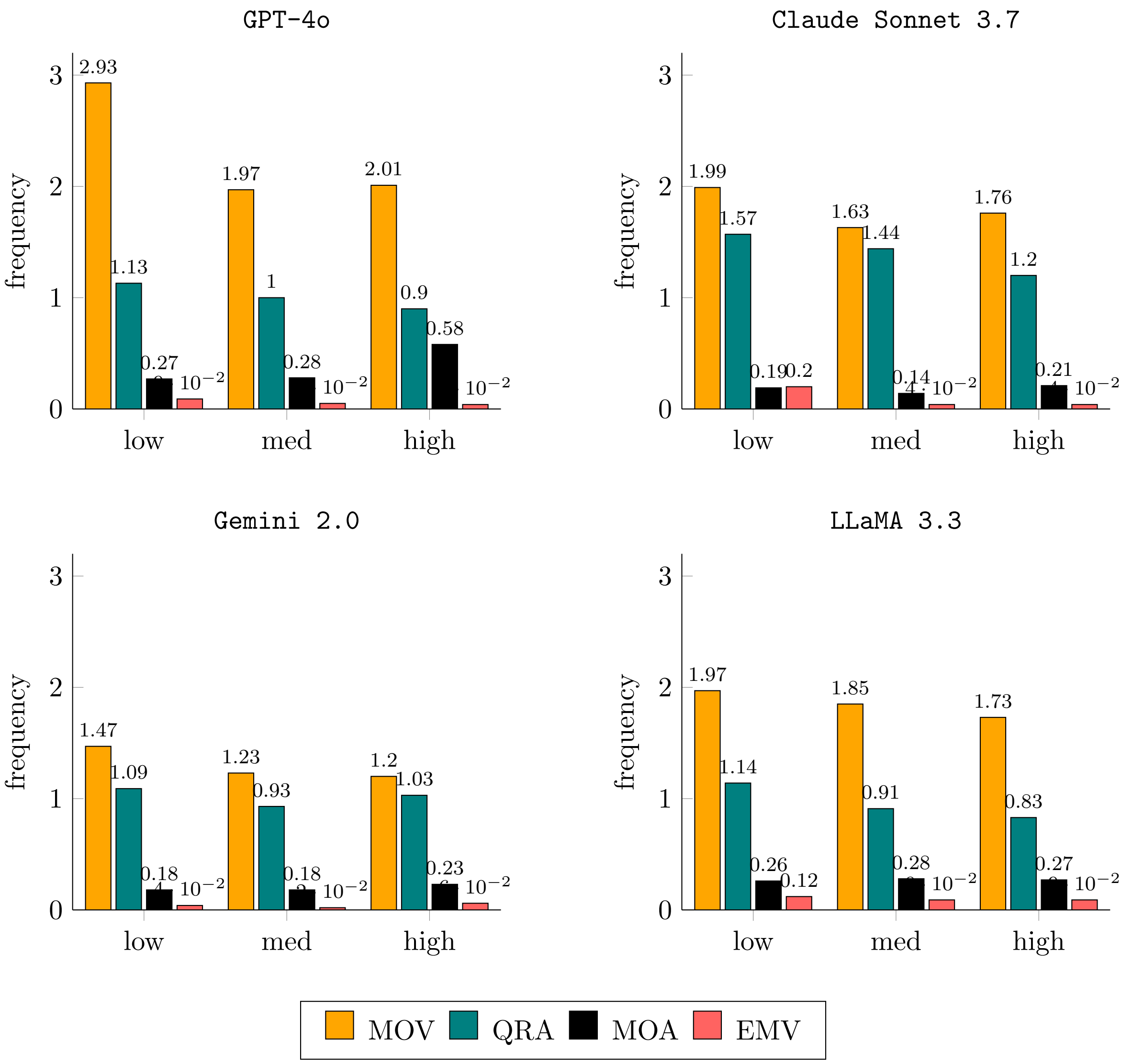

Figure 7: Normalized hedge-type frequencies per 100 tokens, by model and context stake.

An additional analysis was performed using n-grams, where each of the collected samples was also scanned for the ten most common uni-, bi-, tri-, and four-grams using Python's `collections` library. Here, one-, two-, three-, and four-word phrases in each dataset were extracted and counted. Tables 11, 12, 13, and 14 show the 10 most frequent $n-$grams for each model tested (minus bare pronouns and monofunctional prepositions). Even though it is beyond the scope of this thesis to analyse the details of other common hedging strategies LLMs may use outside the narrowed range in Table 1, it is still curious to see the differences and overlaps across the models' lexicons. Notice that most phrases listed are directly or indirectly related to hedging.

IDX	UNIGRAM	COUNT	BIGRAM	COUNT	TRIGRAM	COUNT	FOURGRAM	COUNT
0	if	2.090	would you	0.453	would you like	0.314	would you like me	0.302
1	or	1.042	sure if	0.432	like me to	0.308	id be happy to	0.095
2	not	0.989	about the	0.343	not sure if	0.246	want to make sure	0.089
3	about	0.980	want to	0.334	but im not	0.166	not entirely sure if	0.083
4	but	0.767	like me	0.308	sure if the	0.145	not sure if the	0.071
5	have	0.719	if the	0.302	not entirely sure	0.118	youre trying to decide	0.062
6	sure	0.707	if you	0.293	might be worth	0.167	let me know what	0.056
7	would	0.654	need to	0.284	let me know	0.161	but im not sure	0.053
8	might	0.598	not sure	0.272	might want to	0.098	would you prefer to	0.053
9	can	0.598	but i	0.210	id be happy	0.095	might be worth checking	0.050

Table 11: Most frequent $n-$grams spontaneously produced by `Claude 3.7`. Normalised per 100 tokens.

IDX	UNIGRAM	COUNT	BIGRAM	COUNT	TRIGRAM	COUNT	FOURGRAM	COUNT
0	if	1.828	want to	0.624	just want to	0.250	want to make sure	0.179
1	just	1.418	let me	0.397	want to make	0.195	just want to make	0.103
2	sure	1.029	okay so	0.389	not sure if	0.187	okay heres what id	0.103
3	okay	0.997	not sure	0.302	okay let me	0.123	heres what id say	0.103
4	not	0.787	sure if	0.286	okay heres what	0.115	just want to be	0.064
5	can	0.779	make sure	0.286	heres what id	0.103	dont want you to	0.060
6	or	0.771	just want	0.258	what id say	0.103	better safe than sorry	0.056
7	want	0.747	if you	0.242	want to be	0.103	see if i can	0.052
8	but	0.727	about the	0.207	make sure were	0.099	can we just doublecheck	0.052
9	about	0.679	if its	0.199	just in case	0.095	not 100 sure if	0.048

Table 12: Most frequent $n-$grams spontaneously produced by `Gemini 2.0`. Normalised per 100 tokens.

IDX	UNIGRAM	COUNT	BIGRAM	COUNT	TRIGRAM	COUNT	FOURGRAM	COUNT
0	if	2.295	might be	0.657	might be a	0.280	might be a good	0.265
1	or	1.323	if the	0.560	good idea to	0.275	what do you think	0.143
2	might	0.926	sure if	0.361	just in case	0.254	not entirely sure if	0.127
3	can	0.830	let me	0.351	let me know	0.254	reaching out to the	0.102
4	any	0.728	good idea	0.305	not sure if	0.219	let me know if	0.097
5	recommend	0.657	if you	0.275	sure if the	0.173	not sure if the	0.097
6	about	0.621	idea to	0.275	out to the	0.163	get back to you	0.087
7	not	0.616	not sure	0.275	do you think	0.158	make an informed decision	0.081
8	have	0.585	about the	0.270	might be worth	0.158	might be best to	0.076
9	check	0.580	just in	0.254	not entirely sure	0.143	do you think hey	0.071

Table 13: Most frequent $n-$grams spontaneously produced by `GPT-4o`. Normalised per 100 tokens.

IDX	UNIGRAM	COUNT	BIGRAM	COUNT	TRIGRAM	COUNT	FOURGRAM	COUNT
0	`if`	1.966	`want to`	0.879	`not sure if`	0.358	`want to make sure`	0.262
1	`can`	1.294	`sure if`	0.527	`want to make`	0.265	`let me just check`	0.181
2	`not`	1.189	`not sure`	0.427	`let me just`	0.232	`would you like me`	0.132
3	`sure`	1.081	`make sure`	0.397	`dont want to`	0.190	`but i want to`	0.117
4	`or`	1.023	`let me`	0.343	`would you like`	0.184	`not entirely sure if`	0.098
5	`want`	1.023	`can you`	0.307	`like me to`	0.132	`do you want to`	0.078
6	`but`	0.989	`dont want`	0.298	`but im not`	0.120	`take a closer look`	0.078
7	`just`	0.828	`can we`	0.289	`but i want`	0.117	`but im not sure`	0.075
8	`check`	0.734	`check the`	0.280	`not entirely sure`	0.114	`but at the same`	0.069
9	`have`	0.686	`but i`	0.274	`do you think`	0.102	`at the same time`	0.069

Table 14: Most frequent $n-$grams spontaneously produced by `Llama 3.3`. Normalised per 100 tokens.

VIII CLOSING THE GAP An attentive reader may have noticed that the present study has measured how human comprehenders *perceive* hedging and how LLMs *produce* hedging. This is primarily because the direction of the interaction of interest has been *from* machines *to* humans. However, it shall not be forgotten that LLMs do not produce language in a vacuum. They learn to produce language—including that of uncertainty—from us. To complete the puzzle of understanding the human-computer interaction on uncertainty, it is hence also enriching to conduct a brief survey of LLM preferences in a set-up similar to that of the experiment in Part I.

Hedge Actionability Preference Survey. With that in mind, this study also conducted a brief survey of each examined LLM, where the models were asked the same questions as human participants in Part I. Each tested LLM was presented with all combinations of the selected hedges in the exact same prompts as were presented to human participants (see Table 4). Each model was asked about the same combination 5 times to account for potential variability in generation. All requests were made through separate calls to prevent context bleeding. Recall that the experimental items included 4 hedges injected in 5 prompts across 3 contexts. Given that every hedge i faced another hedge j in the comparison set-up 5 times per prompt (due to repetitions) in 10 different set-ups (5 prompts times 2 possible positions), a total of 50 responses were collected for any pair.

Before diving deeper, the win rates were computed for all models separately. These were simple computations of how many times a hedge i was preferred to hedge j over the total number of games between i and j without additional adjustments. This means that any reciprocal entries in the output matrices, `[i][j]` and `[j][i]`, add up to 1 (as one of i, j must have won). These values are summarised in Table 15.

	MOV	MOA	QRA	EMV
MOV	-	55.33%	34.67%	54.00%
MOA	44.67%	-	39.33%	50.00%
QRA	65.33%	60.67%	-	58.67%
EMV	46.00%	50.00%	41.33%	-

(a) Claude 3.7

	MOV	MOA	QRA	EMV
MOV	-	50.00%	43.33%	53.33%
MOA	50.00%	-	45.33%	52.00%
QRA	56.67%	54.67%	-	50.67%
EMV	46.67%	48.00%	49.33%	-

(b) Gemini 2.0

	MOV	MOA	QRA	EMV
MOV	-	54.67%	45.33%	62.00%
MOA	45.33%	-	51.33%	62.00%
QRA	54.67%	48.67%	-	52.67%
EMV	38.00%	38.00%	47.33%	-

(c) GPT-4o

	MOV	MOA	QRA	EMV
MOV	-	50.00%	42.00%	45.33%
MOA	50.00%	-	46.67%	49.33%
QRA	58.00%	53.33%	-	55.33%
EMV	54.67%	50.67%	44.67%	-

(d) LLaMA 3.3

Table 15: Pairwise win rate matrices for different models.

As previously suggested, however, these values by themselves are of little interest. Their value lies in the similarity of these entries to the matrix of human preferences. To analyse the above data, this study first computed a win rate matrix of human preferences across contexts (initially outlined in Table 7d). The result of the calculations is summarised in Table 16.

	MOV	MOA	QRA	EMV
MOV	-	50.303%	41.284%	55.502%
MOA	49.697%	-	52.490%	62.859%
QRA	58.716%	47.510%	-	59.517%
EMV	44.498%	37.141%	40.483%	-

Table 16: Reference matrix of human preferences.

Normalised difference was used to understand the values of LLMs' preferences (outlined in Table 15) in the context of human preferences (outlined in Table 16). This metric, computed using Equation 6, uses an absolute difference between any two win rates and scales them over their sum. The results show the difference relative to the magnitude of the values, with 0 meaning that the probabilities are identical and values approaching one indicating the largest possible difference. The relative differences for each model are outlined in Figure 8. A parallel computation is performed for relative differences for the models combined between different stake levels. The computed win rates are outlined in Table 17. The relative differences are shown in Figure 9.

$$\text{normalised difference} = \frac{|p_{LLM} - p_{human}|}{(p_{LLM} + p_{human})} \tag{6}$$

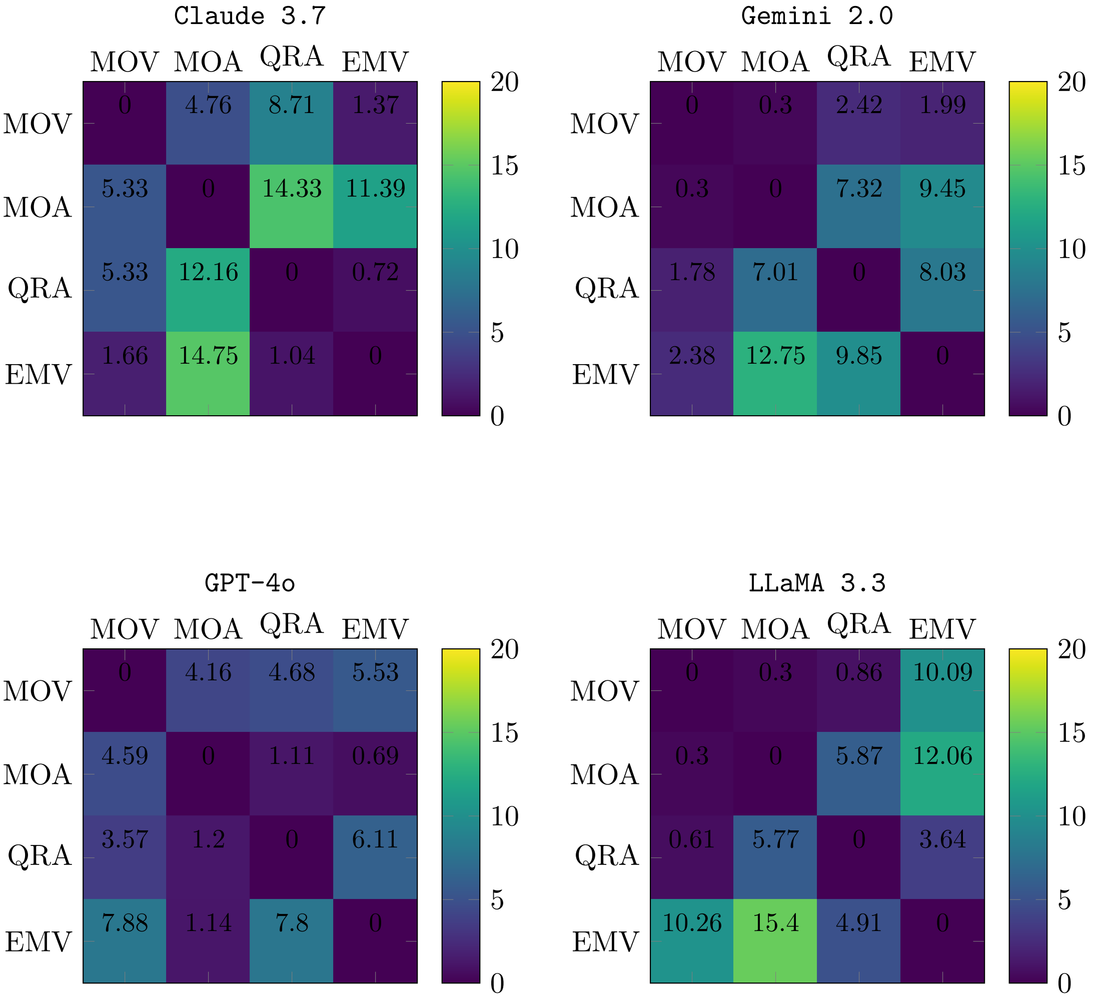

Figure 8: Pairwise preference matrices for all stakes combined for Claude 3.7, Gemini 2.0, GPT-4o, and LLaMA 3.3.

	MOV	MOA	QRA	EMV
MOV	-	46.50%	35.00%	50.50%
MOA	53.50%	-	36.50%	51.00%
QRA	65.00%	63.50%	-	61.50%
EMV	49.50%	49.00%	38.50%	-

(a) Low

	MOV	MOA	QRA	EMV
MOV	-	67.00%	43.00%	51.50%
MOA	33.00%	-	44.00%	39.00%
QRA	57.00%	56.00%	-	47.50%
EMV	48.50%	61.00%	52.50%	-

(b) Medium

	MOV	MOA	QRA	EMV
MOV	-	44.00%	46.00%	59.00%
MOA	56.00%	-	56.50%	70.00%
QRA	54.00%	43.50%	-	54.00%
EMV	41.00%	30.00%	46.00%	-

(c) High

	MOV	MOA	QRA	EMV
MOV	-	52.50%	41.33%	53.67%
MOA	47.50%	-	45.67%	53.33%
QRA	58.67%	54.33%	-	54.33%
EMV	46.33%	46.67%	45.67%	-

(d) All

Table 17: Pairwise win rate matrices for different context strengths.

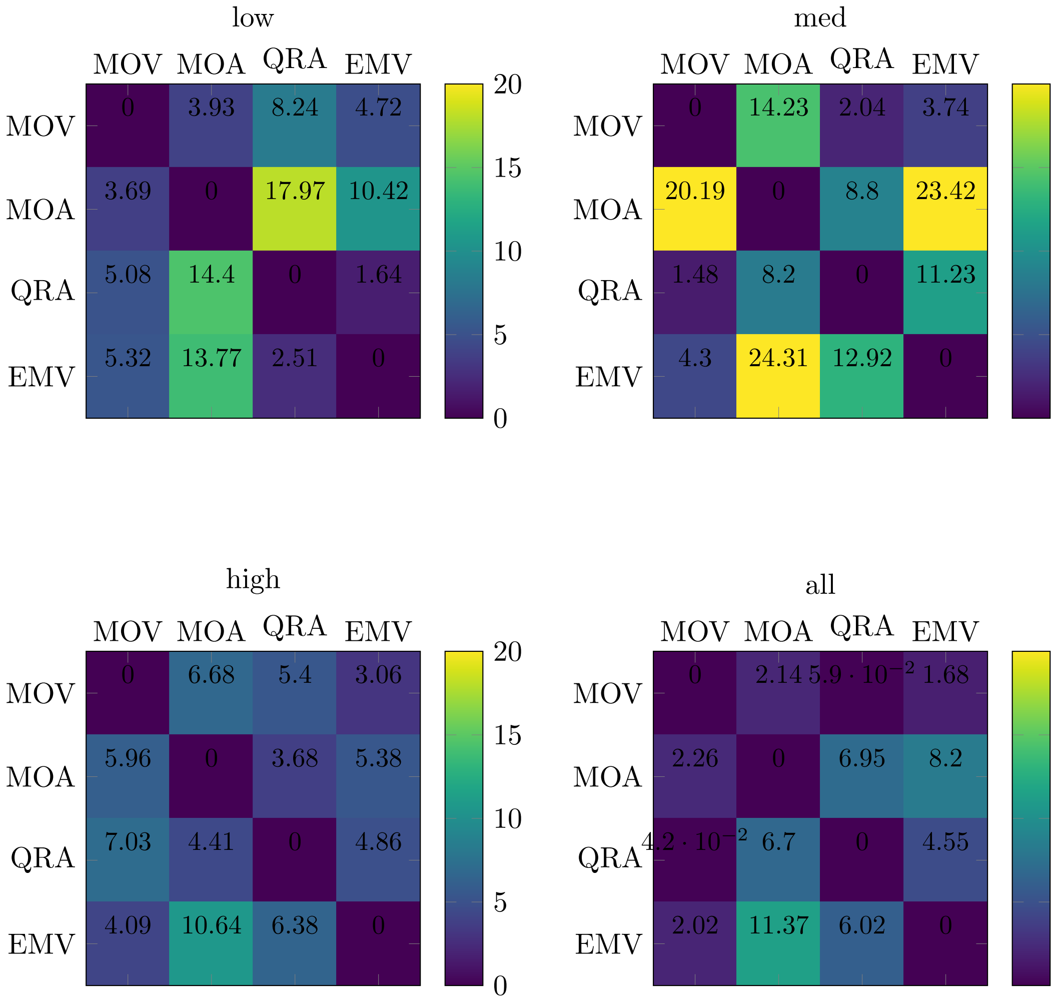

Figure 9: Aggregated pairwise preference matrices for all models by stake level: low, med, high, and all stake levels combined.

The above confusion matrices show how different the processed LLM preferences were from human-evaluated data. To read these matrices correctly, one should understand that small values mean that the win rate distribution was identical between human and LLM data. For example, if two hedges i, j were considered as highly competitive and yielded an outcome of 50% each by humans, then about the same perceived win rate assigned by LLMs will result in dark purple values in the above matrices. On the other side, if the win rate is close to 50% for each player but there is a strictly dominant player identified in human data, then the value will shine more yellow.

Additionally, Figures 8 and 9 compare LLMs' potentially biased preferences to human data. However, how biased these preferences are compared to choices drawn from the binomial distribution at random is not obvious. To add this extra comparison layer, this study ran 10,000 simulations with matrices filled with values drawn from the binomial distribution at random and computed a normalised difference between these simulation matrices and human data. Figure 10 shows, for comparison, what the confusion matrix would look like if the data were drawn from a random binomial distribution (i.e. if the models selected one player over another uniformly at random) and compared to human evaluations. The larger a confusion value `[i][j]` is in Figures 8 or 9 compared to its counterpart in 8, the more we can speak of systematic bias.

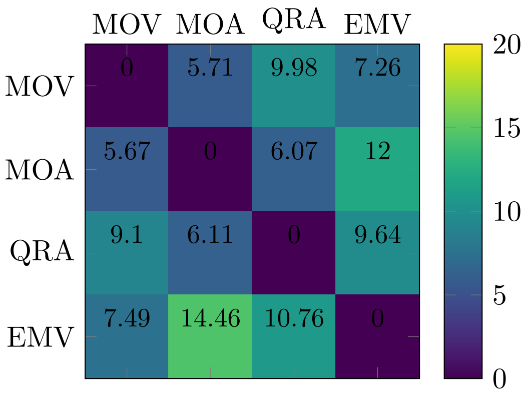

Figure 10: Confusion matrix of random binomial data, normalised.

Additionally, this study used the raw values to compute the maximum likelihood estimate of model parameters was computed using Python's `choix` library (similar to Table 9). All scores were compared to the findings from Part I using the cosine similarity formula, where a value of 1 indicates that the output vectors are identical, whereas -1 indicates that the vectors point in the exact opposite directions. Table 18 shows how the maximum likelihood estimate for each model across contexts relates to the maximum likelihood estimate identified by human participants. No consistent similarities were identified, but most models show positive (i.e. above random) similarities in preferences across contexts.

	low	med	high	all
`Claude 3.7`	0.533	-0.327	0.777	0.480
`Gemini 2.0`	0.160	-0.643	0.796	0.658
`GPT-4o`	0.970	-0.783	0.883	0.859
`Llama 3.3`	-0.251	0.193	0.375	0.311
all	0.550	-0.490	0.865	0.735

Table 18: Cosine similarity between LLM assessment and human preferences (see Table 9). Values closer to 1 indicate greater similarity, whereas values closer to -1 indicate perfectly inverse values.

Human Hedge Production. This study also conducted a brief corpus analysis to determine the frequency with which speakers produce individual hedging strategies in everyday speech. The data used to perform this analysis came from the Corpus of Contemporary American English (COCA) (Davies, 2009), which consists of over 17 million n-grams of language produced between the years 1990 and 2019. All n-grams were searched for tokens corresponding to lexical items in Table 2, used throughout the previous analyses in this study. The search yielded 1,858 results, which were subsequently manually filtered for false positives (e.g. *about* in the n-gram *talked about* was excluded for it is not a valid hedge).

After the data processing, 1,506 n-grams remained. The n-grams were grouped by the hedging strategies, and their respective counts were tallied up. Figure 11 displays the final aggregated frequencies of the investigated categories. Even though a significantly higher use of MOV hedges is obvious from the data, no other obvious correlations are found with human preferences or LLM data—except perhaps that MOV ranks highest (as observed in LLM spontaneous production) and that EMV comes last (as observed in LLM production as well as human preference). Crucially, the order or distribution does not correspond to production data observed in Figure 7. Please note that this summary is illustrative and is unlikely to reflect the actual distribution used to train LLMs per se. This data may not generalise on language intended to communicate uncertainty.

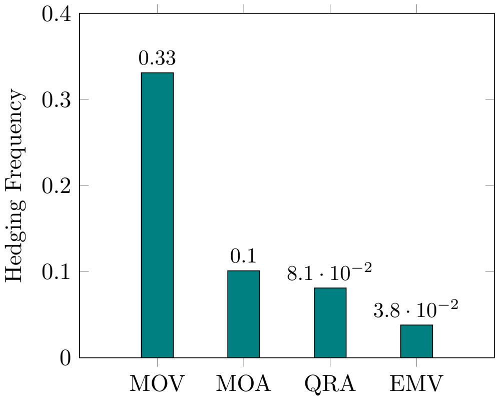

Figure 11: Number of hedging tokens normalised per 100 tokens.

IX DISCUSSION In Part I, this study explored the human perception of hedging. The findings from this part propose that not all hedging strategies, when split by grammatical category, are born equal. More specifically, a statistically significant hierarchy was found and formally described in Equation 4. These results propose that it is not only the semantic but also the grammatical function of various hedging strategies that makes a difference to how sentences containing these hedging strategies are perceived. Concurrently, the findings add another dimension to our understanding of hedging: literature has so far understood that the usage of hedging requires very fine pragmatic skills, but these findings suggest that syntactic skills complement pragmatics in conveying uncertainty beyond merely constructing grammatical structures.

In Part II, this study also explored how LLMs hedge across contexts, and whether the hedging strategies they use at all reflect human perception of actionability as stakes change. Given the results from Part I, it could be expected that if LLMs are well calibrated to human understanding, higher-ranked hedging strategies (i.e. more actionable expressions) would be more readily available in low-stake contexts and vice versa. This is because in high-stake contexts, any decisions are typically more consequential, which demands that any decision-making process be less imposing, more human-controlled, and transparent. However, this pattern was not confirmed.

This study additionally observed some interesting phenomena outside the experiment's intent. First, all tested LLMs show a reliable increase in the number of tokens produced as the stakes rise. At the same time, the frequency of hedging decreases as stakes rise. This combination of findings is particularly concerning in the context of Steyvers et al. (2025), which shows that human comprehenders tend to consider longer responses more reliable by default, even if the extra length does not bring any relevant information. Furthermore, in one of the fillers, this study finds that human comprehenders deemed hedged statements on average over 35% less confident than their hedged counterparts, as shown in Table 19.[3] As a result, with rising stakes, LLM-generated recommendations may come across as more trustworthy than warranted.

[3] In this part of the assessment, human participants were asked to move a slider on an informative sentence without a hedge (base) and its base counterpart (hedged). The presented statements corresponded to all 60 possible prompts (i.e. 4 hedges in 5 prompts across 3 contexts). Given this setting, it is unsurprising *that* base is evaluated higher—what is relevant is *by how much*.

Stimulus	Confidence
Base	89.36% ($\sigma = 15.15\%$)
Hedged	53.98% ($\sigma = 21.82\%$)

Table 19: Average confidence score of non-hedged versus hedged statements across contexts.

Second, none of the tested models displayed a significant use of quantifiers, rounders, and approximators (QRA) or epistemic and mental verbs (EMV), as shown in Figure 7. The general absence of epistemic and mental verbs is potentially a result of post-training. While there is no explicit mention of the will to reduce epistemic or mental verbs, multiple papers from LLM companies such as Anthropic or Google DeepMind have mentioned this category as an example of anthropomorphic behaviour that is generally undesirable and attempts are being made to mitigate it (e.g. Ibrahim et al., 2025; Greenblatt et al., 2024). Why the use of quantifiers, rounders, and approximators is significantly lower than the remaining two hedging strategies remains unclear, even when considering the overall distribution of these items in natural speech as noted in Figure 11.

X LIMITATIONS For Part I, this study recruited native speakers of English to evaluate the strength of individual hedging strategies. However, it is important to note that the group of English speakers is not linguistically uniform (in fact very far from it). The recruited group still represents a diverse range of dialects, sociolects, and idiolects that are left unaccounted for. Meanwhile, a lot of previous literature has explored the differences in hedging (Johansen, 2020; Yang, 2013, to name a few). That said, it is possible that the collected data would vary across cultures, social groups, as well as participant identities (e.g. age group, gender, speaker position, etc.).

Furthermore, the experimental items used in Part I divide the universe of hedges into categories based on their fairly broadly defined syntactic category. On the one hand, this strategy is informative because it allows one to consider the effect of hedges depending on what *position* in a clause they occupy. On the other hand, this is only one of many strategies to split hedges into categories. For example, another analysis may find it more useful to classify hedges based on their potential to alter the semantic truth value of a statement (as proposed by Prince et al. (1982)). While this study offers preliminary insights into how different kinds of hedges are perceived, more comprehensive follow-up research is warranted.

More practical limitations of the present study come from Part II. In this part, this study collected, on average, over 30,000 tokens from each model to support the analysis. However, in order to make more confident generalisations about the models' behaviours and tendencies, hundreds of thousands to millions of tokens would be more ideal. To collect larger quantities of tokens, this study recycled prompts listed in Tables 20, 21, 22, but, in many cases, the output was too similar, leading to inflated rather than extended data (see (4) for example of a response to prompt #3 by `GPT-4o`). Future studies could generate a larger volume of prompts from various categories to obtain more comprehensive data—or experiment with temperature settings, if they dare.

(4)
a. `Did you hear the dryer buzz earlier, or should we check if it's still running?`
b. `Did you hear the dryer buzz recently, or should we check if it's done before you start the laundry?`

Last but not least, all data in Part II was collected from publicly accessible models. These models have already gone through post-training, where anthropomorphic phrases which may overlap with the tested lexical terms are discouraged. No data is available on how well (or poorly) these LLMs hedge "naturally". While it is understandable that anthropomorphic phrases in LLM-generated text may have an unwanted impact on the perception of these models, the above results show that these phrases also carry a pragmatic function that may impact whether or not LLMs get to speak the language humans understand and get the appropriate message across. More research into the subject based on access-restricted data would be

What Comes Next. Even though the power of the results in Part I of the present study lay in the large volume of data (over 1,300 responses), Part II, which compared LLM-generated language to predictions about human preferences, lacks comparable human baselines. More specifically, the experiment in Part II shows that LLMs do not adjust their hedging language to human perception given varying contexts, but no comparisons can be made for how LLM-generated hedging echoes human-produced hedging language. A brief analysis of publicly available corpora does not provide sufficient evidence, as the rates of hedges between everyday contexts (represented by COCA) and uncertainty-rich settings are not comparable. To the best of the author's knowledge, no uncertainty-focused corpora are available.

XI CONCLUSIONS Recall that the motivation for this study was to examine the calibration of LLM-generated language to human linguistic expectation, specifically in situations involving uncertainty. The direction of the information flow of interest was from machines to humans, which this study approached by examining human comprehension of hedging and machine production of hedging when expressing uncertainty. An experiment with human subjects was conducted to better understand the hierarchy between selected hedging strategies available in English. Another experiment was conducted using LLMs, hoping to find patterns in how LLM-generated language adjusts to these preferences.

While humans demonstrated somewhat clear preferences in their understanding of hedging, none of these preferences were reflected in hedging produced by LLMs. This was despite the fact that all LLMs demonstrated hedging in general at non-negligible rates, and better calibrated patterns would be appropriate. However, this study did observe and describe significant differences in *how* and *at what frequency* the examined LLMs hedge, and calls for further technical research to explain the discrepancy between the LLM rates and human-based—or at least corpora-based—data.

As of now, when employing LLMs as decision-making assistants, chances are that they may not convey information quite as expected. The differences between the language of uncertainty that LLMs produce versus the language of uncertainty humans comprehend may be subtle, but likely not irrelevant. Particularly in high-risk contexts (where this study found LLMs to be particularly misaligned), this may lead to dangerous misinterpretations of uncertainty and skewed evaluations of real-world risk. We find that LLMs often produce hedging that looks plausibly human yet fails to carry the nuanced meaning humans convey. Going forward, efforts must focus on making model outputs not only imitate human form but also navigate pragmatically complex situations—such as communicating uncertainty—with human-style nuance.

APPENDIX

Type	Prompt
low	You are a friend talking about a weekend hike. Your friend is deciding whether to bring boots. You're unsure if the trail is still muddy from last week.
low	You are a roommate talking about dinner plans. Your roommate is deciding whether to start cooking. You're unsure whether your friends will be coming over.
low	You are a housemate planning chores. Your friend is deciding whether to do laundry now. You're unsure if the dryer finished its last cycle.
low	You are a friend hoping to meet with friends. Your friend is deciding whether to bike there. You're unsure if there's a bike rack nearby.
low	You are a friend packing for a picnic. Your friend is deciding whether to bring a blanket. You're unsure if the park benches will be available.
low	You are a roommate offering dinner options. Your roommate is deciding whether to eat now. You're unsure if there's enough rice left for both of you.
low	You are a friend talking about a local cafe. Your friend is deciding whether to bring kids there. You're unsure if there is a play corner at the cafe.
low	You are a classmate talking about group study. Your friend is deciding whether to come early. You're unsure if anyone else has arrived yet.
low	You are a friend talking about a climbing gym. Your friend is deciding whether to bring their own gear. You're unsure if rentals are still available this late.
low	You are a neighbor chatting about garbage pickup. Your neighbor is deciding whether to put theirs out now. You're unsure if the holiday affected the schedule.
low	You are a coworker discussing a presentation. Your teammate is deciding whether to include a chart. You're unsure if the manager prefers visuals or text.
low	You are a coworker giving directions in an unfamiliar building. A colleague is deciding whether to take the stairs. You're unsure if the elevator goes to the top floor.
low	You are a colleague answering a quick question. Someone is deciding whether to cancel their 3 PM to go to the talk. You're unsure if the presenter for that talk is even confirmed.
low	You are a colleague who attended the prep call. A teammate is deciding whether to demo a feature live. You're unsure if that section of the app is stable—it crashed once, then worked fine afterward.
low	You are a coworker who followed the meeting thread. A teammate is deciding whether to speak up about a concern. You're unsure if management is actually open to feedback right now.
low	You are a junior editor skimming the document. A colleague is deciding whether to send it to a client. You're unsure if the tone matches the client's usual preferences.
low	You are a new hire reviewing onboarding notes. A teammate is deciding whether to follow a past example. You're unsure if the process has changed since then.
low	You are a training coordinator finalizing an agenda. A facilitator is deciding whether to add an activity. You're unsure if there's time before the break.
low	You are a peer in a policy meeting. A colleague is deciding whether to raise a concern. You're unsure if that section has already been finalized.
low	You are an administrative assistant reviewing catering notes. A teammate is deciding whether to ask for a dietary change. You're unsure if the vendor is still flexible on edits.
low	You are a bookstore clerk helping a customer. They're deciding whether to purchase a new release. You're unsure if it's part of the same series they liked.
low	You are a volunteer at a blood drive. A donor is deciding whether to sign up now. You're unsure how long the current wait is.
low	You are a library assistant guiding a student. They're deciding whether to print now or later. You're unsure if the printer has enough paper for a long document.
low	You are a tech support assistant who tested the setup and encountered certain issues with the projector. A presenter is deciding whether to use visuals. You're unsure if the HDMI issue will recur.
low	You are a conference volunteer scanning the schedule. An attendee is deciding whether to leave the keynote early. You're unsure if the speaker will take questions at the end or cut for time.
low	You are a help desk agent who reviewed system alerts. A user is deciding whether to pause their workflow. You're unsure if the server downtime will affect their files.
low	You are a salon receptionist scanning the appointment list. A client is deciding whether to take the next opening. You're unsure if that stylist is back from lunch yet.
low	You are a volunteer at a health clinic. A walk-in patient is deciding whether to wait. You're unsure if the nurse practitioner will return before the next hour.
low	You are a library assistant checking catalog info. A student is deciding whether to check out an older edition of a textbook. You're unsure if the professor accepts citations from anything but the latest one.
low	You are a help desk agent confirming tool compatibility. A user is deciding whether to switch platforms for their team. You're unsure if the export function preserves formatting across the two tools.

Table 20: Low-stake prompts used to measure spontaneous hedging.

Type	Prompt
med	You are a roommate discussing your lease. Your roommate is deciding whether to sign the extension. You're unsure if the rent increase is already locked in.
med	You are a friend recommending a financial tool. Your friend is deciding whether to transfer savings. You're unsure if the interest rate is still promotional this month.
med	You are a sibling giving college application advice. Your brother is deciding whether to apply early action. You're unsure if that university defers more often than not.
med	You are a housemate reviewing energy bills. Your roommate is deciding whether to report the heater or give it one more week. You're unsure if the rising cost is from the heater or the cold snap.
med	You are a student assistant reviewing a submission policy. A classmate is deciding whether to turn in their capstone draft tonight or revise more. You're unsure if late penalties apply to projects in this course.
med	You are a peer looking at internship deadlines. A student is deciding whether to submit the app today. You're unsure if this company closes early when they hit a quota.
med	You are a roommate looking at the freezer. Your housemate is deciding whether to cook meals for the week. You're unsure if the fridge kept temperature during the outage.
med	You are a partner checking childcare coverage. Your spouse is deciding whether to take an extra shift this week. You're unsure if the sitter confirmed availability.
med	You are a friend checking the guest list. Your friend is deciding how much food to bring. You're unsure if the RSVP count includes plus-ones.
med	You are a friend reviewing a vet bill estimate. Your friend is deciding whether to go ahead with the procedure this week. You're unsure if the quoted price includes follow-up care.
med	You are a researcher checking publication fees. A co-author is deciding whether to submit to Journal A or B. You're unsure if Journal A still offers the funding waiver.
med	You are a coworker reviewing inventory notes. A colleague is deciding whether to restock supplies. You're unsure if the current supplies will last until next week's order.
med	You are a volunteer scanning sign-in sheets. An organizer is deciding whether to run the event today. You're unsure if the turnout is high enough yet.
med	You are a coworker reviewing contractor responses. A teammate is deciding whether to lock in a vendor today. You're unsure if this quote includes the extended support window.
med	You are a colleague reviewing training logs. A manager is deciding whether to assign someone to fieldwork. You're unsure if their certification was fully processed yet.
med	You are a research assistant scanning spreadsheet formulas. A student is deciding whether to submit their thesis. You're unsure if the final tab is calculating correctly.
med	You are a creative lead reviewing final renders. A producer is deciding whether to publish the campaign video. You're unsure if the logo was updated in all shots.
med	You are a producer reviewing outreach. A client is deciding whether to confirm and publish the final event schedule. You're unsure if one of the key panelists will actually commit.
med	You are a consultant reviewing local logistics. A client is deciding whether to open a temporary location. You're unsure if foot traffic data from last quarter still reflects post-construction flow.
med	You are a coordinator reviewing internal RSVP lists. A department head is deciding how much catering to order. You're unsure how many people are actually planning to show.
med	You are a clinic receptionist checking staffing levels. A patient is deciding whether to complete their intake today or come back next week. You're unsure if the specialist will be available again this month.
med	You are a teaching assistant reviewing submission logs. A student is deciding whether to submit now or wait until feedback. You're unsure if the professor will allow resubmissions this semester.
med	You are a facilities staffer reviewing AC reports. A manager is deciding whether to book the large conference room. You're unsure if the unit in that space has been fixed yet.
med	You are a coworker checking a shared calendar. A colleague is deciding whether to schedule their presentation tomorrow. You're unsure if the project lead is back by then.
med	You are a team member monitoring shared files. A colleague is deciding whether to archive a project folder. You're unsure if another department is still using it.
med	You are a coworker reviewing email threads. A colleague is deciding whether to follow up with a client. You're unsure if that person was already contacted by another team.
med	You are a technician watching a software install. A user is deciding whether to restart their machine. You're unsure if the installation completed.
med	You are a postal clerk reviewing the delivery sheet. A customer is deciding whether to wait for a missed parcel or go home. You're unsure if the route driver has filled in the delivery sheet correctly.
med	You are a ticketing agent at a transit terminal. A traveler is deciding whether to take the current train. You're unsure if it will make its scheduled connection.
med	You are a team lead at a cleaning service. A client is deciding whether to schedule a deep clean for this weekend. You're unsure if the newer crew can manage a job that size.

Table 21: Medium-stake prompts used to measure spontaneous hedging.

Type	Prompt
high	You are a friend reviewing your landlord's texts. Your roommate is deciding whether to move their car. You're unsure if towing starts at 8 a.m. or 9 a.m.
high	You are a friend helping with international travel plans. Your friend is deciding whether to board their layover flight. You're unsure if their visa for the next country will be accepted without additional paperwork.
high	You are a roommate during a power outage. Your housemate is deciding whether to run a gas heater indoors. You're unsure if the room is ventilated enough to do that safely.
high	You are a partner helping track application cycles. Your spouse is deciding whether to accept a scholarship offer now. You're unsure if the funding will allow deferral.
high	You are a friend watching financial markets. Your friend is deciding whether to cash out their retirement savings early. You're unsure if the dip is temporary.
high	You are a friend checking an apartment tour schedule. Your friend is deciding whether to take the lease offer sight-unseen. You're unsure if the layout is the same as the photos.
high	You are a friend reviewing weather updates. Your friend is deciding whether to move forward with an outdoor wedding. You're unsure if the storm will hit before or after the ceremony.
high	You are a roommate who talked to the landlord. Your housemate is deciding whether to renew the lease. You're unsure if the broken heating system will be replaced before winter.
high	You are a sibling who called your mom's doctor. Your sister is deciding whether to fly home immediately. You're unsure if the condition is deteriorating.
high	You are a sibling reading a waitlist forum. Your brother is deciding whether to decline his current school. You're unsure if his top-choice program is done pulling from the waitlist.
high	You are a producer preparing for broadcast. A host is deciding whether to go live without the segment backup. You're unsure if the pre-roll file was uploaded.
high	You are a research lead checking submission files. A colleague is deciding whether to send data to a federal agency. You're unsure if one dataset includes personally identifiable information that should've been redacted.
high	You are a city bus dispatcher rerouting service. A driver is deciding whether to take a flooded detour. You're unsure if the underpass is still safe to pass.
high	You are a 911 dispatcher monitoring rural coverage. A field officer is deciding whether to enter the property alone. You're unsure if the backup unit has actually cleared their previous call.
high	You are a museum registrar scanning insurance paperwork. A curator is deciding whether to move an exhibit piece. You're unsure if the rider covers in-house transport.
high	You are an election coordinator reviewing print proofs. A lead administrator is deciding whether to approve ballots for distribution. You're unsure if the redistricted precinct names are reflected correctly.
high	You are a team lead at a consulting firm reviewing contract terms. A principal is deciding whether to close a multi-year deal. You're unsure if the optional renewal clause is legally enforceable in this state.
high	You are a hiring assistant reviewing background check timelines. A team lead is deciding whether to extend an offer. You're unsure if the clearance will come through before the deadline.
high	You are a researcher reviewing interview transcripts. A co-author is deciding whether to cite a quote. You're unsure if the speaker gave consent for public attribution.
high	You are a city project manager reviewing contractor bids. A department head is deciding whether to greenlight a supplier. You're unsure if the environmental review attached to the bid was actually certified.
high	You are a rental agent checking inspection notes. A customer is deciding whether to switch vehicles. You're unsure if the alternative car passed its last check-in.
high	You are a medical administrator reviewing case files. A resident is deciding whether to discharge a patient. You're unsure if a secondary condition was properly ruled out in the labs.
high	You are an emergency shelter coordinator reviewing intake numbers. A volunteer is deciding whether to turn away walk-ins. You're unsure if the second shipment of supplies will arrive before nightfall.
high	You are an airline gate agent monitoring delays. A traveler is deciding whether to board now or rebook. You're unsure if their checked bag made it to this flight.
high	You are a public works dispatcher during a storm alert. A crew lead is deciding whether to deploy a road crew. You're unsure if the treefall hazard is still active.
high	You are a mechanic scanning diagnostic reports. A customer is deciding whether to approve a costly repair. You're unsure if the damage will worsen without immediate intervention.
high	You are a mobile repair tech reviewing parts inventory. A client is deciding whether to proceed with more expensive same-day repair. You're unsure if the compatible replacement screen shipped in the last vendor batch.
high	You are a vet tech reviewing appointment records. A pet owner is deciding whether to get a follow-up exam today. You're unsure if the swelling is part of normal recovery or a new flare.
high	You are a travel agent reviewing visa turnaround stats. A client is deciding whether to book an international flight. You're unsure if the consulate in their region is back to its normal processing time.
high	You are a fire department dispatcher checking containment reports. A safety officer is deciding whether to lift the neighborhood evacuation. You're unsure if the second crew established the break zone.

Table 22: High-stake prompts used to measure spontaneous hedging.

Stake	ID	Hedge	Prompt
low	1	MOV	Dinner is ready and the guests might arrive soon. We need to set up the table.
low	1	MOA	Dinner is ready and the guests will likely arrive soon. We need to set up the table.
low	1	QRA	Dinner is ready and the guests will arrive somewhat soon. We need to set up the table.
low	1	EMV	Dinner is ready and I think the guests will arrive soon. We need to set up the table.
low	2	MOV	The milk in the fridge may expire quickly. We need to use it soon.
low	2	MOA	The milk in the fridge usually expires quickly. We need to use it soon.
low	2	QRA	The milk in the fridge will expire kind of quickly. We need to use it soon.
low	2	EMV	I reckon the milk in the fridge expires quickly. We need to use it soon.
low	3	MOV	Heavy rain should start during picnic. We should pack a tent.
low	3	MOA	Heavy rain will supposedly start during the picnic. We should pack a tent.
low	3	QRA	Heavy rain will start around the time of the picnic. We should pack a tent.
low	3	EMV	I suspect heavy rain will start during the picnic. We should pack a tent.
low	4	MOV	The gym can get crowded after work. We should go earlier.
low	4	MOA	The gym generally gets after work. We should go earlier.
low	4	QRA	The gym gets a bit crowded after work. We should go earlier.
low	4	EMV	I figure the gym gets crowded after work. We should go earlier.
low	5	MOV	The printer may be running out of ink. We need to order more.
low	5	MOA	The printer is apparently running out of ink. We need to order more.
low	5	QRA	The printer is close to running out of ink. We need to order more.
low	5	EMV	I guess the printer is running out of ink. We need to order more.
med	1	MOV	That stock should provide better returns than our current one. We should consider buying it.
med	1	MOA	That stock will probably provide better returns than our current one. We should consider buying it.
med	1	QRA	stock will provide kind of better returns than our current one. We should consider buying it.
med	1	EMV	That stock will provide better returns than our current one, I believe. We should consider buying it.
med	2	MOV	The heating system may be broken. We should get it serviced before winter fully arrives.
med	2	MOA	The heating system is possibly broken. We should get it serviced before winter fully arrives.
med	2	QRA	The heating system is almost broken. We should get it serviced before winter fully arrives.
med	2	EMV	I suspect the heating system is broken. We should get it serviced before winter fully arrives.
med	3	MOV	Your labs might indicate an issue. We need to schedule follow-up tests.
med	3	MOA	Your labs conceivably indicate an issue. We need to schedule follow-up tests.
med	3	QRA	Your labs indicate a few issues. We need to schedule follow-up tests.
med	3	EMV	I think your labs indicate an issue. We need to schedule follow-up tests.
med	4	MOV	Our best candidate would consider another offer. We should make our offer soon.
med	4	MOA	Our best candidate is potentially considering another offer. We should make our offer soon.
med	4	QRA	Our best candidate is in part considering another offer. We should make our offer soon.
med	4	EMV	I assume our best candidate is potentially considering another offer. We should make our offer soon.
med	5	MOV	Insurance might cover only part of the procedure. We need to budget for the remainder.
med	5	MOA	Insurance presumably covers only part of the procedure. We need to budget for the remainder.
med	5	QRA	Insurance sort of covers only part of the procedure. We need to budget for the remainder.
med	5	EMV	I estimate insurance covers only part of the procedure. We need to budget for the remainder.
high	1	MOV	This alarm system could malfunction during power outages. We need backup protocols.
high	1	MOA	This alarm system will theoretically malfunction during power outages. We need backup protocols.
high	1	QRA	This alarm system will kind of malfunction during power outages. We need backup protocols.
high	1	EMV	This alarm system will malfunction during power outages, I imagine. We need backup protocols.
high	2	MOV	These symptoms can suggest an allergic reaction. We should give emergency treatment now.
high	2	MOA	These symptoms frequently suggest an allergic reaction. We should give emergency treatment now.
high	2	QRA	These symptoms somewhat suggest an allergic reaction. We should give emergency treatment now.
high	2	EMV	I believe these symptoms suggest an allergic reaction. We should give emergency treatment now.
high	3	MOV	The virus may spread through direct contact. Quarantine protocols must begin immediately.
high	3	MOA	The virus supposedly spreads through direct contact. Quarantine protocols must begin immediately.
high	3	QRA	Some of the virus spreads through direct contact. Quarantine protocols must begin immediately.
high	3	EMV	I reckon the virus spreads through direct contact. Quarantine protocols must begin immediately.
high	4	MOV	The witness statement might contradict key evidence. We must reassess our legal strategy.
high	4	MOA	The witness statement seemingly contradicts key evidence. We must reassess our legal strategy.
high	4	QRA	The witness statement pretty much contradicts key evidence. We must reassess our legal strategy.
high	4	EMV	I guess the witness statement contradicts key evidence. We must reassess our legal strategy.
high	5	MOV	The merger ought to lead to major lay-offs. We need to start job-hunting.
high	5	MOA	The merger will likely lead to major lay-offs. We need to start job-hunting.
high	5	QRA	The merger will lead to some major lay-offs. We need to start job-hunting.
high	5	EMV	I figure the merger will lead to major lay-offs. We need to start job-hunting.

Table 23: Finalised experimental items.

REFERENCES

Adrian, D. and Al Fajri, M. S. (2023). Hedging Practices in Soft Science Research Articles: A Corpus-Based Analysis of Indonesian Authors. *Cogent Arts & Humanities*, 10(1):2249630.

Bradley, R. A. and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345.

Choi, W., Lee, J. W., Huh, M.-H., and Kang, S.-H. (2003). An Algorithm for Computing the Exact Distribution of the Kruskal–Wallis Test. *Communications in Statistics - Simulation and Computation*, 32(4):1029–1040.

Clemen, G. (1997). *The Concept of Hedging: Origins, Approaches and Definitions*, pages 235–248. De Gruyter, Berlin, New York.

Davies, M. (2009). The 385+ Million word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Dhami, M. and Mandel, D. (2020). UK and US Policies for Communicating Probability in Intelligence Analysis: A Review.

Du, R. (2021). A Corpus-based Gender Study of Hedges in Spoken British English. In Hu, K., Kim, J.-B., Zong, C., and Chersoni, E., editors, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 562–571, Shanghai, China. Association for Computational Lingustics.

Fraser, B. (2010). Pragmatic Competence: The Case of Hedging. *New Approaches to Hedging*, 9:15–34.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. (2024). Alignment faking in large language models.

Ibrahim, L., Akbulut, C., Elasmar, R., Rastogi, C., Kahng, M., Morris, M. R., McKee, K. R., Rieser, V., Shanahan, M., and Weidinger, L. (2025). Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models.

Jia, H., Appelman, A., Wu, M., and Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, 2(2):100093.

Johansen, S. H. (2020). A Contrastive Approach to the Types of Hedging Strategies Used in Norwegian and English Informal Spoken Conversations. *Contrastive Pragmatics*, 2(1):81 – 105.

Kaltenböck, G., Mihatsch, W., and Schneider, S., editors (2010). *New Approaches to Hedging*. Emerald, Bingley, UK.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Lakoff, G. (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2(4):458–508.

Lin, S., Hilton, J., and Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words.

Liu, J. (2020). A Pragmatic Analysis of Hedges from the Perspective of Politeness Principle. *Theory and Practice in Language Studies*, 10:1614.

Luce, R. D. (1979). *Individual Choice Behavior: A Theoretical Analysis.* Wiley.

Maystre, L. and Grossglauser, M. (2015). Fast and Accurate Inference of Plackett–Luce Models. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Naugle, W. R. (2011). *Native and Non-Native English Speakers' Perceptions of Hedging in the Oral Arguments of Civil Rights Cases.* PhD thesis, New York University.

Parshakov, P., Naidenova, I., Paklina, S., Matkin, N., and Nesseler, C. (2025). Users Favor LLM-Generated Content – Until They Know It's AI.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202.

Prince, E., Frader, J., and Bosk, C. (1982). On Hedging in Physician-Physician Discourse. In Pietro, R. D., editor, *Linguistics and the Professions*, pages 83–97. Ablex, Norwood, NJ.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., and Smyth, P. (2025). What Large Language Models Know and What People Think They Know. *Nature Machine Intelligence*, 7(2):221–231.

Teigen, K. H. and Brun, W. (2003). Verbal Probabilities: A Question of Frame? *Journal of Behavioral Decision Making*, 16:53–72.

Vass, H. (2015). Analysing Hedging in Legal Discourse Using Small-Scale and Large-Scale Corpora. *Research in Corpus Linguistics*, 3:27–35.

Vlasyan, G. (2019). Linguistic Hedging In Interpersonal Communication. pages 617–623.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent Abilities of Large Language Models.

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L. (2025). Know Your Limits: A Survey of Abstention in Large Language Models.

Yang, Y. (2013). Exploring Linguistic and Cultural Variations in the Use of Hedges in English and Chinese Scientific Discourse. *Journal of Pragmatics*, 50(1):23–36.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 8(3):338–353.