Information Design with Large Language Models[*]

Paul Dütting, Google Research

Safwan Hossain[†], Harvard University

Tao Lin, Harvard University

Renato Paes Leme, Google Research

Sai Srivatsa Ravindranath, Harvard University

Haifeng Xu, University of Chicago

Song Zuo, Google Research

Abstract

Information design is typically studied through the lens of Bayesian signaling, where signals shape beliefs based on their correlation with the true state of the world. However, Behavioral Economics and Psychology emphasize that human decision-making is more complex and can depend on how information is framed. This paper formalizes a language-based notion of framing and bridges this to the popular Bayesian-persuasion model. We model framing as a possibly non-Bayesian, linguistic way to influence a receiver's belief, while a signaling/recommendation scheme can further refine this belief in the classic Bayesian way. A key challenge in systematically optimizing in this framework is the vast space of possible framings and the difficulty of predicting their effects on receivers. Based on growing evidence that Large Language Models (LLMs) can effectively serve as proxies for human behavior, we formulate a theoretical model based on access to a framing-to-belief oracle. This model then enables us to precisely characterize when solely optimizing framing or jointly optimizing framing and signaling is tractable. We substantiate our theoretical analysis with an empirical algorithm that leverages LLMs to (1) approximate the framing-to-belief oracle, and (2) optimize over language space using a hill-climbing method. We apply this to two marketing-inspired case studies and validate the effectiveness through analytical and human evaluation.

1 Introduction

Information design is a core concept in microeconomics and decision theory that considers the strategic communication of information from one party (i.e., sender) to shape the decisions of another (i.e., receiver) [1]. While information design has been extensively studied through different lenses, the *Bayesian persuasion* model, pioneered by Kamenica and Gentzkow [19], considers the decision-maker's belief influenced in a Bayesian manner and involves carefully engineering a recommendation (or more generally a signal) that is correlated with the world state. These quantitative correlations are all that matter in such theoretical models.

This perspective stands in strong contrast to established findings in Behavioral Economics and Psychology, which emphasize that human decision-making is more complex and can depend, for instance, on how the information is framed. In the landmark work of Tversky and Kahneman [31], they observe that humans, when asked to choose between two options to combat the outbreak of a hypothetical disease, decide differently depending on how the downstream effects of the two options

[*] This work features both a theoretical framework, and empirical results. For the experiments, we use GPT-4o-mini [28] for the LLM portions of the framework, and SciPy [32] for the analytical parts. All experiments were conducted by the Harvard University group.

[†] Correspondence to shossain@g.harvard.edu

are phrased. To explain this seemingly irrational behavior, they argue that the raw information combines with societal norms to affect the perception of “acts, outcomes, and contingencies” in the decision-maker.

More generally, there is ample evidence that “persuasion in the field” is richer than the existing theoretical models (see, e.g., the influential survey of DellaVigna and Gentzkow [8]), and that (i) tangential information given before a recommendation; (ii) text or phrasing used to support a recommendation; (iii) and even non-textual cues such as color, background images, or fonts all impact the effectiveness of persuasive communication. Dealing with these aspects of persuasion—which we refer to as *framing*—is extremely challenging due to (i) the vast, primarily linguistic, space of relevant framings and (ii) the difficulty in systematically understanding how framing influences a decision-maker. However, the advent of Large Language Models (LLMs) offers a new approach to addressing these twin challenges. These systems are adept at understanding the structure of natural language. Further, a growing line of work in economics (Horton [18], Brand et al. [2], Leng [20], Dillion et al. [9] and Manning et al. [24]) argues that LLMs can be used to simulate human behavior and perspectives on a host of problems, or even be tuned as agents that can be delegated to make decisions for humans [15]. It is thus timely to revisit this challenge.

1.1 Our Contribution

In this work, we build the theoretical foundations of information design with LLMs, and showcase how state-of-the-art LLMs can be used to optimize persuasive communication. To this end, we propose a formal model that combines the classic Bayesian persuasion model with the aforementioned framing perspective. This model enables a systematic investigation of which variants of the problem are intractable even under optimistic LLM-inspired assumptions, and which variants are more benign and therefore more promising. Guided by these insights, we build on publicly available LLMs to implement a practical hill-climbing approach and demonstrate its efficacy through experiments and human evaluations.

Our Model. Our basic setup parallels that of Bayesian Persuasion, with a sender (she) and a receiver (he). A state of the world $\omega \in \Omega$ is drawn from a distribution $\mu_0 \in \Delta(\Omega)$. The sender observes ω, while the receiver does not. The receiver has a set of actions $\mathcal{A}$ that affect the possibly misaligned utilities of the two players. The receiver chooses an action $a^\star \in \mathcal{A}$ that maximizes his expected utility given his belief. The sender seeks to influence the receiver's decision by inducing a different posterior belief.

For a given problem instance, we consider a possible framing space C, where each framing $c \in C$ leads to a receiver belief $\mu_c \in \Delta(\Omega)$. By choosing a framing $c \in C$, the sender induces the receiver's prior belief to be μ_c, a process that may be Bayesian or non-Bayesian, and may be determined by societal norms. In short, we model framing as affecting belief formation [12, 10], rather than changing the perception about the payoffs, but are agnostic to the exact mechanism of this process. In addition, as in classic Bayesian persuasion, the sender may commit to a recommendation (or signaling) scheme, which constitutes a possibly randomized mapping from world states to recommendations (signals). We assume that the belief update induced by a recommendation scheme is Bayesian, mirroring the Bayesian persuasion literature [19, 11].

Problem Variants. Our problem formulation, which combines classic Bayesian persuasion with framing, gives rise to three possible variants of optimization problems for the sender:

(a) *Recommendation-only:* The framing (and thus the receiver's initial belief) is fixed; the sender may only optimize over the recommendation scheme.

(b) *Framing-only:* The sender's recommendation scheme is fixed; they may only optimize over the framing.

(c) *Joint Framing and Recommendation:* The sender may jointly optimize over both the framing and the recommendation scheme.

The first setting corresponds to facing a receiver with a fixed but possibly distinct (from the sender) prior belief. This direction is captured by the large literature on Bayesian persuasion. By contrast, the key thrust of this work is to study settings (b) and (c), both theoretically and empirically. Setting (b) is relevant where the sender is bound to an information revelation scheme they have committed to in the past (e.g., a multi-year advertising strategy or regulation restrictions), but can change the framing (e.g., endorsement, wording, etc.); setting (c) represents a sender with full freedom to choose both.

Theoretical Results. Inspired by influential recent work in economics by Horton [18], Brand et al. [2], Dillion et al. [9], we abstract the ability of LLMs to proxy for human behavior by assuming access to a noisy *framing-to-belief oracle* (or belief oracle for brevity) that maps a framing $c \in C$ to a belief $\mu_c \in \Delta(\Omega)$. The primary goal of our theoretical investigation is to identify variants of the problem that are amenable to optimization when LLMs can play this crucial role. Section 3 investigates approach (b), optimizing the framing for a fixed recommendation scheme, while approach (c), jointly optimizing the framing and recommendation scheme, is studied in Section 4.

Our results reveal that approach (b) poses severe challenges, even under generous assumptions. Specifically, we show that the sender's utility will generally be a discontinuous function of the belief μ_c induced by framing: slight errors in the LLM-inspired belief oracle can thus have significant implications. We also show that the problem of finding the optimal framing is computationally hard (NP-hard), even to approximate. In contrast, we show that jointly optimizing the framing and the recommendation strategy is a more favorable problem from an optimization perspective. Particularly, we show that the sender's utility as a function of the framing-induced belief μ_c and the corresponding optimal recommendation strategy is a continuous function, making the setting graceful to errors. We also give a quasi-polynomial-time approximation scheme (QPTAS) for the joint optimization problem. These findings support the conclusion that the joint optimization problem tends to be the more tractable problem for systematic optimization, inspiring our experiments. Overall, our work characterizes the unique structural and optimization landscape of this generalized model of information design.

Empirical Results. In Section 5, we propose a practical hill-climbing inspired method for the joint optimization problem. LLMs, used to both optimize the framing and serve as the oracle mapping framing to belief, are combined with analytical methods that optimize the recommendation strategy for a given belief. To demonstrate the efficacy of this approach, we present two case studies: a realtor pitching himself (framing) to prospective buyers and deciding on the property recommendation strategy, and a clothing brand developing an advertising campaign that involves a slogan and description of the product line (framing) and the discount/sale strategy.

In both case studies, we observe that the proposed approach is highly effective when measuring the effectiveness by the LLM's reaction to the framing. To further validate our findings we also present a study with human subjects, and find that while the reaction of the human subjects is less pronounced and shows more variance, our predictions are directionally correct and also well-approximate the magnitude of the effect.

Broader Implications. This work formally integrates behavioral framing effects into information design. In the pre-LLM era, designing persuasive communication was necessarily human-driven and relied on heuristics and costly focus groups since automating the generation of framing narratives posed significant challenges. We take the first steps toward exploring a systematic, algorithmically driven alternative that leverages LLMs for this setting. We characterize exactly when they hold major promise – joint optimization of framing and recommendation – and when they still face computational challenges – framing-only optimization for fixed recommendation. For the former, our empirical studies yield a simple, scalable procedure that enables organizations to systematically select textual framings alongside quantitative recommendations. Looking ahead, as generative AI technology progresses, we see ample room to both improve these implementations and extend them to non-textual communication mediums like images or video.

1.2 Additional Related Work

Algorithmic information design. The algorithmic study of information design has attracted significant recent interest. This literature starts from the complexity-theoretic study initiated by Dughmi and Xu [11], and lately has integrated many aspects of machine learning to address unknowns in the setting [4, 14, 22]. Haghtalab et al. [16] study a communication game where the sender persuades a receiver by providing "anecdotes", instead of committing to recommendation strategies, assuming a known specific model of how anecdotes influence the receiver's behavior. While their anecdote effect shares a similar spirit to our framing effect, we do not assume any specific framing effect. The use of LLM allows us to model real-world framing effects that cannot be easily characterized theoretically.

Information design with LLMs. Among recent works on the interface of information design and LLMs, Harris et al. [17] study a learning-theoretic question about learning an optimal recommendation strategy by querying an LLM that simulates a receiver with unknown prior belief. This differs from our aim of introducing a new dimension, i.e., framing, to information design. Li et al. [21] explore how Bayesian persuasion mechanisms can be implemented in natural language, while Wu et al. [33] develop methods for generating persuasive marketing content that is both linguistically coherent and grounded in product attributes. Both works share our emphasis on bridging formal persuasion models with natural language generation, but differ in the focus: while they design practical systems for persuasive text generation, our work provides a theoretical framework that explicitly integrates framing into information design and studies its tractability when combined with signaling.

Additional evidence for framing effects. In addition to the landmark study of Tversky and Kahneman [31], there are numerous studies in Behavioral Economics and Psychology that provide evidence for framing effects. For instance, Ellingsen et al. [12] show that naming the standard prisoner's dilemma game differently (e.g., as the "Community Game" or "Stock Market Game") will lead to different players behaviors despite playing the same underlying game. These behavioral studies are consistent with "the hypothesis that social frames are coordination devices... enter people's beliefs rather than their preferences"[12]. Nelson and Oxley [27] observe similar belief influence by framing in persuasion problems. These behavioral studies motivate our principled study of using framing in information design.

LLMs as proxys for human behavior. Our work subscribes to the recent studies that use LLMs as proxies of human agents to understand the social norms they carry and/or the economic decisions they make [18, 2, 20, 9]. Our work fits this general theme, but is different from these works in research questions.

2 Model

Bayesian Persuasion and Signaling. A fundamental model of information design is the classic Bayesian persuasion problem between two rational players: a *sender* (persuader, with she/her pronouns) and a *receiver* (decision maker, with he/him pronouns). Both players' utilities depend on a *state of the world* $\omega \in \Omega$. The receiver must take an action a from some finite set $\mathcal{A}$, which, along with the state ω, jointly determines the utilities of both players.[1] Formally, the sender's utility function is $u : \mathcal{A} \times \Omega \to \mathbb{R}$ and the receiver's utility function is $v : \mathcal{A} \times \Omega \to \mathbb{R}$. The sender does not act, but possesses an information advantage: she is assumed to perfectly observe the realized state of the world ω, whereas the receiver only knows the prior distribution $\mu_0 \in \Delta(\Omega)$ of the state. For simplicity, we make a few mild *regularity* assumptions. First, μ_0 has full support, i.e., $\mu_0(\omega) > 0$ for every state $\omega \in \Omega$. Second, every action $a \in \mathcal{A}$ is strictly inducible: that is, for every $a \in \mathcal{A}$, there is some belief $\mu \in \Delta(\Omega)$ wherein action a is the unique best response action for the receiver, i.e., $\mathbb{E}_{\omega\sim\mu}[v(\omega, a)] > \mathbb{E}_{\omega\sim\mu}[v(\omega, a')]$ for all $a' \in \mathcal{A} \setminus \{a\}$. Third, some of our computational results are about additive approximation, which requires players' utilities to be bounded. Hence without loss of generality we assume both players' utilities u, v are normalized to be within $[0, 1]$.[2]

The standard sender-receiver interaction in Bayesian persuasion is through a sender-designed *signaling scheme* that partially reveal the realized state ω to the receiver. Formally, for a signaling space $\mathcal{S}$, the sender designs and commits to a randomized mapping $\pi : \Omega \to \Delta(\mathcal{S})$, where $\pi(s|\omega)$ specifies the probability of sending signal $s \in \mathcal{S}$ when the realized state is ω. The receiver, upon observing a signal s sampled from this signaling scheme, updates their belief over Ω and takes the expected-utility-maximizing action. As a textbook assumption in this literature, ties are assumed to be broken in favor of the sender. In the special case where the signaling space $\mathcal{S}$ equals $\mathcal{A}$, the signaling scheme $\pi : \Omega \to \Delta(\mathcal{A})$ is also called a *recommendation scheme*, where signals correspond to action recommendations. A recommendation scheme π is *obedient* if whenever π recommends an action $a \in \mathcal{A}$ to the receiver, a is indeed optimal based on the receiver's posterior belief. We will consider general signaling schemes as well as recommendation schemes in this work.

Framing. The core hypothesis of this work is that, in real-world persuasion, the receiver's belief is not only influenced by the signaling scheme as in classic Bayesian persuasion models, but also significantly shaped by how the signals are described to receivers. To capture the latter effect, we employ the seminal work of Tversky and Kahneman [31] for studying how the description of the decision-making problem affects agents' psychology of choice, a framework widely known as *framing.* Adapted to our setting, framing corresponds to natural language phrases or descriptions that accompany, describe, or convey the signal; thus, the set of possible framings, denoted by C, can encompass all possible coherent text within some linguistic and semantic constraints defined by the problem instance. To connect framing to belief, let function $\ell : C \to \Delta(\Omega)$ (which we coin a *belief oracle*) denote the mapping from a framing c to the corresponding belief μ_c it induces in the receiver. Let set $B = \{\ell(c) : c \in C\}$ denote all inducible beliefs. This mapping abstracts out the belief update procedure, which could be Bayesian or non-Bayesian. From a non-Bayesian perspective, the framing may simply form some receiver belief inherited from common sense in human languages. To exposit the Bayesian perspective, one can imagine that the receiver has some initial belief about the state

[1] We use 0 index for actions and states. That is, $\mathcal{A} = \{a_0, \ldots, a_{|\mathcal{A}|-1}\}$, and $\Omega = \{\omega_0, \ldots, \omega_{|\Omega|-1}\}$.

[2] Any utilities bounded within $[-A, B]$ can without loss of generality be normalized to $[0, 1]$.

ω, which was then Bayesian updated to μ_c based on certain "societally consensed" signaling $\sigma(c|\omega)$ after observing the framing c. In any case, how the mapping ℓ was formed is not concerned in the mathematics of our model. However, in Section 5, we empirically demonstrate that such a framing-to-belief mapping ℓ can be robustly approximated by LLMs, serving as a justification for this modeling primitive. In Sections 3 and 4, we also analyze the robustness of the optimization problem to inaccuracies of the belief oracle ℓ. For these studies, we allow the belief oracle to have small errors and let ℓ_ϵ denotes the function that satisfies $|\ell_\epsilon(c) - \ell(c)| \leq \epsilon$ for every $c \in C$, i.e., ℓ_ϵ is inaccurate up to at most ϵ for any c.

Natural-language-generated framing space C is discrete by nature, though it is enormous. In this case, the corresponding inducible belief set B is also discrete. We will also consider the relaxation to continuous framing space; in this case, we assume that the inducible beliefs B form a *convex* subset of the simplex $\Delta(\Omega)$. That is, if two beliefs μ_1, μ_2 can be induced by some framing, there are ways of framing to induce any belief in between. It is valuable to consider the continuous relaxation for two reasons. First, the geometry of inducible beliefs enables more structural characterization of the optimal design. Second, while this may not precisely map to the ultimately discrete linguistic framing space in practice, the richness of human language renders this a reasonable approximation. We formalize these notions below.

Definition 1 (Framing Space)**.** *For a framing space C and a mapping from framings to induced beliefs $\ell : C \to \Delta(\Omega)$, let $B = \{\ell(c) : c \in C\}$ denote the set of inducible beliefs. In a* discrete *framing space, both C and B are discrete. Correspondingly, in a* continuous *framing space, C is continuous and B is assumed to be a convex subset of the belief simplex $\Delta(\Omega)$.*

Sender-Receiver Interactions and the Equilibrium. The interaction between sender and receiver is consistent with prior literature on persuasion, except that the sender's policy now additionally has a framing strategy. The model is formally described below.

Definition 2 (Information Design with Framing)**.** *The tuple $\big(c \in C,\ \pi : \Omega \to \Delta(\mathcal{S})\big)$ is denoted as the information design policy with framing. The sender first commits to a tuple $\big(c,\ \pi\big)$, and upon state realization ω, the receiver observes the pair $(c \in C, s \sim \pi(\cdot|\omega))$ and updates their belief from $\mu_c = \ell(c)$ to a posterior $\mu_c(\omega \mid s) = \frac{\mu_c(\omega)\pi(s \mid \omega)}{\sum_{\omega' \in \Omega} \mu_c(\omega')\pi(s \mid \omega')}$. The receiver then takes a best-response action that maximizes his expected utility under the updated belief $\mu_c(\cdot \mid s)$:*

$$a^*_{c,\pi,s} \in \operatorname*{arg\,max}_{a \in \mathcal{A}} \sum_{\omega \in \Omega} \mu_c(\omega \mid s) v(a, \omega) = \operatorname*{arg\,max}_{a \in \mathcal{A}} \sum_{\omega \in \Omega} \mu_c(\omega)\pi(s \mid \omega) v(a, \omega). \tag{1}$$

To map the above abstract model to our running example of advertising, this means the advertiser as the sender could choose a description or slogan (the *framing* strategy) that accompanies their product, based on its hidden features/states (the state) or discount/buy recommendation (the signal). In Proposition 1, we will show that randomizing over framing does not increase the sender's expected utility in all our problem variants. Hence, it is without loss of generality to use a single/fixed framing. The sender's goal is thus to choose an optimal strategy that maximizes their expected utility assuming that the receiver will best-respond to the sender's choice of information design strategy. Formally, this corresponds to a *Stackelberg equilibrium.*

We study two variants of the sender's information design problem with framing that complements prior works on Bayesian persuasion. The first is the optimization of framing c under a given/fixed signaling scheme (coined the *framing-only strategy*). Such an instance is specified by $\mathcal{I} = (\mu_0, u, v, \pi)$ where π is a given signaling scheme. The second is the joint optimization of the framing c and signaling scheme π, coined the *joint framing-signaling strategy.* Here, an instance is defined by $\mathcal{I} = (\mu_0, u, v)$. Note that the third possible variant, optimizing the signaling scheme π under a fixed framing (i.e., a fixed receiver belief) is essentially captured by the classical Bayesian Persuasion framework [19] and thus not the focus of our work. In all cases, the sender is selecting a strategy to maximize their ex-ante utility for a best-responding receiver.

Definition 3 (Equilibrium)**.** *The sender's ex-ante utility for an information design policy* (c, π) *is:*

$$\mathbb{E}_{\omega\sim\mu_0,\, s\sim\pi(\cdot|\omega)}\big[u(a^*_{c,\pi,s}, \omega)\big] \;=\; \sum_{\omega\in\Omega} \mu_0(\omega) \sum_{s\in S} \pi(s \,|\, \omega) u(a^*_{c,\pi,s}, \omega). \tag{2}$$

where $a^*_{c,\pi,s}$ *is the receiver's best-response action defined in Eq.* (1)*. The sender's (Stackelberg) equilibrium strategy is*

- *for the framing-only strategy space,* $\arg\max_{c\in C} \mathbb{E}\big[u(a^*_{c,\pi,s}, \omega)\big]$*;*
- *for the joint strategy space,* $\arg\max_{c\in C,\, \pi:\Omega\to\Delta(\mathcal{S})} \mathbb{E}\big[u(a^*_{c,\pi,s}, \omega)\big]$*.*

3 Framing-Only Strategy Space

This section analyzes the conceptually simpler situation where we optimize the framing while fixing the signaling scheme. Formally, given an information design instance $\mathcal{I} = (\mu_0, u, v, \pi)$, the sender looks to find the optimal framing c^* in the Stackelberg equilibrium (Definition 3). We first specify an important distinction. The classic literature on Bayesian persuasion typically considers obedient signaling schemes with signal space $\mathcal{S}$ equal to the action space $\mathcal{A}$, due to a revelation-principle style argument [19, 11]. But in our setting, the signaling scheme π is given exogenously and not within the designer's control, while the receiver's prior is optimized by the designer. As such, signals from π cannot be freely interpreted as direct action recommendations, so we consider a general signaling space.

We start by showing that randomizing over framings does not increase the expected utility of the sender under *any* signaling scheme π. This means that it is without loss of generality to focus on deterministic framing both here and in the forthcoming section that studies jointly optimizing π and c.

Proposition 1. *For any instance* $\mathcal{I}$ *with a given signaling scheme* π*, the optimal sender utility can always be achieved by some deterministic framing* c^**.*

As a simple corollary, for a discrete framing space, the optimal framing can be computed in a polynomial time proportional to the framing space size, by enumerating the framing $c \in C$ and evaluating its sender utility.

Corollary 1. *For a discrete framing space* C*, the optimal framing* c^* *can be computed in* $|C| \cdot \mathrm{poly}(|\Omega|, |\mathcal{A}|, |\mathcal{S}|)$ *time.*

In practice (e.g., our running example of advertising), however, framing encompasses contextually relevant natural language expressions, hence the space C is enormous. This gives rise to a structurally more interesting question: are there algorithms that scale more gracefully with respect to the size of the framing space C? To study this problem, we turn to the relaxed proxy problem that considers framing optimization in continuous space.

3.1 Optimizing Framing in a Continuous Space

We turn to the continuous framing space with a convex belief space $B = \{\ell(c) : c \in C\} \subseteq \Delta(\Omega)$. In the most optimistic case, the sender can use framing to induce *any* belief in the simplex: $B = \Delta(\Omega)$. This means that the framing space C is no longer a parameter, and the instance size is completely defined in terms of $|\Omega|$, $|\mathcal{A}|$, and $|\mathcal{S}|$.

Unfortunately, we show that even in the optimistic setting of $B = \Delta(\Omega)$, selecting the optimal framing is NP-Hard (in the parameter $|\Omega|$). Since the sender utility given any belief can be efficiently computed (as discussed above), we formally prove the hardness for algorithms that receive as input the instance parameters (u, v, μ_0, π) and must return the sender utility under an optimal induced belief μ_c^*. This can be formally stated as follows:

$$\begin{aligned} \max_{\mu_c \in B = \Delta(\Omega)} \quad U_\pi(\mu_c) &= \sum_{\omega \in \Omega} \sum_{s \in \mathcal{S}} \mu_0(\omega) \pi(s|\omega) u(a^*_{\mu_c, s}, \omega) \\ \text{s.t.} \qquad a^*_{\mu_c, s} &\in \arg\max_{a \in \mathcal{A}} \sum_{\omega' \in \Omega} \mu_c(\omega') \pi(s|\omega') v(a, \omega'), \quad \forall s \in \mathcal{S}. \end{aligned} \tag{3}$$

We now connect the problem of optimizing framing under a fixed signaling scheme to a classic problem in algorithmic game theory, *Bayesian Stackelberg game* (Definition 4). We then show in Proposition 2 that our optimal framing problem can be converted to finding the equilibrium of this game.

Definition 4 (Bayesian Stackelberg game). *A* Bayesian Stackelberg Game (BSG) *consists of a leader with action space $\mathcal{A}_\ell$ and a follower with action space $\mathcal{A}_f$ and unknown type θ drawn from a known distribution $P \in \Delta(\Theta)$. With type-dependent leader utility $u_\ell(a_\ell, a_f, \theta)$ and follower utility $u_f^\theta(a_\ell, a_f)$, the leader commits to a mixed strategy $x \in \Delta(\mathcal{A}_\ell)$, and the receiver will best respond. The leader's optimal (equilibrium) utility is given by:*

$$\begin{aligned} &\max_{x \in \Delta(\mathcal{A}_\ell)} \sum_{\theta \in \Theta} P(\theta) \sum_{a_\ell \in \mathcal{A}_\ell} x(a_\ell) u_\ell(a_\ell, a_f^*(\theta, x), \theta) \\ &\quad \text{s.t. } \; a_f^*(\theta, x) \in \arg\max_{a_f \in \mathcal{A}_f} \sum_{a_\ell \in \mathcal{A}_\ell} x(a_\ell) u_f^\theta(a_\ell, a_f), \quad \forall \theta \in \Theta. \end{aligned}$$

Proposition 2. *The continuous framing-only optimization problem* (3) *can be reduced to finding the equilibrium of a Bayesian Stackelberg game.*

Proof. Let $P(s) = \sum_{\omega \in \Omega} \mu_0(\omega) \pi(s|\omega)$. Consider the BSG where the leader's action space is the space of states Ω, and the follower has $|\mathcal{S}|$ types where the probability of type s is $P(s)$. The utility functions of the leader and the follower are

$$\tilde{u}(a, s) = \sum_{\omega \in \Omega} \frac{\mu_0(\omega) \pi(s|\omega)}{P(s)} u(a, \omega), \qquad \tilde{v}^s(a, \omega) = \pi(s|\omega) v(a, \omega)$$

respectively. Note that the leader's utility $\tilde{u}(a, s)$ depends on the follower's action a and type s, but not the leader's action ω. By definition, the leader's optimization problem in the BSG is

$$\begin{aligned} \max_{x \in \Delta(\Omega)} U_P(x) &= \sum_{s \in \mathcal{S}} P(s) \tilde{u}(a^*(s, x), s) \\ \text{s.t. } \; a^*(s, x) &\in \arg\max_{a \in \mathcal{A}} \sum_{\omega' \in \Omega} x(\omega') \tilde{v}^s(a, \omega'), \quad \forall s \in \mathcal{S}, \end{aligned} \tag{4}$$

namely,

$$\max_{x\in\Delta(\Omega)} U_P(x) = \sum_{s\in\mathcal{S}} P(s) \sum_{\omega\in\Omega} \frac{\mu_0(\omega)\pi(s|\omega)}{P(s)} u(a^*(s,x),\omega) \tag{5}$$
$$\text{s.t. } \; a^*(s,x) \in \arg\max_{a\in\mathcal{A}} \sum_{\omega'\in\Omega} x(\omega')\pi(s|\omega')v(a,\omega'), \;\; \forall s\in\mathcal{S}.$$

We note that the optimization problems (5) and (3) are equivalent, where x corresponds to μ_c and $a^*(s,x)$ corresponds to $a^*_{\mu_c,s}$. Therefore, we conclude that the framing-only optimization problem can be reduced to a Bayesian Stackelberg game. □

The reduction from framing-only optimization to Bayesian Stackelberg games suggests that one approach to finding the optimal framing is to use existing computational methods for BSG (e.g., [29]). However, Conitzer and Sandholm [5] show that a family of BSGs are computationally intractable (NP-hard). Our Theorem 1 proves that any BSG in that family can be converted to a continuous framing-only optimization problem. The proof is technical and formally given in Appendix A.2; but we sketch the high-level intuition below. Combined with Proposition 2, our analyses show that framing-only optimization is (bi-directional) equivalent to a subset of BSGs that are known to be computationally hard. Hence, the framing-only optimization problem is NP-hard and not easily solvable even with a perfect oracle that maps framing to belief.

Theorem 1 (NP-hardness)**.** *For a framing-only instance $\mathcal{I} = (\mu_0, u, v, \pi)$ with convex framing-induced belief space B, there is no additive Fully Polynomial-Time Approximation Scheme (FPTAS) for computing the optimal framing belief $\mu^*_c \in B$, unless $P = NP$. This hardness holds even when $B = \Delta(\Omega)$.*

Proof Sketch. Conitzer and Sandholm [5] show that a family of BSGs where the followers have binary actions is NP-hard. We show that any such BSG can be cast into a framing-only optimization problem $\mathcal{I}$ with $|\Omega| = |\mathcal{A}_\ell| + |\Theta| + 1$ states, $|\mathcal{A}| = |\Theta||\mathcal{A}_f| + 2$ actions, $|S| = |\Theta|$ signals, and a continuous framing space $B = \Delta(\Omega)$. Specifically, we create a state for each leader action, ω_{a_ℓ}, and each follower type ω_θ. We create receiver actions for each binary action a follower of a type θ can take – i.e. a^θ_i for all θ. When the receiver sees a signal s_θ (which is proxying type θ in BS), we want them to only consider actions a^θ_0, a^θ_1, which should directly correspond to follower θ's utility in taking action 0 or 1 in the BSG. Since receiver utility in persuasion does not explicitly depend on type θ, the states ω_θ are used to achieve this effect. The receiver is heavily penalized for taking an action $a^{\theta'}_*$ at state ω_θ. We carefully select the sender utilities and add additional dummy states and actions to ensure that (1) on the optimal μ_c, the persuasion sender places sufficient weight on states corresponding to ω_θ to ensure that the receiver takes this type-consistent action, and (2) the persuasion objective captures the type-dependent Bayesian Stackelberg objective. The inapproximability stems from a more careful analysis of the original result of [5]. □

3.2 The Effect of Framing on Sender Utility

Next, we analyze how much the sender can improve their utility via manipulating framing with a given signaling scheme π. To build intuition for this section's result, we start with an example to demonstrate how much belief shift via framing could lead to meaningful improvement in the sender's utility – is a substantial belief change necessary?

The following variant of the well-known prosecutor-judge example in Kamenica and Gentzkow [19] shows that the answer is *No*, meaning that a small change in framing can significantly change the sender's utility. Specifically, consider a Bayesian persuasion setting where a defendant may be either *innocent* or *guilty* (two possible states of the world), and the judge (receiver) decides whether to *acquit* or *convict* the defendant, receiving utility 1 for the just action and 0 otherwise. The prosecutor (sender) observes the defendant's true state and can signal accordingly to maximize their utility, which is 1 for a conviction regardless of the state. Now suppose the prosecutor, when innocent, always recommends acquittal, and when guilty, recommends either action with probability 0.5. If the prosecutor's prior belief over the states is $[0.67, 0.33]$ and the judge shares this belief (as is the case in Bayesian Persuasion), then the prosecutor achieves utility 0. If, however, the prosecutor can use framing (style of argument/language) to slightly alter the judge's belief to $[\frac{2}{3}, \frac{1}{3}]$, then the prosecutor's utility under the same signaling scheme jumps to 0.66! This is because the sender's utility as a function of the receiver's belief is not continuous for the given signaling scheme. This discontinuity highlights the power of leveraging framings: for a fixed signaling scheme, slightly altering the receiver's belief by framing can have a major impact on the receiver's action and hence the sender's utility.

To formalize this, let $U_\pi(\mu)$ denote the sender's expected utility (3) under fixed signaling scheme π when the receiver's prior belief is μ. In the prosecutor-judge instance, this function is discontinuous at $\mu = [\frac{2}{3}, \frac{1}{3}]$. We show below that such discontinuities occur in general instances almost surely for a large class of signaling schemes, specifically, schemes in which some signal s is sent with positive probability at every state. In other words, schemes not satisfying this condition must be "very revealing": upon observing any signal s, the receiver can rule out at least some state(s) with full confidence. The proof of this result is in Appendix A.3.

Proposition 3 (Discontinuous sender utility)**.** *Consider any signaling scheme π in which there exists a signal $s_0 \in \mathcal{S}$ such that $\pi(s_0|\omega) > 0$ for every state $\omega \in \Omega$. Then the sender's expected utility $U_\pi(\mu)$ as a function of the receiver's prior belief μ is* generally discontinuous *in the following sense: for any μ_0, for u and v sampled from any continuous distribution over utility functions, $U_\pi(\mu)$ is discontinuous in $\mu \in \Delta(\Omega)$ with probability 1. This holds even if the sender utility $u(a, \omega)$ is independent of ω.*

This discontinuity also implies that the sender's utility is highly sensitive to errors in the belief oracle ℓ. Suppose with access to an imperfect oracle ℓ_ε, the optimal framing $\hat{c}$ happens to induce a belief $\hat{\mu}$ such that $U_\pi(\cdot)$ was discontinuous at $\hat{\mu}$. Then if this framing is deployed, even though the true induced prior μ^* is within distance ε to $\hat{\mu}$ (i.e., $|\mu^* - \hat{\mu}| \leq \varepsilon$), the discontinuity means that the realized utility can be arbitrarily far from the utility achieved under the imperfect oracle, regardless of how small ε is. In particular, there does not exist a scalar λ such that $|U_\pi(\hat{\mu}) - U_\pi(\mu^*)| \leq \lambda\varepsilon$. Indeed, this discontinuity is the fundamental reason for the hardness result shown in Theorem 1.

4 Framing-Signaling Joint Strategy Space

We now consider the joint design of framing and signaling scheme. Formally, given an instance $\mathcal{I} = (\mu_0, u, v)$, the sender selects a framing $c \in C$ and a signaling scheme π to maximize her own utility. This is a *strict* generalization of the standard Bayesian persuasion model, which is the special case where C is a singleton set hence only the signaling scheme π is a variable. As shown by Proposition 1 of Section 3, there is no benefit in randomizing over framings, so we will focus on deterministic framing henceforth.

The ability to design π offers the sender more freedom as compared to the framing-only strategy. Indeed, a key challenge in the restricted framing-only case was the inability to interpret signals as obedient action recommendations. It turns out that for the design of joint strategy, it is without loss of generality to focus on action-recommending signaling schemes, as we show below (the proof is deferred to Appendix B).

Observation 1. *To compute the optimal joint framing-signaling strategy (c^*, π^*), it suffices to consider signaling scheme π with a direct signal space, i.e., $\mathcal{S} = \mathcal{A}$, in which each signal $a \in \mathcal{A}$ recommends an obedient action to the receiver.*

Observation 1 allows us to restrict attention to the design of direct signaling schemes without loss of generality. This succinct representation gives rise to the following corollary about polynomial-time solvability under discrete framing space. The algorithm simply enumerates all possible $c \in C$, and for each c computes the corresponding optimal signaling scheme via the standard Bayesian persuasion linear program [11].

Corollary 2. *For a discrete framing space C, the optimal joint strategy (c^*, π^*) can be computed in $|C| \cdot \mathrm{poly}(|\mathcal{A}|, |\Omega|)$ time.*

Similar to the preceding section, next we answer two key questions: (1) the effect of framing on the sender's utility and the sensitivity of this utility to errors in the belief oracle, and (2) the efficient computability of the optimal joint strategy with this oracle. Our results highlight key differences between joint design and framing-only design. The joint design problem turns out to be more tractable.

4.1 The Effect of Framing on Sender Utility

To compare with Proposition 3, we start our analysis by analyzing the continuity property of the optimal sender utility as a function of the framing-induced receiver prior belief μ. It is not difficult to see that, in the optimal joint design, given any framing-induced receiver prior belief μ, the sender's corresponding optimal signaling scheme is the one that optimally accompanies this receiver prior belief. We denote this signaling scheme as π^*_μ and the resulting sender utility $U^*(\mu)$ as a function of the framing-induced receiver prior belief μ. Observation 1 shows that $U^*(\mu)$ can be efficiently computed for any $\mu \in \Delta(\Omega)$ using the following linear program, with the constraints referred to as the *obedience* constraints as is standard in the literature [1]:

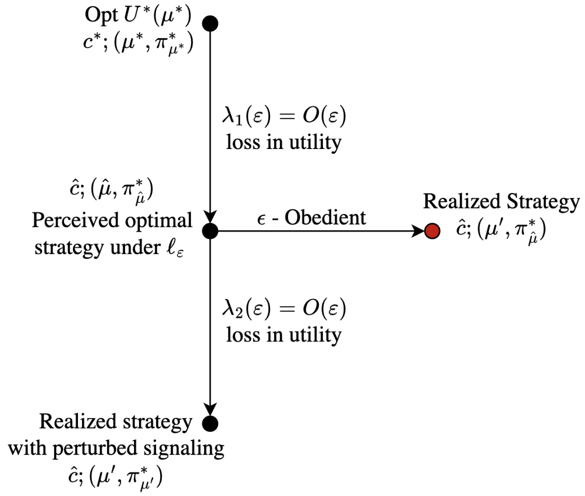
$$\begin{aligned} U^*(\mu) \ :=\ & \max_{\pi:\Omega\to\Delta(\mathcal{A})} \sum_{\omega\in\Omega}\sum_{a\in\mathcal{A}} \mu_0(\omega)\pi(a|\omega)u(a,\omega) && (6) \\ & \text{s.t.} \quad \forall a, a' \in \mathcal{A}\times\mathcal{A} : \sum_{\omega\in\Omega} \mu(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0. && (7) \end{aligned}$$

The $U^*(\mu)$ above tracks the sender's utility as a function of the receiver's prior belief μ under an optimal signaling scheme. It is worthwhile to compare $U^*(\mu)$ to the earlier function $U_\pi(\mu)$ introduced in Section 3 to capture the same quantity but with a fixed signaling scheme π. In contrast to the discontinuity of $U_\pi(\mu)$ shown in Proposition 3, our following Theorem 2 (proof deferred to Appendix B.2) shows that once π becomes part of the strategy space, the sender's optimal utility becomes continuous in the receiver's prior belief μ.

Theorem 2 (Continuous sender utility)**.** *The sender's utility $U^*(\mu)$, as defined in Eq.* (6)*, is a locally Lipschitz continuous function of the induced receiver belief μ within the interior of the belief simplex $\Delta(\Omega)$.*

Proof Sketch. The high-level idea is a sensitivity analysis for the linear program (6)(7), where we want to show that the optimal objective $U^*(\mu)$ of the linear program cannot change a lot when the parameter μ is slightly perturbed. In particular, let π^*_μ be an optimal solution of the linear program when the parameter is μ. We modify π^*_μ slightly to be another solution $\tilde{\pi}$ that satisfies the obedience constraint (7) simultaneously for all parameters μ' that are close to μ (in particular, $\|\mu' - \mu\|_1 \leq \varepsilon$). Such modification is possible due to the conditions that (1) every action $a \in \mathcal{A}$ of the receiver is strictly inducible by some belief; (2) $\mu(\omega) > 0$ for every $\omega \in \Omega$. Since the modification is small, the utility of $\tilde{\pi}$ is only slightly worse than the utility of π^*_μ, which is $U^*(\mu)$. So, the optimal objective $U^*(\mu')$ for μ' cannot be too much worse than $U^*(\mu)$, which establishes the continuity of the $U^*(\cdot)$ function. See details in Appendix B.2. □

Theorem 2 has useful implications on understanding the sender's utility loss due to a noisy belief oracle ℓ_ε. To illustrate this, it is helpful to relax the obedience constraint. Formally, we say a recommended action a is *ε-obedient* for a receiver with prior belief μ if for every action $a' \in \mathcal{A}$, $\sum_\omega \mu(\omega)\pi(a|\omega)[v(a,\omega) - v(a',\omega)] \geq -\varepsilon$.[3] Now let $(\mu^*, \pi^*_{\mu^*})$ denote the optimal joint strategy under the perfect oracle, with framing c^* inducing $\mu^* = \ell(c^*)$. Choosing c^* under the noisy oracle ℓ_ε will result in a perceived belief $\ell_\varepsilon(c^*)$ within the ball $B_\varepsilon(\mu^*) = \{\mu \in \Delta(\Omega) : \|\mu - \mu^*\| \leq \varepsilon\}$, and due to Theorem 2, the perceived utility will be within $\lambda_1\varepsilon$ of the optimal, where λ_1 is the Lipschitz continuity parameter of U^*, hence $|U^*(\ell_\varepsilon(c^*)) - U^*(\mu^*)| \leq \lambda_1\varepsilon$. If, under a noisy oracle, we solve $\max_\mu U^*(\mu)$ as defined in Program (6) (or Program (8) below for joint design) with exact obedience constraint (i.e., 0-obedience), then the optimal framing $\hat{c} \in C$, with corresponding perceived belief $\ell_\varepsilon(\hat{c}) = \hat{\mu}$ satisfies: $U^*(\hat{\mu}) \geq U^*(\ell_\varepsilon(c^*)) \geq U^*(\mu^*) - \lambda_1\varepsilon$.


Opt $U^*(\mu^*)$
$c^*; (\mu^*, \pi^*_{\mu^*})$
$\lambda_1(\varepsilon) = O(\varepsilon)$
loss in utility
$\hat{c}; (\hat{\mu}, \pi^*_{\hat{\mu}})$
Perceived optimal
strategy under ℓ_ε
ϵ - Obedient
Realized Strategy
$\hat{c}; (\mu', \pi^*_{\hat{\mu}})$
$\lambda_2(\varepsilon) = O(\varepsilon)$
loss in utility
Realized strategy
with perturbed signaling
$\hat{c}; (\mu', \pi^*_{\mu'})$

When using framing $\hat{c}$ in practice, however, the actual belief induced by $\hat{c}$ is $\mu' = \ell(c) \neq \hat{\mu}$ where $||\mu' - \hat{\mu}|| \leq \varepsilon$. Since the signaling $\pi^*_{\hat{\mu}}$ is unchanged, in shifting the induced belief from $\hat{\mu}$ to μ', it is evident that the obedience constraints, which were exactly satisfied for $\hat{\mu}$, are violated by at most ε for μ' – in other words, the deployed scheme is ε-obedient. If the sender was aware of the oracle noise and wished to be conservative, they could slightly modify the $\pi^*_{\hat{\mu}}$ scheme to make it exactly obedient for μ': this is feasible because μ' is close to $\hat{\mu}$, but will cause an additional $\lambda_2(\varepsilon)$ loss to the sender's utility, where $\lambda_2(\varepsilon) = O(\varepsilon)$. We illustrate this visually on the left and state it formally below:

Corollary 3. *The realized utility in facing an ε-obedient receiver under a joint optimal strategy based on a noisy oracle ℓ_ε is at most $\lambda_1\epsilon = O(\varepsilon)$ away from the optimal utility for an exactly obedient receiver. In slightly modifying the signaling scheme of this strategy, the strategy can be exactly obedient and at most $\lambda_1(\varepsilon) + \lambda_2(\varepsilon)$ away from optimal, with $\lambda_2(\varepsilon) = O(\varepsilon)$.*

We will focus on the design of ε-obedient signaling schemes throughout this section. This is primarily for mathematical convenience since under mild regularity assumptions, the loss in obedience constraint can be transferred into the loss in objectives with the same order. The main idea is to compute strictly obedient schemes by adding an $O(\varepsilon)$ amount of margin to obedience constraints, hence this scheme is still obedient even when the receiver's prior belief is off by an ε amount. What remains is then to bound the sender's utility loss due to imposing a stricter

[3] As per convention, ε-obedience means the obedience constraints may be violated by at most ε. The notation $(-\varepsilon)$-obedience thus means the obedience constraints are strictly satisfied with an ε margin.

obedience constraint, which can be bounded as $O(\epsilon)$ under mild regularity assumptions on the problem instance. See relevant studies in, e.g., [34, 22].

4.2 Optimizing the Joint Strategy under a Continuous Framing Space

Under discrete framing space, the optimal joint strategy can be computed in time linear in $|C|$, the size of framing space. This space in reality can be enormous, hence it is more insightful to consider continuous yet naturally structured framing space. Specifically, we consider the case where the framing-induced belief space $B = \{\ell(c) : c \in C\} \subseteq \Delta(\Omega)$ is convex. This leads to the following optimization problem with a linear objective subject to *bi-linear* constraints:

$$\max_{\mu \in B} U^*(\mu) = \max_{\mu \in B,\, \pi:\Omega \to \Delta(\mathcal{A})} \sum_{\omega \in \Omega} \sum_{a \in \mathcal{A}} \mu_0(\omega)\pi(a|\omega)u(a,\omega) \tag{8}$$

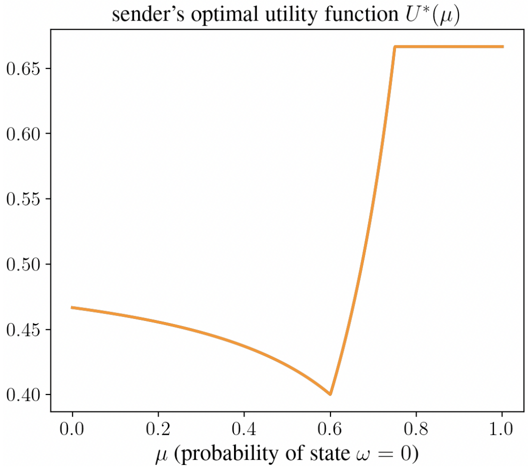
$$\text{s.t.} \qquad \forall\, a, a' \in \mathcal{A} \times \mathcal{A} : \sum_{\omega \in \Omega} \mu(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0.$$

Bi-linear optimization problems are well-known to be challenging to solve in general. In the following example, we illustrate this challenge by demonstrating the non-convexity and non-concavity of $U^*(\mu)$ even in simple instances.

Example 1 (Non-convexity and non-concavity of $U^*(\mu)$)**.** *Consider an instance with* 2 *states* $\Omega = \{0, 1\}$ *and 3 actions* $\{0, 1, 2\}$, *with the following utility matrices for the sender and the receiver (rows are actions, columns are states):*

$$u = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0.2 & 0.2 \end{bmatrix}, \qquad v = \begin{bmatrix} 0.65 & 0.15 \\ 0.60 & 0.30 \\ 0.10 & 0.50 \end{bmatrix}.$$

The sender has prior $\mu_0 = (\frac{1}{3}, \frac{2}{3})$ *for the two states. We use the probability of state* 0 *to denote the receiver's belief* $\mu \in [0, 1]$*. The sender's optimal utility function* $U^*(\mu)$ *is plotted to the right. It is continuous but not convex, concave, or quasi-concave.*



The example above hints at potential barriers for the optimization problem. Indeed, recall that in the framing-only strategy setting of Section 3, there is no FPTAS for the problem, unless P = NP, even when there is no constraint on the induced belief space B, i.e., $B = \Delta(\Omega)$ is the entire simplex. But interestingly, we obtain a positive result for unconstrained belief space in the joint design case: there is a *bi-criteria* FPTAS with $\frac{1}{|\Omega|}$ loss in objective and ε loss in obedience constraint with poly $(|\Omega|, |A|, 1/\varepsilon)$ running time. Further, when the sender's utility is state-independent and only depends on the receiver's action (e.g., the prosecutor-judge example), the exactly optimal policy can be computed in polynomial time.

Theorem 3 (Positive results for unconstrained belief space)**.** *For the optimal joint strategy design with unconstrained belief space* $B = \Delta(\Omega)$*, the following hold:*

- *A* $\left(1 - \frac{1}{|\Omega|}\right)$ *multiplicative-approximation of the optimal joint strategy utility can be computed in poly-time for a receiver with* ε*-obedience;*

- *If the sender's utility is state-independent, i.e.,* $\forall a, \forall(\omega, \omega')$, $u(\omega, a) = u(\omega', a)$, *then the exact optimal strategy can be computed in poly-time.*

Proof. Beginning with the first claim, recall that we consider sender utilities to be positive (this is without loss of generality since the sender utility is linear in $u(a, \omega)$, allowing us to normalize as needed). Let $a^u(\omega) = \arg\max_a u(a, \omega)$ and $a^v(\omega) = \arg\max_a v(a, \omega)$ denote the optimal action for the sender and receiver at state ω respectively. Further, let $\omega_{min} = \arg\min_\omega \mu_0(\omega) u(a^u(\omega), \omega)$; it captures the "least" important state for the sender assuming sender-optimal action at each state.

Since $B = \Delta(\Omega)$, consider using framing to induce a belief $\mu_c(\omega_{min}) = 1 - \varepsilon$ and $\frac{\varepsilon}{n-1}$ for all other states; let the signaling scheme π deterministically recommend the received optimal action at ω_{min} and sender optimal actions at all other states. In other words:

$$\pi(a^v(\omega_{min})|\omega_{min}) = 1 \quad \text{and} \quad \forall \omega \neq \omega_{min} \, : \, \pi(a^u(\omega)|\omega) = 1.$$

Since the sender's utility is non-negative, if the receiver follows the recommended actions outlined by this scheme, the sender is guaranteed to achieve at least the following utility:

$$\sum_{\omega \neq \omega_{min}} \mu_0(\omega) u(a^u(\omega), \omega) \geq u_{max} - \frac{1}{|\Omega|} u_{max} = \left(1 - \frac{1}{|\Omega|}\right) u_{max}$$

where we note that the maximal possible utility achievable by the sender is $u_{max} = \sum_\omega \mu_0(\omega) u(a^u(\omega), \omega)$ and by the pigeonhole principle, $\mu_0(\omega_{min}) u(a^u(\omega_{min}), \omega_{min}) \leq \frac{1}{|\Omega|} u_{max}$. We now show that following the recommended actions is ε-obedient for the receiver. Indeed, for a recommended action a and any other action a', the ε-obedience expression for this pair under the scheme π is:

$$(1 - \varepsilon)\pi(a|\omega_{min})[v(a, \omega_{min}) - v(a', \omega_{min})] + \sum_{\omega \neq \omega_{min}} \frac{\varepsilon}{n-1} \pi(a|\omega)[v(a, \omega) - v(a', \omega)] \; \geq \; -\varepsilon.$$

When the receiver gets recommended action $a^v(\omega_{min})$, this expression becomes at least $(1 - \varepsilon)$ $[v(a^v(\omega_{min}), \omega_{min}) - v(a', \omega_{min})] - \varepsilon \geq -\varepsilon$ since at state ω_{min}, action $a^v(\omega_{min})$ is optimal for the receiver. Conversely, if the receiver is recommended some action $a \neq a^v(\omega_{min})$, then the expression is: $\sum_{\omega \neq \omega_{min}} \frac{\varepsilon}{n-1} \pi(a|\omega)[v(a, \omega) - v(a', \omega)] \geq -\varepsilon$ since the utilities are bounded to $[0, 1]$. So, ε-obedience is satisfied.

For the second claim, we can make use of an additional assumption: the sender utility is state-independent. This means that their utility is maximized if the receiver takes some action a^u at *all* possible states. We also recall from Section 2 that any action of the receiver is inducible – that is, for any action $a \in \mathcal{A}$, there exists some belief μ_a wherein taking a is optimal for the receiver. For the sender optimal action a^u, let μ_{a^u} be the belief where a^u is optimal for the receiver. Indeed, μ_{a^u} can be computed using the following set of linear constraints:

$$\forall a' \in \mathcal{A} \, : \, \sum_\omega \mu(\omega)[v(a^u, \omega) - v(a', \omega)] \geq 0.$$

Since $B = \Delta(\Omega)$, we can choose a framing c to induce belief μ_{a^u}. We accompany this framing with an uninformative signaling scheme π. Such a scheme reveals no information about the realized state. For example, for a given action a_1, $\pi(a_1|\omega) = 1$ for all ω is an uninformative scheme. Under this joint strategy, the receiver's belief is always μ_{a^u}, where their best-response is to take the sender optimal action a^u. This is thus an optimal strategy for the sender. □

Theorem 3 studies unconstrained belief space. Next, we turn to the constrained belief space and present a *bi-criteria* Quasi-Polynomial Time Approximate Scheme (QPTAS) in Theorem 4.

While the hardness of the exact optimization problem is an interesting open question, we view the existence of a QPTAS as evidence of better tractability of the joint optimization problem compared to the framing-only variant in Section 3, which was shown to be reducible from the *Independent-Set* problem that has no known QPTAS.

Theorem 4 (QPTAS for constrained belief space)**.** *For any instance $\mathcal{I} = (\mu_0, u, v)$ with convex belief space $B \subseteq \Delta(\Omega)$, for any $\varepsilon > 0$, there exists a* $\text{poly}(|\Omega|^{\frac{\log|\mathcal{A}|}{\varepsilon^2}})$*-time algorithm that can compute an ε-obedient joint strategy with at least as much utility as the optimal joint strategy under exact obedience.*

Proof. Let $\mu^* \in B$ and $\pi^* : \Omega \to \Delta(A)$ be the optimal induced prior belief and signaling scheme for this instance under exact obedience constraints. If we were to draw n samples from μ^*, the resulting empirical distribution, denoted $\hat{\mu}$, would be an n-uniform distribution – i.e., each entry of $\hat{\mu}$ is a multiple of $\frac{1}{n}$. Because $\mathbb{E}\big[\sum_\omega \hat{\mu}(\omega)\pi^*(a|\omega)[v(a,\omega) - v(a',\omega)]\big] = \sum_\omega \mu^*(\omega)\pi^*(a|\omega)[v(a,\omega) - v(a',\omega)] \geq 0$, by Hoeffding's inequality, we have

$$\forall a, a' \in \mathcal{A} \times \mathcal{A}, \quad \Pr\Big[\sum_\omega \hat{\mu}(\omega)\pi^*(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] < -\varepsilon\Big] \leq \exp\left(-2n\varepsilon^2\right).$$

Taking a union bound over all $|\mathcal{A}|^2$ pairs (a, a'), we have $\sum_\omega \hat{\mu}(\omega)\pi^*(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq -\varepsilon$ satisfied for all pairs of (a, a') with probability at least $1 - |\mathcal{A}|^2 \exp\left(-2n\varepsilon^2\right)$, hence $\hat{\mu}$ in conjunction with π^* satisfies ε-obedience constraints. Pick $n = \frac{\log|A|}{\varepsilon^2}$. The probability $1 - |\mathcal{A}|^2 \exp\left(-2n\varepsilon^2\right)$ will be positive. This means that there must exist an n-uniform distribution $\hat{\mu}$ satisfying ε-obedience in conjunction with π^*. Note that the sender's utility under the $(\hat{\mu}, \pi^*)$ strategy is the same as the (μ^*, π^*) strategy because the sender's utility depends on π^* but the receiver's prior belief.

Now consider the following algorithm: enumerate over all n-uniform distributions in B, and for each, solve the optimal signaling linear program (6)-(7) but with a relaxed ε-obedience constraint. Return the solution with the best sender utility. This solution must be weakly better than the ε-obedience solution $(\hat{\mu}, \pi^*)$ mentioned above, which is therefore weakly better than the optimal solution (μ^*, π^*).

We then consider the runtime of the algorithm. The runtime depends on the number of n-uniform distributions and the time to check whether each distribution is included within the inducible belief set B. Since B is convex, checking this inclusion can be done in poly-time. As for the number of n-uniform distributions, since the probability of each state ω can take on n possible values $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$ and the sum must equal 1, it is equivalent to placing n elements into $|\Omega|$ distinct buckets. Thus, the number of possible distributions is at most $\binom{n+|\Omega|-1}{|\Omega|-1}$. For a fixed $|\Omega|$ and n growing large, this quantity is a polynomial in n of degree $|\Omega|-1$, which is clearly upper bounded by $O(|\Omega|^n)$. For a fixed n and as $|\Omega|$ grows large, we have $\binom{n+|\Omega|-1}{|\Omega|-1} = \binom{n+|\Omega|-1}{n} \leq \frac{(n+|\Omega|-1)^n}{n!} = O(|\Omega|^n)$. Thus, the runtime of this algorithm is bounded by $\text{poly} \cdot O(|\Omega|^n) = \text{poly}(|\Omega|^{\frac{\log|\mathcal{A}|}{\varepsilon^2}})$. □

5 Empirical Studies with Large Language Models

Two key aspects of our theoretical exploration of framing and signaling are worth noting: (a) our model rests on access to a (possibly noisy) framing-to-belief oracle $\ell : c \to \mu_c$; (b) our theoretical studies highlight the computational challenges of optimizing over the framing space C, even with access to such an oracle. Note that the first issue is primarily a challenge when the decision-maker/receiver is human. When AI agents are used as the decision-maker, an increasingly common

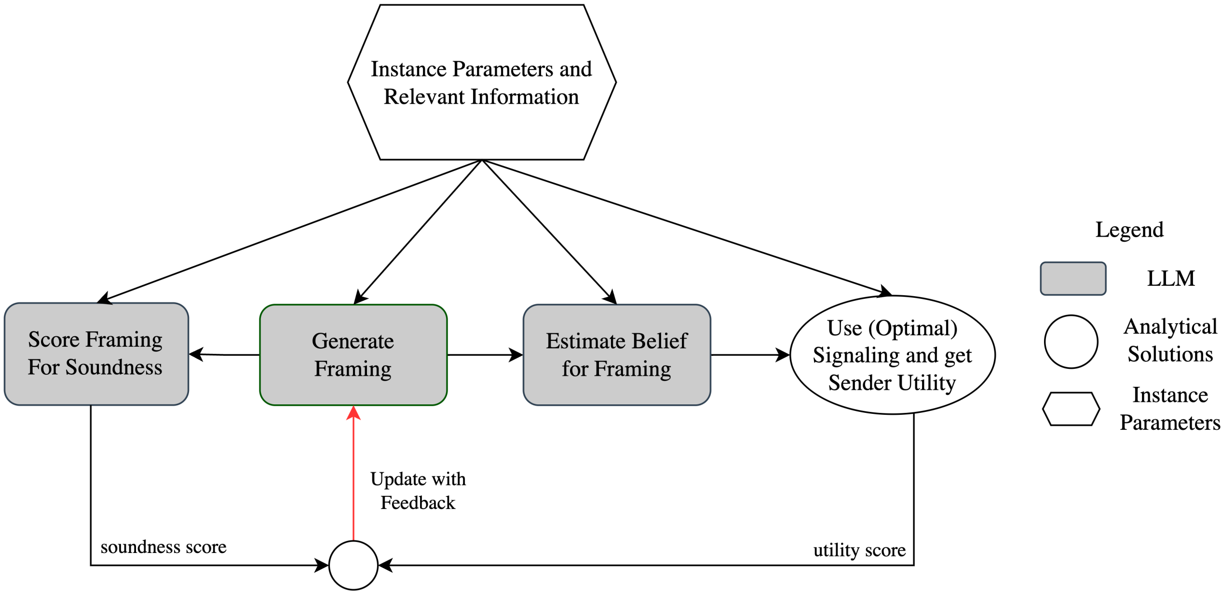

Instance Parameters and Relevant Information
Score Framing For Soundness
Generate Framing
Estimate Belief for Framing
Use (Optimal) Signaling and get Sender Utility
Update with Feedback
soundness score
utility score
Legend
LLM
Analytical Solutions
Instance Parameters

Figure 1: Diagram of our proposed framework for optimizing framing and signaling. It includes LLMs searching the framing space, verifying it for correctness, and generating framing-induced beliefs. It also includes poly-time analytical solvers to compute optimal signaling for a given belief.

scenario, this is an easier task [15]. In any case, this section highlights that empirical methods leveraging Large Language Models (LLMs) can address these two issues in full generality. Specifically, we empirically show that LLMs can not only search over the framing space efficiently, but can also be used to approximately capture the mapping ℓ, even for human decision makers.

5.1 Methodology

We present our proposed approach for optimizing in the framing space, either by itself or jointly with signaling, in Figure 1. LLMs perform two crucial roles here: generating framing and refining it based on feedback, and estimating the induced belief of a (possibly human) receiver given any framing. These are complemented with an additional LLM module that verifies the soundness of any framing, and analytical solvers to compute the sender's utility given a quantitative receiver belief and signaling scheme. We discuss these roles and the overall methodology below:

Instance information: Information relevant for optimal framing and signaling is not just the quantitative parameters like (μ_0, u, v) but also qualitative descriptions about the setting, the receiver and sender, their preferences, and any related meta information. This is because the belief induced by framing is not an explicit mathematical process. Rather, it captures how a string describing some aspect of the instance will be perceived by a given receiver, factoring in social norms, environmental, and personal factors.

Estimating the belief oracle ℓ: A key responsibility of LLMs in our flow is estimating how a framing would influence a given receiver's belief. In settings where decision-making has been delegated to LLMs, this involves simply querying the LLM and the approximation is essentially exact. When the underlying decision-maker is a human agent, a nascent line of economics research argues that LLMs can approximate this agent in various settings [18, 24], including as a statistical proxy to model human beliefs [26]. This approach involves endowing the LLM with information about the human agent it is modeling - in our case, this is the relevant information about the receiver, including demographics, preferences, and backgrounds. Second, what we precisely require from the LLM is a quantitative value on the receiver's belief. This can be gathered by either: asking the LLM to generate one of $|\Omega|$ tokens corresponding to each state and recording the log probabilities, or asking it to directly return numerical probabilities. Cruz et al. [7] show that on distributional (non-factual) questions, the first approach leads to accurate (correctly returns the most-likely outcome) but highly-uncalibrated answers (the log-probabilities are far from the true distribution), while eliciting numerical probabilities in a chat-style prompt results in better outcomes. We thus elicit numerical probabilities in our framework and comment on experimental observations about this approach in Section 5.2.

Validation of framing: Using LLMs to generate framing risks hallucinations; in our context, this would be any information in the framing that is incorrect or inconsistent. For example, if asked to design a framing for a Nike basketball shoe, the LLM generating a blurb highlighting their collaboration with a non-existent NBA team or player would be incorrect. In general, the *soundness* may be more nuanced than a binary outcome; the framing could take certain liberties that, while not blatantly incorrect, may be undesired. As such, we propose using an LLM to score soundness with a real value between 0 and 1. This LLM module is given in-context information about the instance along with the generated framing. To calibrate the scores, we specify a few labeled (*framing*, *score*) examples in the prompt since few-shot approaches have been successful in the literature [23, 3]. The correctness score is part of the feedback to the framing-generating LLM.

Computing sender utility: When the strategy space is framing-only, the signaling scheme is given, and computing the sender utility for a framing-induced belief μ_c is just carrying out the algebra in equations (1) and (2). In the case of a joint strategy space, computing the optimal signaling scheme for a given receiver belief can be solved by the linear program specified in (6). In either case, poly-time analytical approaches can compute the sender's utility given a belief μ_c.

Generating framing: Building on the success of in-context learning via "textual gradients" [30], we propose that an LLM generate a framing based on instance-relevant information and a language-specified task, and iteratively refine it through feedback. The relevant information includes profiles of the receiver, their preferences, and those of the sender. We find it sufficient to present this information qualitatively. The task description defines key parameters for the framing, such as word count and style, while also outlining the refinement process and feedback. For each generated framing, the induced prior is estimated and then used to compute the corresponding sender utility; this is scaled by the soundness score. This final quantitative score is supplemented with the reasoning behind the generated belief and soundness score to construct the feedback string. The LLM's context is updated with this feedback, prompting it to generate a refined framing.

5.2 A Real-Estate Case Study

To demonstrate our proposed optimization framework, we consider the following scenario: A realtor (sender) works with a potential home-buyer (receiver) and may show them houses with various features (the world states). The realtor observes the true state of the property, and signals the buyer through an action recommendation (buy or not buy). The buyer observes some description of the realtor (the framing) and the recommendation signal. Both of these influence their belief about a property this realtor would show/specialize in. The buyer takes their optimal action based on this belief and their utility. Numerical and prompting details about the instance are in Appendix C; but in general, the realtor wants the buyer to purchase expensive homes and not cheap ones; the buyer wants cheap homes that fit their desired criteria.

Searching the Framing Space. How should the realtor pitch themselves to the potential buyer? How does this change depending on the buyer? This relates to finding the optimal framing for a given instance. More precisely, we consider two instances of this problem, corresponding to two potential buyers: "Henry" and "Lilly". Both instances share the same realtor, "Jeremy", and the instance description contains all relevant information about Jeremy, alongside profiles of Henry and Lilly. See below for the descriptions contained in the instance information:

- Realtor Jeremy: *Jeremy Hammond is Male and 42. Worked with the our firm for 2 years, Worked previously as a realtor for 6 years, and a contractor before that. Lives with his wife and 3 kids and a dog and a cat in Downtown Boston. Hobbies include playing the drums, spending time with kids, hiking, and backyard gardening. Active member of his Home Owners ́ Association.*
- Buyer Henry: *Henry lives in Boston and is an avid outdoors person who enjoys hiking and being in nature. For him, a "good" house has low maintenance, affords easy access to trails, biking, running etc, and far from hustle of the main city. He is single and lives by himself - so he is indifferent to school districts, etc. A bad house is generally one in a very family-oriented neighborhood with stingy HOA rules, maintenance, lawn care expectations and so on. For him, cheap is anything less that costs less $500,000, with expensive being houses above this.*
- Buyer Lilly: *Lilly is moving to Boston with her husband, 3 young kids and a dog. She and her family are looking for a spacious house in the suburbs with good schools for their kids, a nice yard for her dog, and friendly community-focused neighbours. This is what constitutes a "good" house for her. Smaller homes, those in not-so-great school zones, or those in busy and loud areas of the city near Downtown are "bad" in her eyes. For them, anything costing less that $650,000 is considered cheap, with those above considered expensive.*

A good framing is a personalized description of Jeremy for the pertinent buyer that induces a favourable (to the realtor) belief about the types of houses Jeremy could show. For both buyers, there are four possible states, corresponding to the product of (good, bad) and (cheap, expensive). To illustrate the full range of our framework, we consider the joint optimization setting where both the framing and signaling scheme can be optimized. For this experiment, we use GPT-4o-mini [28] for the LLM portions of the framework, and SciPy [32] for the analytical parts. The results for each instance are presented in Tables 1 and 2. We compare the best framing found by our framework with the following baselines: no framing wherein the receiver and sender have the same prior, using the default profile of Jeremy contained in the instance description as the framing, and the optimal joint strategy when any belief in $\Delta(\Omega)$ can be induced (i.e., Program (8) with $B = \Delta(\Omega)$). This last baseline is computed analytically.

We note that in both instances, the LLM generated framing produced a higher utility than using the default description and the standard persuasion baseline where both parties share the same utility. In analyzing the generated responses, we observe that these LLM framing selectively highlight aspects of Jeremy's profile that may appeal to each buyer, while omitting that which does not. For example, the framing for Henry emphasizes Jeremy's love of the outdoors and pitches his contractor background as helpful in finding low-maintenance properties. The framing for Lilly on the other hand, frames this background as helpful in finding spacious properties and highlights him being a dog-owner, just like Lilly. Interestingly, the utility generated by the optimal framing for Henry comes close to the theoretical optimal when *any* belief is inducible.

Estimating Receiver Belief under Framing. The second key role that LLMs play in our framework is quantifying the receiver's belief for a given framing – indeed, this is instrumental to the results presented above. To what extent does the LLM capture the beliefs of the decision-

Table 1: Framing Optimization Result for the Henry Instance

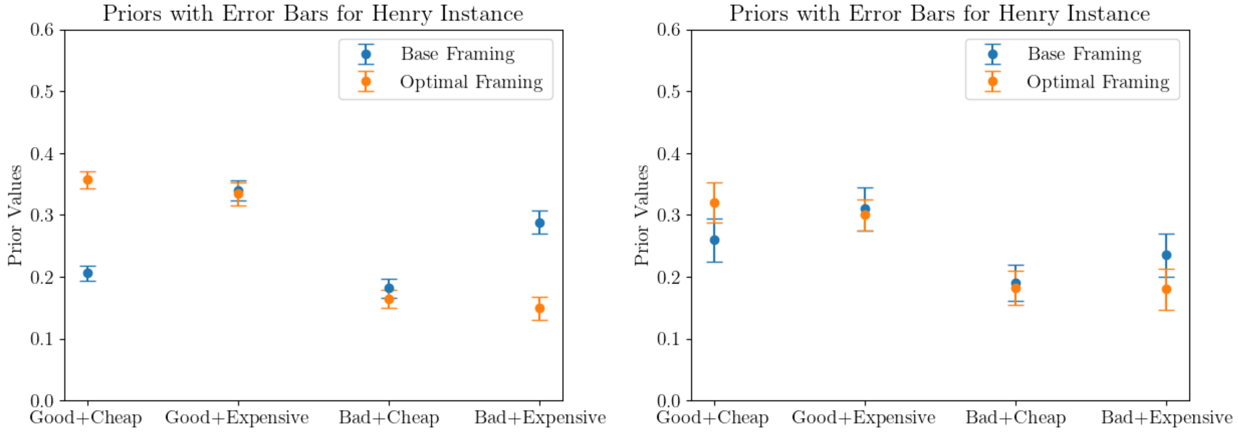
Framing for the "Henry" Instance	**Utility**
No framing - receiver prior equal to sender prior	0.28
Realtor Jeremy Profile from the Instance Description	0.30
Best LLM Framing: *Meet Jeremy Hammond, a dedicated realtor with over 8 years of experience, specializing in finding the perfect homes for outdoor enthusiasts like you. Living in Downtown Boston, Jeremy understands the balance between city life and access to nature. With a background as a contractor, he ensures that every property meets your low-maintenance needs. When he's not helping clients, you can find him hiking local trails or enjoying his backyard garden. Trust Jeremy to help you discover a home that complements your active lifestyle while staying within your budget.*	0.40
Analytical Upper Bound (Optimal Joint Strategy with $B = \Delta(\Omega)$)	0.41

Table 2: Framing Optimization Result for the Lilly Instance

Framing for the "Lilly" Instance	**Utility**
No framing - receiver prior equal to sender prior	0.33
Realtor Jeremy Profile from the Instance Description	0.33
Best LLM Framing: *Introducing Jeremy Hammond, a seasoned realtor with 8 years dedicated to helping families find their dream homes in Boston's suburbs. With a rich background as a contractor, Jeremy excels in identifying spacious, family-friendly properties with excellent school districts—just what you need for your kids. As a fellow dog owner, he knows the importance of a great yard and a welcoming neighborhood. Trust Jeremy to leverage his local expertise and commitment to family values as he guides you to affordable yet quality homes that fit your family's lifestyle.*	0.42
Analytical Upper Bound (Optimal Joint Strategy when $B = \Delta(\Omega)$)	0.46

makers they are proxying for? And in scenarios where decision-making has been delegated to LLMs, to what extent are the LLM-generated beliefs (which can essentially be considered ground truth) consistent with the actions they take? We aim to study these two questions within the context of our real estate case study, starting with the first.

We recruit 50 American participants via Prolific and gather their beliefs under the same framing and instructions given to LLMs for the real estate case study. Each survey fixes a buyer profile (Henry or Lilly) and a seller framing (description of the realtor, Jeremy). Participants were then asked to specify the probabilities the described buyer would assign to each of four possible house types. Specifically, there were four multiple-choice questions (corresponding to each possible state), with each answer option covering a probability range in 10% increments. We first report results for the Henry case (in Figure 2), comparing priors from the base realtor description with the LLM-generated optimal description.

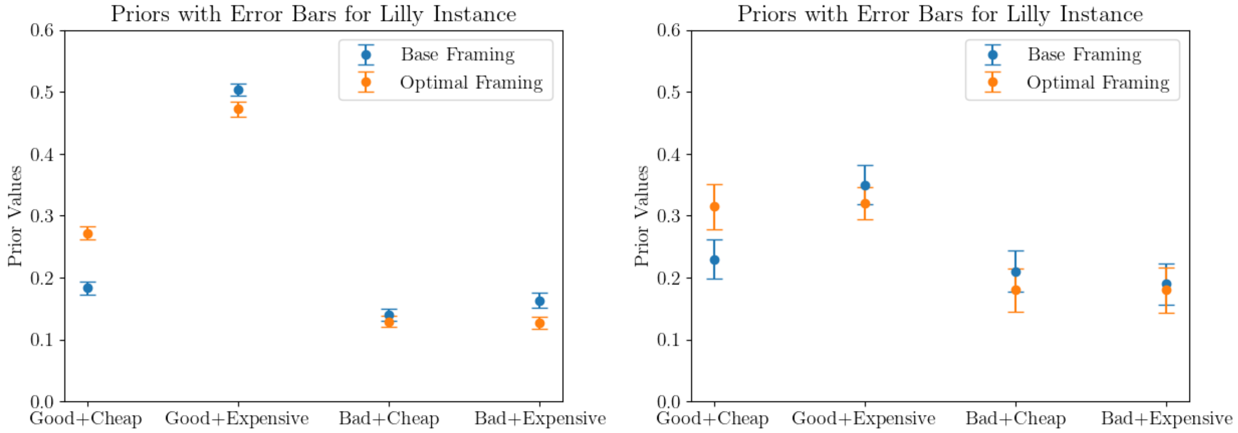


(a) **LLM Generated** belief distribution for Henry

(b) **Prolific Users'** belief distribution for Henry

Figure 2: Means (across 50 runs) from both LLM and human-generated beliefs with 90% confidence intervals on the **Henry** instance.

We notice that human responses have higher variance than GPT-4o at default temperatures.[4] The magnitude of belief change due to framing is smaller in human subjects as compared to LLMs. That said, the LLM proxy does capture the aggregate human responses at an ordinal level. Both rank "Good+Cheap" as more likely under the optimal framing as compared to the base framing and rank "Bad+Expensive" as more likely under the base framing than the optimal one. Further, the probabilities for "Good+Expensive" and "Bad+Cheap" outcomes do not shift much in either plot. We observe a similar trend for the Lilly instance, whose results are presented in Figure 3.



(a) **LLM Generated** belief distribution for Lilly

(b) **Prolific Users'** belief distribution for Lilly

Figure 3: Means (across 50 runs) from both LLM and human-generated beliefs with 90% confidence intervals on the **Lilly** instance.

[4] Although variance can be made 0 by using a low temperature, this can result in less effective outputs. We use the OpenAI default temperature of 0.7 in all our experiments.

Overall, this initial experiment indicates that LLM proxies can capture high-level trends present in human responses, but may be more confident and extreme in their results. The numeric differences in generated priors can be seen as the error in the belief oracle, whose implications we discussed in earlier sections. Given that human evaluation is both expensive and time-consuming, we envision a practical deployment of our system to do "cheap" iterations using LLM evaluators interspersed with more "expensive" validation loops using humans to calibrate errors.

Table 3: Results of 50 independent runs of the consistency experiment for the optimal framing generated for each instance.

Instance	**Pre-Belief Action**	**Post-Belief Action**	**Optimal Action for Generated Belief**
Henry	43 "buy" and 7 "not buy"	50 "buy" and 0 "not buy"	50 "buy" and 0 "not buy"
Lilly	50 "buys" and 0 "not buy"	50 "buy" and 0 "not buy"	50 "buy" and 0 "not buy"

We now turn to our second question, which is especially relevant when LLMs are themselves the decision makers: Do LLMs act in a way that is consistent with the beliefs they generate? First, we prompt the LLM with the receiver utility and the optimal realtor framing and ask what action the receiver would take given just this information. Note that this is *before* the LLM was asked to generate any beliefs. This reflects the initial instinct of the LLM, and we denote it as the *instinctive action*. Second, we consider the LLM *after* generating the belief as per our framework. We maintain the requisite prompts and the generated belief in-context, and give the LLM the receiver utility and ask it to make a decision. This *post-belief action*, is compared against both the instinctive action and the optimal action for each generated belief.[5] As we see from Table 3, post-belief actions are always consistent with the optimal action for each prior. The instinctive pre-belief actions are slightly less so for Henry, but perfectly match for Lilly.

5.3 An Advertising Case Study

To further evaluate our model and the concept of framing, we conduct a second empirical case study focused on advertising. We consider a clothing brand, *Himalaya* (inspired by Patagonia), known for durable, high-quality gear favored by outdoor enthusiasts. The brand is launching a new outerwear line targeting the style-conscious, casually active ath-leisure market. Their advertising campaign aims to attract this new demographic without compromising brand identity and includes: (1) a new motto, (2) a description of the new line, and (3) discount offers. The first two constitute the framing, while the third represents the signal. The full setup is detailed below.

There are four product categories based on two attributes: *(trendy, durable), (trendy, not durable), (not trendy, durable), (not trendy, not durable)*. Framing refers to the slogan and description used in the ad campaign. Buyers can choose from three actions: "buy on sale", "buy at regular price", or "not buy", and there are three corresponding signals: "Ad with Discount", "Ad without Discount", and "No Ad".[6] Both the user and brand receive zero utility from not buying. As a realistic constraint, trendy and durable products are never on sale[7], but still provide positive utility to both parties if purchased at regular price. The user receives negative utility from buying

[5] This optimal action can have randomness, since for each run, a different belief can be generated, which has a possibly different optimal action. In practice, however, this is not observed.

[6] "No Ad" means no slogan or description is shown. "Ad with Discount" presents the slogan and description, but no discount.

[7] This also ensures there is no dominant action for either party, making the instance non-degenerate.

durable but unfashionable products at regular price but is indifferent between buying on sale and not buying. If a product is neither trendy nor durable, the user prefers not to buy, while the brand prefers to sell it on sale. The numerical utilities are given below with rows denoting actions and columns states.

$$\text{Brand Utility} = \begin{bmatrix} \text{N/A} & 1 & 0.3 & 0.8 \\ 2.5 & 2 & 1.0 & 0.5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \;;\; \text{Target Constomer Utility} = \begin{bmatrix} \text{N/A} & 1 & 0 & -0.5 \\ 1 & 0.6 & -1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We describe the target customer and brand to the LLM as follows. These are used to generate new framing (the customer description is also used to estimate priors). The description of the outerwear line are taken directly from Patagonia's description of their rain-jackets.

- Target Demographic: *Our new target demographic are fashion-aware average mall consumers, who are casually into an active lifestyle. They are middle class, but are willing to pay a slight premium for quality and style. This is a segment that the athleisure market has dominated of late, with brands like Lululemon, Nike being the main players. Consumers here like the idea of performance wear (e.g., for hiking or skiing) but are not deeply familiar with or motivated by technical characteristics. What matters most is whether the clothes looks stylish in everyday environments like schools, cafés, or city streets. Functionality and durability is a nice bonus, but aesthetic appeal primarily drives their interest.*
- Brand Description: *Himalaya is fairly well known brand in the outdoor enthusiast, mountaineering, and adventure community. It has a reputation for bulletproof build quality and performance, and valuing sustainability. It is launching a new outerware line, that includes parkas, ski-jackets and ski-pants, windbreakers, and thermal layers. All products here are made with 100% post-consumer recycled nylon ripstop and without PFAS. They meet H2No Performance Standard for waterproofness and breathability. Fabric and inner membrace have durable water repellent (DWR) finish.*
- True Prior: $\mu_0 = [0.225, 0.125, 0.5, 0.15]$.[8]

In Table 4, we compare the current Patagonia slogan and verbatim description of one of their most popular jacket, against the best jointly generated slogan and description by the LLM. We present the sender utility that can be achieved using the optimal signaling for each slogan and description induced prior. We note that the LLM-generated description eschews significant technical details and specifications while still highlighting the durability and functionality of the product.

6 Discussion

This paper connects the rich literature on Bayesian signaling with mature ideas from behavioral economics and psychology which posit that the linguistic and contextual framing of information plays an important role in shaping the perceptions and beliefs of decision-makers. Traditionally, costly methods such as focus groups were required to explore the link between framing and belief formation. However, the emergence of LLMs provides a more efficient, systematic, and cost-effective alternative. Our work experimentally demonstrates this approach and taking this belief-generating process as a given, we further investigate the optimization properties of this problem. Our theoretical results demonstrate that while slight changes in framing can significantly improve sender utility in many settings, determining the optimal framing, with or without signaling, is a challenging problem. Here too, LLMs offer respite. We propose an optimization framework that uses LLMs

[8] This was obtained by asking the LLM about their prior over the current Patagonia products.

Table 4: Framing Optimization Result for the Advertising Instance

Framing (Slogan + Description) for New Clothing Line	**Utility**
Patagonia Slogan: *Build the best product.* Patagonia Product Desc: *Whether you're seeking boulders and alpine views or stalking rainbow trout, this low-profile, waterproof jacket will keep you dry and provide rain protection from the fork in the trail to the places less traveled, and back again. This 3-layer shell meets our H2No® Performance Standard for exceptional waterproof/breathable protection, and the fabric, membrane, and durable water repellent (DWR) finish are made without intentionally added PFAS.*	1.18
LLM Slogan: *Fashionably Functional: Gear Up for Your Next Adventure.* LLM Description: *Unveil a new era of outerwear with Himalaya's latest collection, designed for the style-savvy adventurer. Our parkas, ski jackets, and windbreakers are made from 100% post-consumer recycled nylon, perfectly marrying sustainability with modern aesthetics. With innovative waterproof technology and a tailored fit, these pieces are versatile enough for both urban exploration and outdoor escapades. Embrace every journey with confidence, knowing your gear is as stylish as it is reliable, providing the perfect companion for every adventure.*	1.40

to efficiently search the framing space, leveraging their ability to understand linguistic structure and learn in-context.

This work opens many interesting directions for future research. On the theoretical side, it remains open to determine the computational complexity of the framing-signaling joint optimization problem (we conjecture this to be NP-Hard). Empirically, more work is needed to better understand how LLM-generated beliefs match those of humans. Or how satisfied humans are with LLM-proxied decision-making in settings relevant to signaling. Answering these questions may require careful and nuanced human-subject experiments. Lastly, besides Bayesian persuasion, it would be instructive to combine the rich perspective of framing with other important signaling models such as cheap-talk [6, 13] and mediation [25].

References

[1] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[2] James Brand, Ayelet Israeli, and Donald Ngwe. Using gpt for market research. *Harvard Business School Marketing Unit Working Paper*, (23-062), 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[4] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020.

[5] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

[6] Vincent P. Crawford and Joel Sobel. Strategic Information Transmission. *Econometrica*, 50 (6):1431, November 1982.

[7] André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2024. URL https://openreview.net/forum?id=qrZxL3Bto9.

[8] Stefano DellaVigna and Matthew Gentzkow. Persuasion: empirical evidence. *Annu. Rev. Econ.*, 2(1):643–669, 2010.

[9] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

[10] James N Druckman. On the limits of framing effects: Who can frame? *The journal of politics*, 63(4):1041–1066, 2001.

[11] Shaddin Dughmi and Haifeng Xu. Algorithmic Bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, Cambridge MA USA, June 2016. ACM.

[12] Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, 76(1):117–130, 2012.

[13] Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic perspectives*, 10(3): 103–118, 1996.

[14] Yiding Feng, Wei Tang, and Haifeng Xu. Online bayesian recommendation with no regret. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 818–819, 2022.

[15] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

[16] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, Markus Mobius, and Divyarthi Mohan. Communicating with Anecdotes (Extended Abstract). *15th Innovations in Theoretical Computer Science Conference (ITCS)*, 287, 2024.

[17] Keegan Harris, Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. Algorithmic persuasion through simulation: Information design in the age of generative ai. *arXiv preprint arXiv:2311.18138*, 2023.

[18] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[19] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[20] Yan Leng. Can llms mimic human-like mental accounting and behavioral biases? *Available at SSRN 4705130*, 2024.

[21] Wenhao Li, Yue Lin, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. Verbalized Bayesian Persuasion, 2025. URL `https://arxiv.org/abs/2502.01587`.

[22] Tao Lin and Ce Li. Information Design with Unknown Prior (Extended Abstract). *In 16th Innovations in Theoretical Computer Science Conference (ITCS)*, 325:72:1–72:1, 2025.

[23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train prompt and predict: A systematic survey of prompting methods in nlp. *ACM Computing Surveys*, 55(9):1–35, 2023.

[24] Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

[25] Roger B. Myerson. Game theory: analysis of conflict. *The President and Fellows of Harvard College, USA*, 66, 1991.

[26] Keiichi Namikoshi, Alex Filipowicz, David A Shamma, Rumen Iliev, Candice L Hogan, and Nikos Arechiga. Using llms to model the beliefs and preferences of targeted populations. *arXiv preprint arXiv:2403.20252*, 2024.

[27] Thomas E Nelson and Zoe M Oxley. Issue framing effects on belief importance and opinion. *The journal of politics*, 61(4):1040–1067, 1999.

[28] OpenAI. Chatgpt (gpt-4o-mini). Online, 2025. URL `https://openai.com`. Large language model.

[29] Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordóñez, and Sarit Kraus. Playing Games for Security: An Efficient Exact Algorithm for Solving Bayesian Stackelberg Games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 895–902. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[30] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

[31] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

[32] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[33] Jibang Wu, Chenghao Yang, Simon Mahns, Yi Wu, Chaoqi Wang, Hao Zhu, Fei Fang, and Haifeng Xu. Grounded Persuasive Language Generation for Automated Marketing, 2025. URL `https://arxiv.org/abs/2502.16810`.

[34] You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to Persuade on the Fly: Robustness Against Ignorance. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 927–928, Budapest Hungary, July 2021. ACM.

A Omitted Proofs in Section 3

A.1 Proof of Proposition 1

Let $P_C \in \Delta(C)$ be a distribution over the framing space. Consider the sender's expected utility under this distribution (we consider the framing space to be discrete here, but the result immediately holds for the continuous setting by replacing $\sum_c$ with $\int_c$):

$$\mathbb{E}_{P_C}\big[u(a^*_{c,\pi,s},\omega)\big] = \sum_{c\in C} P_C(c) \sum_{\omega\in\Omega}\sum_{s\in\mathcal{S}} \mu_0(\omega)\pi(s|\omega)u(a^*_{c,\pi,s},\omega).$$

Let $c_{\max} = \arg\max_{c\in C}\sum_\omega\sum_s \mu_0(\omega)\pi(s|\omega)u(a^*_{c,\pi,s},\omega)$. Then it is clear that using framing $c_{\max}$ upper-bounds the expected utility achieved from the randomized strategy. Formally:

$$\mathbb{E}_{P_C}\big[u(a^*_{c,\pi,s},\omega)\big] \leq \sum_{\omega\in\Omega}\sum_{s\in\mathcal{S}} \mu_0(\omega)\pi(s|\omega)u(a^*_{c_{\max},\pi,s},\omega) = \mathbb{E}\big[u(a^*_{c_{\max},\pi,s},\omega)\big].$$

So, there is no need to use randomized framing strategies.

A.2 Proof of Theorem 1

We will show that finding the optimal utility in a specific class of Bayesian Stackelberg games (BSG) can be reduced to our problem of computing the optimal sender utility by optimizing only framing/receiver prior (OF). Conitzer and Sandholm [5] prove that the former problem is NP-Hard. Specifically, it is hard to compute the optimal utility for the following class of BSG problems, which they show is sufficient to ensure that the Independent Set problem can be reduced to it.

- Follower has binary actions (a_0, a_1) with positive bounded utility: $u^\theta_f(a_\ell, a_f) \in [0, v^{max}]$, where $v^{max} \leq |\mathcal{A}_\ell|$.
- Follower always has utility 1 for a_0. That is, $\forall\theta, a_\ell$, $u^\theta_f(a_\ell, a_0) = 1$.
- Leader utility is binary and does not depend on the leader's action (only the follower): $u_\ell(a_f) \in \{0, 1\}$.
- The least probable type occurs with non-zero probability: $\min_\theta P(\theta) \triangleq P_{min} \geq 0$.

For any given instance of the BSG with the above characteristics, denoted by $\mathcal{I}_{BS}$, with optimal solution x^* achieving optimal leader utility $\mathrm{BS}(\mathcal{I}_{BS}, x^*)$, we will give a poly-time construction of an OF problem instance $\mathcal{I}'_{OF}$ whose optimal solution μ^*_c is such that $\mathrm{OF}(\mathcal{I}'_{OF}, \mu^*) = \mathrm{BS}(\mathcal{I}_{BS}, x^*)$. Hence if the OF problem can be efficiently solved, it would imply efficient solving of the class of BSG problem described above, which is known to be NP-Hard. For a given instance $\mathcal{I}_{BS} = (\Theta, \mathcal{A}_\ell, \mathcal{A}_f, P(\theta), u_\ell, u_f)$, we first construct an intermediate instance $\mathcal{I}_{OF} = (\Omega, C, \mathcal{A}, S, \mu_0, u, v, \pi)$ as follows:

- The state space for $\mathcal{I}_{OF}$ is: $\Omega = \{\omega_{a_\ell}\}_{a_\ell\in\mathcal{A}_\ell} \cup \{\omega_\theta\}_{\forall\theta\in\Theta} \cup \widetilde{\omega}$
- The receiver's action space is: $\mathcal{A} = \{a^\theta_0\}_{\forall\theta\in\Theta} \cup \{a^\theta_1\}_{\forall\theta\in\Theta} \cup \widetilde{a}_1 \cup \widetilde{a}_2$
- The signal space is: $S = \{s_\theta\}_{\theta\in\Theta}$.

For this instance and $\varepsilon > 0$, we specify the sender's prior μ_0, the fixed signaling scheme π and the sender and receiver utilities u, v as follows:

- Prior μ_0: $\mu_0(\widetilde{\omega}) = 1-\varepsilon$; $\forall\theta$, $\mu_0(\omega_\theta) = \frac{\varepsilon}{|\Theta|}$; $\forall a_\ell$, $\mu_0(\omega_{a_\ell}) = 0$
- Signaling π: $\forall\theta, \pi(s_\theta|\widetilde{\omega}) = P(\theta)$, $\pi(s_\theta|\omega_\theta) = 1$; $\forall\theta\neq\theta'$, $\pi(s_{\theta'}|\omega_\theta) = 0$; $\forall\theta, a_\ell$, $\pi(s_\theta|\omega_{a_\ell}) = \frac{1}{|\Theta|}$
- Sender Utility $u(a,\omega)$:
  - $\forall a_\ell, u(a_*^\theta, \omega_{a_\ell}) = u_\ell(a_*)$, where $* \in \{0,1\}$.
  - $\forall\theta, u(a_*^{\theta'}, \omega_\theta) = -L$ if $\theta'\neq\theta$; otherwise $u(a_*^\theta, \omega_\theta) = 0$.
  - $\forall\omega, u(\widetilde{a}_1, \omega) = -N$; $u(\widetilde{a}_2, \omega) = -K$
  - $u(a_*^\theta, \widetilde{\omega}) = u_\ell(a_*)$, for all θ.
- Receiver Utility $v(a,\omega)$:
  - $\forall a_\ell, v(a_*^\theta, \omega_{a_\ell}) = u_f^\theta(a_*, a_\ell)$; $v(\widetilde{a}_1, \omega_{a_\ell}) = -M-1$; $v(\widetilde{a}_2, \omega_{a_\ell}) = 0$
  - $\forall\theta, v(a_*^{\theta'}, \omega_\theta) = -M$ for $\theta\neq\theta'$; $v(a_*^\theta, \omega_\theta) = 0$; $v(\widetilde{a}_1, \omega_\theta) = -M-1$; $v(\widetilde{a}_2, \omega_\theta) = +K$
  - $v(\widetilde{a}_1, \widetilde{\omega}) = +N$; $v(a\neq\widetilde{a}_1, \widetilde{\omega}) = 0$

The high-level intuition for this instance is as follows. When the receiver sees a signal s_θ (which is proxying type θ in BS), we want them to only consider actions a_0^θ, a_1^θ, which directly corresponds to follower utility of type θ in BS. Receiver utility in O, however, does not explicitly depend on θ, but rather on the state ω. Hence we expand the state space to include ω_θ states. Using a fixed signaling scheme, we want to ensure that μ_c always induces a slight belief in the receiver that states ω_θ occurred. The sender is incentivized to do this since otherwise, the receiver could take $a_*^{\theta'}$ actions at state ω_θ (which occurs with non-zero probability), which is very bad for the sender. They also don't want to put too much weight on ω_θ states, lest the receiver take the bad (for sender) $\widetilde{a}_2$ action. Lastly, we add an additional state $\widetilde{\omega}$ to ensure the OF objective captures the BSG objective, which depends on the type. Formally, the OF optimization problem under this instance construction above can be written as:

$$\underset{\mu_c}{\text{maximize}} \;\; \underbrace{(1-\varepsilon)\sum_\theta P(\theta) u(a^*(\mu_c, s_\theta))}_{\text{state } \widetilde{\omega}} + \underbrace{\frac{\varepsilon}{|\Theta|}\sum_\theta u(a^*(\mu_c, s_\theta), \omega_\theta)}_{\text{for states } \omega_\theta \text{ where } \pi(s_\theta|\omega_\theta)=1} \tag{9}$$

$$\text{s.t} \quad a^*(\mu_c, s_\theta) = \arg\max_{a\in\mathcal{A}} \left[\mu_c(\omega_\theta) v(a, \omega_\theta) + P(\theta)\mu_c(\widetilde{\omega}) v(a, \widetilde{\omega}) + \frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c(\omega_{a_\ell}) v(a, \omega_{a_\ell}) \right] \tag{10}$$

We now prove three intermediate results that will specify the necessary relations between the constants used in our $\mathcal{I}_{OF}$ instance and disentangle the key arguments needed for the reduction.

Lemma 1. *If $\mu_c(\omega_\theta) = \frac{v^{max}}{|\Theta|M}$, $\forall\theta$, $\mu_c(\widetilde{\omega}) = 0$, with $v^{max} \leq \frac{M}{1+K}$, then (1) the receiver always chooses between the two action $\{a_0^\theta, a_1^\theta\}$ on receiving signal s_θ and (2) the sender utility is at least 0.*

Proof. Since $\mu_c(\widetilde{\omega}) = 0$, we need not consider the receiver taking action $\widetilde{a}_1$, since it is dominated by some other action at all remaining states. We first show that on some signal s_θ, they will never take action $\widetilde{a}_2$. Indeed, it is obedient for the receiver to take action a_0^θ as opposed to $\widetilde{a}_2$ on receiving a signal s_θ:

$$\mu_c(\omega_\theta)[v(a_0^\theta, \omega_\theta) - v(\widetilde{a}_2, \omega_\theta)] + \frac{1}{|\Theta|}\sum_{a_\ell} \mu_c(\omega_{a_\ell})[v(a_0^\theta, \omega_{a_\ell}) - v(\widetilde{a}_2, \omega_{a_\ell})] \tag{11}$$

$$= -\mu_c(\omega_\theta)K + \frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c(\omega_{a_\ell}) v(a_0^\theta, \omega_{a_\ell}) = \frac{1}{|\Theta|}\left(1 - \frac{v^{max}}{M}\right) - \frac{K v^{max}}{|\Theta| M} \geq 0 \tag{12}$$

where the second equality in the second line follows since at state ω_{a_ℓ}, the receiver utility matches that of the BSG setting - i.e $v(a_0^\theta, \omega_{a_\ell}) = u_f^\theta(a_0, a_\ell)$ - and in the BSG instances we care about, the

receiver always gets utility 1 by taking action a_0, $v(a_0^\theta, \omega_{a_\ell}) = 1$ for all ω_{a_ℓ}. This is greater than or equal to 0 due to our choice of constants satisfying $v^{max} \leq \frac{M}{1+K}$ (we assume ties break in favour of a^θ actions). Thus the receiver will not take action $\widetilde{a}_2$ on any signal s_θ.
Next, we show that the receiver will not take any "incorrect type" actions $a_*^{\theta'}$ on receiving signal s_θ. Suppose by contradiction they take a deviating action $a_*^{\theta'}$. Then they can expect a utility of at most $\frac{v^{max}}{|\Theta|} - \frac{v^{max}}{|\Theta|} = 0$. But we know they can achieve a utility of at least 1 by playing a_0^θ on each signal s_θ. Thus, under the given specifications of μ_c, the receiver will always play action a_*^θ on signal s_θ. Since the sender's utility on such actions is always at least 0 (mainly due to BSG instance having binary leader utility), the sender achieves at least 0 expected utility under this μ_c. □

Lemma 2. *Let $\varepsilon \in (0,1)$, $L > \frac{|\Theta|}{\varepsilon}$, $v^{max} \leq \frac{M}{1+K}$, and $N, K > \frac{1}{(1-\varepsilon)P_{min}}$. Then for an optimal solution μ_c^*, the receiver only takes actions from $\{a_0^\theta, a_1^\theta\}$ when receiving signal s_θ. This holds even if sender utilities are scaled by a positive constant.*

Proof. We partition the cases where this does not hold into three cases and for each, we indicate the suboptimality of μ_c^* with respect to a feasible solution that does conform to the above.

(1) ∃ *a signal s_θ where the receiver takes action $\widetilde{a}_1$.* If this were to occur, the sender utility is at most (note that the max sender utility in our $\mathcal{I}_{OF}$ instance is 1 and in states ω_θ, the maximum utility is 0):

$$\underbrace{-N(1-\varepsilon)P(\theta)}_{\text{on } \widetilde{\omega} \text{ and signal } \theta} + \underbrace{(1-\varepsilon)}_{\text{on } \widetilde{\omega} \text{ and other signals}} - \underbrace{\frac{N\varepsilon}{|\Theta|}}_{\text{on } \omega_\theta \text{ and } s_\theta} \leq -N(1-\varepsilon)P(\theta) + 1 < \frac{-P(\theta)}{P_{min}} + 1 \leq 0 \quad (13)$$

where the last inequality arises from substituting the lower bound of N specified. The sender thus achives negative utility. However, using claim 1, we know of a feasible specification of μ_c under these parameters where the sender can achieve at least 0 utility. Thus the μ_c^* here cannot be optimal. Note that when we scale by a positive constant, the last part of Eq. (13) simply becomes $c\left[\frac{P(\theta)}{P_{min}} + 1\right] \leq 0$ for the same reason as above.

(2) ∃ *a signal s_θ where the receiver takes action $\widetilde{a}_2$.* As before, if this were to occur, the sender utility for this μ_c^* is at most:

$$-K(1-\varepsilon)P(\theta) + (1-\varepsilon) - \frac{K\varepsilon}{|\Theta|} \leq -K(1-\varepsilon)P(\theta) + 1 \leq \frac{-P(\theta)}{P_{min}} + 1 \leq 0 \quad (14)$$

where in the last inequality, we substitute the lower bound of K specified. As before, the sender archives negative utility, even through claim 1 shows it is possible to achieve a utility of 0, indicating suboptimality. Further, it is impervious to positive scaling of sender utilities.

(3) ∃ *a signal s_θ where the receiver takes an action $a_*^{\theta'}$.* If this were to occur, consider the sender utility:

$$(1-\varepsilon)\underbrace{\sum_{s_\theta} P(\theta) u(a^*(\mu_c, s_\theta))}_{\text{at most } 1} + \frac{\varepsilon}{|\Theta|}\underbrace{u(a_*^{\theta'}, \omega_\theta)}_{-L} + \frac{\varepsilon}{|\Theta|}\sum_{s_{\hat{\theta}}} \underbrace{u(a^*(\mu_c, s_{\hat{\theta}}), \omega_{\hat{\theta}})}_{\text{at most } 0} \quad (15)$$

$$\leq (1-\varepsilon) - \frac{L\varepsilon}{|\Theta|} \leq 1 - \frac{L\varepsilon}{|\Theta|} < 0 \quad (16)$$

where the last inequality follows since $\frac{|\Theta|}{\varepsilon} < L$. Again the sender receives negative utility when it is possible to achieve at least 0 utility due to claim 1. As before, if we were to scale by a positive constant c, inequality (16) simply becomes $c\left[(1-\varepsilon) - \frac{L\varepsilon}{|\Theta|}\right] < 0$ which still becomes negative due to the choice of L. □

Lemma 3. *Let $\varepsilon \in (0,1)$, $L > \frac{|\Theta|}{\varepsilon}$, $v^{max} \leq \frac{M}{1+K}$, and $N, K > \frac{1}{(1-\varepsilon)P_{min}}$. Then for an optimal solution μ_c^*, we can construct a solution μ' in poly-time such that $OF(\mathcal{I}_{OF}, \mu^*) = OF(\mathcal{I}_{OF}, \mu')$, $\mu'(\widetilde{\omega}) = 0$, $a^*(\mu_c', s_\theta) \in \{a_1^\theta, a_0^\theta\}$. This holds even when all sender utilities are scaled by a positive constant.*

Proof. From claim 2, we already know that μ_c^* satisfies $a^*(\mu_c^*, s_\theta) \in \{a_1^\theta, a_0^\theta\}$. We now show that any weight μ_c^* places on $\mu(\widetilde{\omega})$ can be shifted without changing this invariant. For each signal s_θ, let a_θ denote the receiver's optimal action for this signal. Then the receiver's obedience for a_θ implies:

$$-\mu_c(\widetilde{\omega})P(\theta)v(a', \widetilde{\omega}) - \mu_c(\omega_\theta)v(a', \omega_\theta) + \frac{1}{|\Theta|}\sum_{a_\ell} \mu_c(\omega)[v(a^\theta, \omega_{a_\ell}) - v(a', \omega_{a_\ell})] \geq 0 \ \forall a' \tag{17}$$

Now consider a μ_c' where $\mu_c'(\widetilde{\omega}) = 0$ and $\mu_c'(\omega \neq \widetilde{\omega}) = \frac{1}{1-\mu_c^*(\widetilde{\omega})}\mu_c^*(\omega)$. This is clearly a valid distribution since $\sum \mu_c(\omega) = \frac{1}{1-\mu_c^*(\widetilde{\omega})}\sum \mu_c^*(\omega) = 1$. When $a' = \widetilde{a}_1$, since the invariant is originally maintained and $v(\widetilde{a}_1, \widetilde{\omega}) = +N$, the negative first term in Eq. (17) becomes 0 and the last two terms (which together must have been positive) are just increased in scale. Hence the invariant is maintained. For any $a' \neq \widetilde{a}_1$, the first term is 0 in Eq. (17), and the adjusted μ_c' simply scales the remaining two terms which must be non-negative. Hence the invariant is always maintained. In other words, $a^*(\mu_c', s_\theta) = a^*(\mu_c^*, s_\theta) \in \{a_0^\theta, a_1^\theta\}$. Lastly, since the choice of μ_c only affects the sender through the decision taken by the receiver, and both μ_c^* and μ_c' lead the receiver to always behave in the same way, the sender utility is unchanged and the claim holds. □

We now prove BSG can be reduced to OF. For a BSG instance $\mathcal{I}_{BS} = (\Theta, \mathcal{A}_\ell, \mathcal{A}_f, P(\theta), u_\ell, u_f)$, we construct an instance $\mathcal{I}_{OF} = (\Omega, \mathcal{A}, S, \mu_0, \pi, u, v)$ as described earlier, in poly-time. Next, consider an instance $\mathcal{I}_{OF}' = (\Omega, \mathcal{A}, S, \mu_0, \pi, \frac{1}{1-\varepsilon}u, v)$, which is identical to $\mathcal{I}_{OF}$, except all sender utilities are now scaled by $\frac{1}{1-\varepsilon}$. Note that claims (1) and (3) depend purely on the receiver utility and sender utilities for a^θ actions at ω_{a_ℓ} states being non-negative and the statement of (2) highlights that it holds when the sender utilities are scaled by a positive constant. In other words, all three claims hold on instance $\mathcal{I}_{OF}'$. We now show that for any optimal μ_c^* to instance $\mathcal{I}_{OF}'$, there exists a feasible x' that archives the same utility on the corresponding BSG instance. Similarly, for an optimal x^* to $\mathcal{I}_{BS}$, there exists a μ_c' that archives the same utility on the corresponding OF instance. This naturally implies $BS(\mathcal{I}_{BS}, x^*) = \text{OF}(\mathcal{I}_{OF}', \mu_c^*)$.

The "$\Longrightarrow$" direction: Suppose we have an optimal μ_c^* for instance $\mathcal{I}_{OF}'$; without loss of generality, we assume $\mu_c^*(\widetilde{\omega}) = 0$ (if this is not the case, we can use Claim 3 to construct it to be so in poly-time). Since at each s_θ, we are guaranteed that $a^*(\mu_c^*, s_\theta) \in \{a_1^\theta, a_0^\theta\}$, the sender utility is simply $(1-\varepsilon)\sum_\theta P(\theta)u(a^*(\mu_c^*, s_\theta))$, which corresponds to the utility at state $\widetilde{\omega}$ (note that $\mu_0(\widetilde{\omega})$ is not 0). For any s_θ, without loss of generality, let a_1^θ denote the optimal action. Then obedience with respect to a_0^θ (the only other action possible since claim 3 disavows all others) implies:

$$\frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c^*(\omega_{a_\ell})[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] + \mu_c^*(\omega_\theta)\underbrace{[v(a_1^\theta, \omega_\theta) - v(a_0^\theta, \omega_\theta)]}_{0} \geq 0 \tag{18}$$

Let $x' \in \Delta^{|\mathcal{A}_\ell|}$ be as follows: $x(a_\ell) = \frac{1}{\sum_{\omega_{a_\ell}'} \mu_c^*(\omega_{a_\ell}')}\mu_c^*(\omega_{a_\ell})$. Clearly this is a valid strategy since $\sum_{a_\ell} x(a_\ell) = 1$. Further, since this is just scaling of the $\mu_c^*(\omega_{a_\ell})$ we have that:

$$0 \leq \sum_{a_\ell} x(a_\ell)[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] = \sum_{a_\ell} x(a_\ell)[u_f^\theta(a_1, a_\ell) - u_f^\theta(a_0, a_\ell)] \tag{19}$$

This implies that the optimal action for a follower of type θ for strategy x', $a_f^*(\theta, x) = a^*(\mu_c^*, s_\theta)$, which is the optimal action for the OF receiver for the optimal framing μ_c^* and signal s_θ. For $* \in \{0, 1\}$, since the sender utility for a_*^θ actions at the $\widetilde{\omega}$ state in $\mathcal{I}_{OF}$ is the same as the leader's utility for action a^* in BS, and we are using $\mathcal{I}'_{OF}$ where this sender utility is scaled by $\frac{1}{1-\varepsilon}$, we have that:

$$\text{OF}(\mu_c^*) = (1-\varepsilon)\sum_\theta P(\theta)u(a^*(\mu_c^*, s_\theta)) = \sum_\theta P(\theta)u_\ell(a_f^*(x', \theta)) = BS(x') \tag{20}$$

The "$\Longleftarrow$" direction: Suppose we have an optimal solution to the x^* to the BSG instance $\mathcal{I}_{BS}$. Then by definition, the obedience condition holds for any type θ and the follower's optimal action. For an arbitrary type θ, let the optimal receiver action be a_1 without loss of generality. Then:

$$\sum_{a_\ell} x^*(a_\ell)[u_f^\theta(a_1, a_\ell) - u_f^\theta(a_0, a_\ell)] \geq 0 \tag{21}$$

Now consider constructing μ'_c as follows. We first set $\mu'_c(\widetilde{\omega}) = 0$ and $\mu'_c(\omega_\theta) = \frac{v^{max}}{|\Theta|M}$ for all θ. Due to claim 1, we already know that under this strategy, the receiver in the $\mathcal{I}_{OF}$ instance will only choose between $\{a_0^\theta, a_1^\theta\}$ upon receiving a signal s_θ - in other words, we need not concern ourselves with actions $\widetilde{a}_1, \widetilde{a}_2$ or any $a_*^{\theta'}$, since these are dominated. Next, we set $\mu'_c(\omega_{a_\ell}) = \left(1 - \frac{v^{max}}{M}\right) x^*(a_\ell)$. Observe that this is a valid distribution since $\sum_\omega \mu'_c(\omega) = \left(1 - \frac{v^{max}}{M}\right) + \frac{v^{max}}{M} = 1$. We then observe since Eq. (21) holds for $x(a_\ell)$, and μ'_c is simply a rescaling of $x(a_\ell)$ on the ω_{a_ℓ} states, and $u_f^\theta(a_*, a_\ell) = v(a_*^\theta, \omega_{a_\ell})$:

$$\sum_{\omega_{a_\ell}} \mu'_c(\omega_{a_\ell})[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] \geq 0 \tag{22}$$

The expression is indeed sufficient to conclude that a_1^θ is optimal for the OF instance receiver on getting signal s_θ since our construction of μ'_c ruled out all other actions except a_*^θ. Since $u_\ell(a_*^\theta, a_\ell) = u_\ell(a_*^\theta) = u(a_*^\theta, \widetilde{\omega})$ in the $\mathcal{I}_{OF}$ instance, and we are using $\mathcal{I}'_{OF}$ where this sender utility is scaled by $\frac{1}{1-\varepsilon}$, we have that:

$$\begin{aligned}\text{BS}(x^*, \mathcal{I}_{BS}) &= \sum_\theta P(\theta)u_\ell(a_f^*(x, \theta)) = (1-\varepsilon)\sum_\theta P(\theta)\frac{1}{(1-\varepsilon)}u_\ell(a_f^*(x, \theta)) & (23)\\ &= (1-\varepsilon)\sum_\theta P(\theta)u(a^*(\mu'_c, s_\theta)) = \text{OF}(\mu', \mathcal{I}'_{OF}) & (24)\end{aligned}$$

where the last equality follows from the fact that $\mu_0(\omega_{a_\ell}) = 0$ and the receiver is always taking actions of type a_*^θ on signal s_θ, wherein we recall that $\pi(s_\theta|\omega_\theta) = 1$ sender utility $u(a_*^\theta, \omega_\theta) = 0$.

We have thus shown that the specific class of Bayesian Stackelberg games proven by Conitzer and Sandholm [5] to be NP-Hard, can be expressed as an instance of the optimal framing problem, whose optimal solution exactly matches that of the BSG instance. The result of [5] in-fact, implies something stronger. They show that for a graph $G = (V, E)$, it is possible to construct a BSG instance of the type above such that the graph has an independent set of size K if and only if the optimal leader utility in the BSG instance is at least $\frac{|E|}{|E|+1} + \frac{K}{|V|(|E|+1)}$.
Their reduction uses $|E| + |V|$ types with the $P_{min} = \frac{1}{|V|(|E|+1)}$. Since the sender utility is binary, there is no independent set of size K if and only if the optimal leader utility $\leq \frac{|E|}{|E|+1} + \frac{K-1}{|V|(|E|+1)}$. This means that any $\frac{1}{2|V|(|E|+1)}$ additive approximation to the optimal leader utility would allow us to solve the K-Independent set problem, which is NP-Hard. Since they have $|E|+|V|$ and $|V|$ leader actions, we can formally state that it is NP-Hard to compute a $\frac{1}{2|\Theta||\mathcal{A}_\ell|}$ additive approximation to the BSG problem.

This additive approximation factor is predicated when the sender utility includes constant $L > \frac{|\Theta|}{\varepsilon}$ and $N, K \geq \frac{1}{(1-\varepsilon)P_{min}}$ for some $\varepsilon \in (0,1)$. To normalize this for utilities in the range $[0,1]$, we must divide by the range. If N or K dominates, then the range is $\frac{1}{(1-\varepsilon)P_{min}} + 1$ and any approximation constant must be greater than $\frac{1}{2|\Theta||\mathcal{A}_\ell|} \cdot \frac{(1-\varepsilon)P_{min}}{1+(1-\varepsilon)P_{min}} \geq \frac{P_{min}(1-\varepsilon)}{4|\Theta||\mathcal{A}_\ell|}$. Now conversely, if L dominates, then the range is $\frac{|\Theta|}{\varepsilon} + 1$ and thus the approximation constant must be greater than $\frac{\varepsilon}{2|\Theta||\mathcal{A}_\ell|(\varepsilon+|\Theta|)} \geq \frac{\varepsilon}{4|\Theta|^2|\mathcal{A}_\ell|}$. In the optimal framing instance we construct for the reduction, $|\Theta| = |S|$ and $|\Omega| \geq |\mathcal{A}_\ell|$. Thus, it is NP-Hard to approximate the OF problem up to an additive $\min\left(\frac{P_{min}(1-\varepsilon)}{2|S||\Omega|}, \frac{\varepsilon}{4|\Theta|^2|\mathcal{A}_\ell|}\right)$ factor.

A.3 Proof of Proposition 3

Let (u, v) be a pair of utility functions sampled from some continuous distribution. Recall that we consider receiver utility functions v such that for every possible action, there is some belief in $\Delta(\Omega)$ such that this action is strictly optimal (inducible). Consider any pair of actions $a_1, a_2 \in A$. Since a_1 is strictly inducible, there must be some state $\omega_1 \in \Omega$ under which $v(a_1, \omega_1) > v(a_2, \omega_1)$. Since a_2 is strictly inducible, there must be some state $\omega_2 \in \Omega$ under which $v(a_1, \omega_2) < v(a_2, \omega_2)$. This means that, if the receiver's prior μ is deterministically on ω_1, then we have

$$\sum_{\omega \in \Omega} \mu(\omega)\pi(s_0|\omega)\Big(v(a_1, \omega) - v(a_2, \omega)\Big) \ > \ 0$$

since $\pi(s_0|\omega_1) > 0$ by assumption. If the receiver's prior μ is deterministically on ω_2, then we have

$$\sum_{\omega \in \Omega} \mu(\omega)\pi(s_0|\omega)\Big(v(a_1, \omega) - v(a_2, \omega)\Big) \ < \ 0$$

since $\pi(s_0|\omega_2) > 0$ by assumption. Then, by the intermediate value theorem, there must exist a prior belief $\tilde{\mu}$ supported on $\{\omega_1, \omega_2\}$ only, namely, $\tilde{\mu} \in B_{\omega_1,\omega_2} = \{\mu \in \Delta(\Omega) \mid \mu(\omega_1) > 0, \mu(\omega_2) > 0, \forall \omega \notin \{\omega_1, \omega_2\}, \mu(\omega) = 0\}$, and an action $a' \neq a_1$ such that the receiver is indifferent between a_1 and a' upon receiving signal s_0:

$$\begin{aligned} 0 \ &= \ \sum_{\omega \in \Omega} \tilde{\mu}(\omega)\pi(s_0|\omega)\Big(v(a_1, \omega) - v(a', \omega)\Big) \\ &= \ \tilde{\mu}(\omega_1)\pi(s_0|\omega_1)\Big(v(a_1, \omega_1) - v(a', \omega_1)\Big) + \tilde{\mu}(\omega_2)\pi(s_0|\omega_2)\Big(v(a_1, \omega_2) - v(a', \omega_2)\Big) \end{aligned}$$

and moreover a' and a_1 are both weakly better than any other actions:

$$a', a_1 \in \arg\max_{a \in \mathcal{A}} \sum_{\omega \in \Omega} \tilde{\mu}(\omega)\pi(s_0|\omega)v(a, \omega).$$

Note that a' may or may not be equal to a_2. Next, consider the receiver's best-response action $\tilde{a}_s^*$ upon receiving any signal $s \neq s_0$, under signaling scheme π and prior $\tilde{\mu}$:

$$\tilde{a}_s^* \in \arg\max_{a \in \mathcal{A}} \sum_{\omega \in \Omega} \tilde{\mu}(\omega)\pi(s|\omega)v(a, \omega).$$

Because v is randomly sampled from a continuous distribution, and $\tilde{\mu}$ already made the receiver indifferent between a' and a_1 at signal s_0, the probability that $\tilde{\mu}$ will make the receiver indifferent between any two actions under signal s is 0. So, $\tilde{a}_s^*$ must be unique for any $s \neq s_0$, with strict inequality

$$\sum_{\omega\in\Omega} \tilde{\mu}(\omega)\pi(s|\omega)v(\tilde{a}^*,\omega) \ > \ \sum_{\omega\in\Omega} \tilde{\mu}(\omega)\pi(s|\omega)v(a,\omega), \quad \forall a \in \mathcal{A}\setminus\{\tilde{a}_s^*\}.$$

This means that, for sufficiently small $\varepsilon > 0$, the receiver's best-response actions under the following two prior beliefs

$$\tilde{\mu}^{+\varepsilon} = (\tilde{\mu}(\omega_1)+\varepsilon, \tilde{\mu}(\omega_2)-\varepsilon, 0, \ldots, 0), \qquad \tilde{\mu}^{-\varepsilon} = (\tilde{\mu}(\omega_1)-\varepsilon, \tilde{\mu}(\omega_2)+\varepsilon, 0, \ldots, 0)$$

will still be $\tilde{a}_s^*$, given signal $s \neq s_0$.

However, given signal s_0, because the receiver is indifferent between a' and a_1 under prior $\tilde{\mu}$, the receiver will strictly prefer action a_1 under prior $\tilde{\mu}^{+\varepsilon}$ and strictly prefer action a' under prior $\tilde{\mu}^{-\varepsilon}$, for sufficiently small $\varepsilon > 0$. This means that the sender's utilities under priors $\tilde{\mu}^{+\varepsilon}$ and $\tilde{\mu}^{-\varepsilon}$ are

$$\begin{aligned} U_\pi(\tilde{\mu}^{+\varepsilon}) \ &= \ \sum_{\omega\in\Omega}\mu_0(\omega)\Big(\sum_{s\in\mathcal{S}\setminus\{s_0\}} \pi(s|\omega)u(\tilde{a}_s^*,\omega) \ + \ \pi(s_0|\omega)u(a_1,\omega)\Big) \\ U_\pi(\tilde{\mu}^{-\varepsilon}) \ &= \ \sum_{\omega\in\Omega}\mu_0(\omega)\Big(\sum_{s\in\mathcal{S}\setminus\{s_0\}} \pi(s|\omega)u(\tilde{a}_s^*,\omega) \ + \ \pi(s_0|\omega)u(a',\omega)\Big). \end{aligned}$$

We see that

$$U_\pi(\tilde{\mu}^{+\varepsilon}) - U_\pi(\tilde{\mu}^{-\varepsilon}) \ = \ \sum_{\omega\in\Omega}\mu_0(\omega)\pi(s_0|\omega)\Big(u(a_1,\omega)-u(a',\omega)\Big).$$

Because we assumed $\mu_0(\omega) > 0$, $\pi(s_0|\omega) > 0$, $\forall\omega\in\Omega$, and the randomly sampled utility function satisfies $u(a_1,\omega) \neq u(a',\omega)$ with probability 1, we have

$$U_\pi(\tilde{\mu}^{+\varepsilon}) - U_\pi(\tilde{\mu}^{-\varepsilon}) \ = \ C \neq 0$$

for some constant $C \neq 0$ independent of ε. This means that $U_\pi(\mu)$ is not continuous at $\tilde{\mu}$.

B Omitted Proofs in Section 4

B.1 Proof of Observation 1

Proof. Consider an unrestricted signal space S, and for an instance $\mathcal{I}$, let (c^*, π^*) denote the optimal strategy, with μ_c^* denoting the framing-induced belief. For this strategy, let $m : A \to S$ denote the correspondence between actions to signals under (μ_c^*, π^*). Then the sender utility is:

$$\sum_\omega \sum_a u(a,\omega) \sum_{s\in m(a)} \pi^*(s|\omega) \tag{25}$$

Consider a scheme $\pi'(a|\omega) = \sum_{s\in m(a)} \pi(s|\omega)$. We note that the receiver takes action a when the receiver observes signal a under this scheme since:

$$\begin{aligned} \forall s \in m(a), \forall a' \ &: \sum_\omega \mu_c^*(\omega)\pi^*(s|\omega)[v(a,\omega)-v(a',\omega)] \geq 0 \\ \implies \forall a' &: \sum_\omega \mu_c^*(\omega)[v(a,\omega)-v(a',\omega)] \sum_{s\in m(a)} \pi^*(s|\omega) \geq 0 \end{aligned}$$

It is thus clear that the sender utility from Eq. (25) in unchanged by using this direct scheme with signal space S equal A as action recommendations. □

B.2 Proof of Theorem 2

Without loss of generality, assume that the utility functions of the sender and the receiver are bounded: $\forall a \in \mathcal{A}, \forall \omega \in \Omega$, $u(a,\omega) \in [0,1], v(a,\omega) \in [0,1]$. Recall that $U^*(\mu)$ is the solution to the linear program outlined in (6). We aim to show that $U^*(\mu)$ is continuous at any $\mu \in \Delta(\Omega)$ satisfying $\mu(\omega) > 0, \forall \omega \in \Omega$. We break this result into a set of intermediate claims.

Lemma 4 (Continuity of posterior)**.** *Let $\pi : \Omega \to \Delta(\mathcal{S})$ be any signaling scheme. Let $\mu, \mu' \in \Delta(\Omega)$ be two receiver beliefs. Let μ_s, μ'_s be the posterior beliefs induced by signal s under π and priors μ, μ' respectively. Suppose* $\min_{\omega\in\Omega}\mu(\omega) \geq p_0 > 0$*. Then,* $\|\mu_s - \mu'_s\|_1 \leq \frac{2}{p_0}\|\mu - \mu'\|_1$.

Proof of Lemma 4. Let $\pi(s) = \sum_{\omega\in\Omega}\mu(\omega)\pi(s|\omega)$ and $\pi'(s) = \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)$ be the probability of signal s under prior μ and μ' respectively. By the definition of μ_s, μ'_s and by triangle inequality,

$$\begin{aligned}\|\mu_s - \mu'_s\|_1 &= \sum_{\omega\in\Omega}\big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\big| \\ &\leq \sum_{\omega\in\Omega}\big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| + \sum_{\omega\in\Omega}\big|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\big|.\end{aligned}$$

For the first term above,

$$\sum_{\omega\in\Omega}\big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| = \sum_{\omega\in\Omega}\tfrac{\pi(s|\omega)}{\pi(s)}|\mu(\omega) - \mu'(\omega)|.$$

We note that, $\forall \omega \in \Omega$,

$$\tfrac{\pi(s|\omega)}{\pi(s)} = \tfrac{\pi(s|\omega)}{\sum_{\omega'\in\Omega}\mu(\omega')\pi(s|\omega')} \leq \tfrac{\pi(s|\omega)}{p_0\sum_{\omega'\in\Omega}\pi(s|\omega')} \leq \tfrac{1}{p_0}. \tag{26}$$

$$\implies \sum_{\omega\in\Omega}\big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| \leq \sum_{\omega\in\Omega}\tfrac{1}{p_0}|\mu(\omega) - \mu'(\omega)| = \tfrac{1}{p_0}\|\mu - \mu'\|_1. \tag{27}$$

For the second term,

$$\begin{aligned}\sum_{\omega\in\Omega}\big|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\big| &= \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\big|\tfrac{\pi'(s)-\pi(s)}{\pi(s)\pi'(s)}\big| \\ &= \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\big|\tfrac{\sum_{\omega'\in\Omega}(\mu'(\omega')-\mu(\omega'))\pi(s|\omega')}{\pi(s)\pi'(s)}\big| \\ &\leq \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\tfrac{\sum_{\omega'\in\Omega}|\mu'(\omega')-\mu(\omega')|\cdot\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)\pi'(s)} \\ &= \|\mu' - \mu\|_1\sum_{\omega\in\Omega}\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\tfrac{\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)} \\ \text{by (26)} \quad &\leq \|\mu' - \mu\|_1\sum_{\omega\in\Omega}\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\tfrac{1}{p_0} = \tfrac{1}{p_0}\|\mu' - \mu\|_1.\end{aligned}$$

Therefore, we obtain $\|\mu_s - \mu'_s\|_1 \leq \frac{2}{p_0}\|\mu' - \mu\|_1$. □

Recall that in the model (Section 2) we assumed "every action $a \in \mathcal{A}$ is strictly inducible" in the receiver. This means that there exists a constant $D > 0$ such that, for every action $a \in \mathcal{A}$, there exists a belief $\eta_a \in \Delta(\Omega)$ for which $\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \geq D > 0$ for every $a' \neq a$.

We now want to show the following: *Suppose the prior $\mu \in \Delta(\Omega)$ satisfies $\mu(\omega) \geq 2p_0 > 0, \forall \omega \in \Omega$. Then, for any prior μ' satisfying $\|\mu' - \mu\|_1 \leq \varepsilon < \min\{p_0, \frac{p_0^2 D}{2}\}$, we have:*

$$\left|U^*(\mu') - U^*(\mu)\right| \leq \frac{4\varepsilon}{p_0^2 D}.$$

This will directly prove the theorem.

Let π^* be the optimal signaling scheme for μ, namely, a solution to the linear program in the definition of $U^*(\mu)$. Let $\pi^*(a)$ be the unconditional probability that π^* sends signal a under prior μ: $\pi^*(a) = \sum_{\omega\in\Omega} \mu(a)\pi^*(a|\omega)$. Let $\mu_a \in \Delta(\Omega)$ be the posterior belief induced by signal a under prior μ:

$$\mu_a(\omega) = \frac{\mu(\omega)\pi^*(a|\omega)}{\pi^*(a)}, \quad \forall \omega \in \Omega.$$

Since π^* is persuasive (the constraint in the linear program), a must be an optimal action for the receiver on posterior μ_a:

$$\mathbb{E}_{\omega\sim\mu_a}[v(a,\omega) - v(a',\omega)] \geq 0, \ \forall a' \neq a.$$

According to inducibility assumption, there exists a belief $\eta_a \in \Delta(\Omega)$ for which $\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \geq D > 0$ for every $a' \neq a$. Consider the convex combination of μ_a and η_a with coefficients $1-\delta, \delta$ (we will choose δ in the end): $\xi_a = (1-\delta)\mu_a + \delta\eta_a$. By the linearity of expectation, a must be better than any other action a' by δD on belief ξ_a:

$$\mathbb{E}_{\xi_a}[v(a,\omega) - v(a',\omega)] = (1-\delta)\mathbb{E}_{\hat{\mu}_a}[v(a,\omega) - v(a',\omega)] + \delta\mathbb{E}_{\eta_a}[v(a,\omega) - v(a',\omega)] \geq \delta D. \tag{28}$$

Let $\xi = \sum_{a\in A} \pi^*(a)\xi_a \in \Delta(\Omega)$, and write μ as the convex combination of ξ and another belief $\chi \in \Delta(\Omega)$:

$$\mu \ = \ (1-y)\xi + y\chi \ = \ \sum_{a\in A}(1-y)\pi^*(a)\xi_a \ + \ y\chi. \tag{29}$$

Lemma 5 (Proposition 1 of Zu et al. [34])**.** *If $\delta \leq p_0$, then there exist χ on the boundary of $\Delta(\Omega)$ and $0 \leq y \leq \frac{\delta}{p_0} \leq 1$ that satisfy* (29)*.*

Since (29) is a convex decomposition of the prior μ, according to [19], there exists a signaling scheme $\tilde{\pi}$ that induces posterior ξ_a with probability $(1-y)\pi^*(a)$, for $a \in \mathcal{A}$, and the posterior that puts all probability on ω with probability $y\chi(\omega)$, for $\omega \in \Omega$. Namely, $\tilde{\pi}$ has signal space $\mathcal{S} = \mathcal{A} \cup \Omega$ and signal probability

$$\tilde{\pi}(s|\omega) = \begin{cases} \frac{(1-y)\pi^*(a)\xi_a(\omega)}{\mu(\omega)} & \text{for } s = a \in \mathcal{A}; \\ \frac{y\chi(\omega)}{\mu(\omega)} & \text{for } s = \omega \in \Omega; \\ 0 & \text{otherwise.} \end{cases}$$

It is not hard to verify that, under prior μ and signaling scheme $\tilde{\pi}$, the posterior induced by signal $a \in \mathcal{A}$ is equal to ξ_a, and the posterior induced by signal ω is the deterministic distribution on ω.

We show that, whenever $\tilde{\pi}$ sends an action recommendation $a \in \mathcal{A}$, the recommendation is persuasive for the receiver under any prior μ' in $B_1(\mu, \varepsilon) = \{\mu' : \|\mu' - \mu\|_1 \leq \varepsilon\}$.

Claim 1. *Suppose $\delta \geq \frac{2\varepsilon}{p_0 D}$. Then, for any prior $\mu' \in B_1(\hat{\mu}, \varepsilon)$, any action recommendation $a \in \mathcal{A}$ from $\tilde{\pi}$ is persuasive.*

Proof. By continuity of posterior (Lemma 4), the posteriors induced by signal a under prior μ and μ' satisfy

$$\|\mu_a - \mu'_a\|_1 \le \tfrac{2}{p_0}\|\mu - \mu'\|_1 \le \tfrac{2\varepsilon}{p_0}.$$

Note that the posterior $\mu_a = \xi_a$, so $\|\xi_a - \mu'_a\|_1 \le \frac{2\varepsilon}{p_0}$. Then, since the receiver's utility is in $[0,1]$, for any action $a' \neq a$,

$$\big|\mathbb{E}_{\omega\sim\mu'_a}[v(a,\omega) - v(a',\omega)] - \mathbb{E}_{\omega\sim\xi_a}[v(a,\omega) - v(a',\omega)]\big| \le \|\mu_a - \xi_a\|_1 \le \tfrac{2\varepsilon}{p_0}.$$

Together with (28), we get

$$\mathbb{E}_{\omega\sim\mu_a}[v(a,\omega) - v(a',\omega)] \ge \delta D - \tfrac{2\varepsilon}{p_0} \ge 0.$$

Thus, the action recommendation a is persuasive. □

Then, we show that the signaling scheme $\tilde{\pi}$ is "close to" π^* in the following sense:

Claim 2. *For any $a \in \mathcal{A}$ and $\omega \in \Omega$, $|\tilde{\pi}(a|\omega) - \pi^*(a|\omega)| \le \frac{\delta}{p_0} + y$.*

Proof. By definition,

$$\begin{aligned}
|\tilde{\pi}(a|\omega) - \pi^*(a|\omega)| &= \Big|\frac{(1-y)\pi^*(a)\xi_a(\omega)}{\mu(\omega)} - \frac{\pi^*(a)\mu_a(\omega)}{\mu(\omega)}\Big| \\
&\le (1-y)\Big|\frac{\pi^*(a)\xi_a(\omega)}{\mu(\omega)} - \frac{\pi^*(a)\mu_a(\omega)}{\mu(\omega)}\Big| + y\cdot\frac{\pi^*(a)\mu_a(\omega)}{\mu(\omega)} \\
&= (1-y)\frac{\pi^*(a)}{\mu(\omega)}\big|\xi_a(\omega) - \mu_a(\omega)\big| + y\cdot\pi^*(a|\omega) \\
&= (1-y)\frac{\pi^*(a)}{\mu(\omega)}\cdot\delta\big|\eta_a(\omega) - \mu_a(\omega)\big| + y\cdot\pi^*(a|\omega) \\
&\le (1-y)\frac{1}{p_0}\cdot\delta\cdot 1 + y\cdot 1 \le \frac{\delta}{p_0} + y.
\end{aligned}$$

□

Let $U(\mu, \tilde{\pi})$ be the sender's expected utility when using signaling scheme $\tilde{\pi}$. Since the action recommendation from $\tilde{\pi}$ are persuasive under prior μ (Claim 1), the receiver takes a when receiving signal a. When receiving any signal ω, the receiver takes some action $a^*_\omega \in \arg\max_{a\in\mathcal{A}} v(a,\omega)$. So,

$$U(\mu,\tilde{\pi}) \;=\; \sum_{\omega\in\Omega}\mu_0(\omega)\Big(\sum_{a\in\mathcal{A}}\tilde{\pi}(a|\omega)u(a,\omega) \;+\; \tilde{\pi}(\omega|\omega)u(a^*_\omega,\omega)\Big).$$

Because we assumed $u(a,\omega) \ge 0$,

$$U(\mu,\tilde{\pi}) \;\ge\; \sum_{\omega\in\Omega}\mu_0(\omega)\sum_{a\in\mathcal{A}}\tilde{\pi}(a|\omega)u(a,\omega) \;=:\; U_{\mathcal{A}}(\tilde{\pi}).$$

where $U_{\mathcal{A}}(\tilde{\pi})$ denotes the expected utility from action recommendation signals, which is also the objective function of the linear program in the definition in $U^*(\mu)$. Note that $U_{\mathcal{A}}(\pi^*) = U^*(\mu)$. We claim that $U_{\mathcal{A}}(\tilde{\pi})$ cannot be too much worse than $U_{\mathcal{A}}(\pi^*)$:

Claim 3. *Given $\delta \ge \frac{2\varepsilon}{p_0 D}$, we have $U_{\mathcal{A}}(\tilde{\pi}) \ge U_{\mathcal{A}}(\pi^*) - \frac{2\delta}{p_0}$.*

Proof. By definition,

$$\begin{aligned} U_{\mathcal{A}}(\tilde{\pi}) &= \sum_{\omega\in\Omega}\mu_0(\omega)\sum_{a\in\mathcal{A}}\tilde{\pi}(a|\omega)u(a,\omega) \\ \text{(by Claim 2)} &\geq \sum_{\omega\in\Omega}\mu_0(\omega)\sum_{a\in\mathcal{A}}\pi^*(a|\omega)u(a,\omega) - \Big(\frac{\delta}{p_0}+y\Big)\underbrace{\sum_{\omega\in\Omega}\mu_0(\omega)\sum_{a\in\mathcal{A}}u(a,\omega)}_{\leq 1} \\ &\geq U_{\mathcal{A}}(\pi^*) - y - \frac{\delta}{p_0} \ \geq \ U(\hat{\pi},\mu_0,\hat{\mu}) - \frac{2\delta}{p_0} \end{aligned}$$

where in the last line we used $y \leq \frac{\delta}{p_0}$ from Lemma 5. □

Because $\tilde{\pi}$ is persuasive for any prior $\mu' \in B_1(\mu,\varepsilon)$ and $U_{\mathcal{A}}(\tilde{\pi}) \geq U_{\mathcal{A}}(\pi^*) - \frac{2\delta}{p_0}$, we have:

$$\begin{aligned} U^*(\mu') \geq \ U(\mu',\tilde{\pi}) \geq \ & U_{\mathcal{A}}(\tilde{\pi}) \\ \geq \ & U_{\mathcal{A}}(\pi^*) - \frac{2\delta}{p_0} \\ = \ & U^*(\mu) - \frac{2\delta}{p_0} \geq \ U^*(\mu) - \frac{4\varepsilon}{p_0^2 D} \end{aligned}$$

where we let $\delta = \frac{2\varepsilon}{p_0 D}$. By a symmetric argument, we also have $U^*(\mu) \geq \ U^*(\mu') - \frac{4\varepsilon}{p_0^2 D}$, which implies $|U^*(\mu') - U^*(\mu)| \leq \frac{4\varepsilon}{p_0^2 D}$.

C Experimental Setup: Real Estate

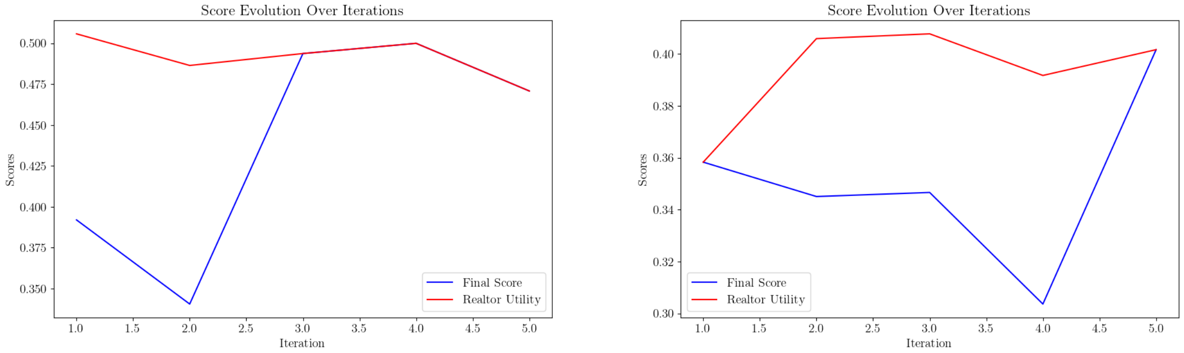
C.1 Instance parameters

Here we include the detailed setup of the real-estate example we used to verify our framework experimentally. Note that the utilities here are not within the range $[0,1]$, but can be normalized to be so without loss of generality. Indeed, the utility values highlighted in Section 5 are normalized utilities according to the $[0,1]$ scale. There are 4 possible states for each instance: $(\textit{good},\textit{cheap})$, $(\textit{good},\textit{expensive})$, $(\textit{bad},\textit{cheap})$, $(\textit{bad},\textit{expensive})$. Each buyer, however, has a different notion of "good" and "cheap" (see the buyer profiles in Section 5.2). We index these states 0 through 3. The instances share the same utilities but have different realtor priors. Rows correspond to "not buy" and "buy".

- Realtor prior for Henry: $[0.1, 0.35, 0.3, 0.25]$
- Realtor prior for Lilly: $[0.2, 0.4, 0.1, 0.3]$
- Realtor Utility: $\begin{bmatrix} 0 & 0 & 0 & 0 \\ -0.25 & 1 & -0.5 & 0.75 \end{bmatrix}$
- Buyer Utility (both Henry and Lilly): $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0.75 & -0.25 & 0.25 & -3 \end{bmatrix}$

C.2 Intermediate Results

We observe that the iterative generation process finds a good framing within 4 to 5 iterations; thereafter, the results it produces becomes poor. The process also tends to find high utility solutions fairly easily, but they often score low on correctness at first. The "final score" in the plots is the sender utility multiplied by the correctness score.



(a) Score over iterations on the Henry Instance (b) Scores over iterations on the Lilly Instance

Figure 4: Scores over multiple iterations of the LLM generating framing. Final score is the product of the utility and the correctness score.

- Henry Instance, Iteration 2: *Meet Jeremy Hammond, a dedicated realtor with 8 years of experience in the Boston area and a strong foundation as a contractor. His expertise in home maintenance and repair ensures you'll find properties that require minimal effort, allowing you to focus on enjoying the great outdoors. As an active community member and nature enthusiast, Jeremy understands the importance of access to parks and trails. He is committed to guiding clients toward quality homes that align with an active lifestyle, making him an excellent choice for discovering properties that fit your needs and budget.*

- Lilly Instance, Iteration 4: *Jeremy Hammond is a knowledgeable realtor in Boston with over 8 years of experience, including 6 years as a realtor and a prior background as a contractor. His hands-on approach allows him to identify quality homes that cater to families. As a father himself, Jeremy appreciates the importance of finding spacious, welcoming neighborhoods. He is actively involved in the community, giving him insights into local schools and amenities. Trust Jeremy to help you navigate the suburban landscape, ensuring you find a home that combines comfort, community, and family-friendly features.*

To give an example of some intermediate results, consider the framing generated at iteration 2 of the Henry instance. It scores well on utility but not on correctness. The justification given by the LLM is it may be leap to consider Jeremy having expertise in home maintenance and repair, despite being a contractor, since no further information was specified. Further it is is not clear whether his 8 years of experience as a realtor were all in Boston (the last 2 were). In iteration 4 of the Lilly instance, the LLM notes that Jeremy has a combined 8 years of experience in real-estate, not the 6 specified in the framing.

C.3 Prompts to Estimating Beliefs from Framing

To estimate the belief for a given framing, we use the following prompt template. The key-words `buyer_name` and `buyer_desc` and `realtor_desc` correspond to the instance parameters mentioned in Section 5.2:

You will be used as a proxy for a (human) person looking to buy a house. You will be given a description of the potential buyer (their preferences, etc) and a description of a real estate agent soliciting clients. You will be asked to provide your responses in a JSON format specified in the prompt.

GENERAL PROBLEM DESCRIPTION: Both the client and realtor are based in Boston. You can imagine a house has the following features: (good, cheap), (good, expensive), (bad, cheap), (bad, expensive). Please see below for what constitutes "good" and "bad" for this buyer that you are acting as a proxy for.

BUYER DESC: `buyer_desc`

REALTOR DESC: `realtor_desc`

TASK DESC: Given your general knowledge about the Boston housing market, it's general pricing/cost-of-living and most importantly, this description of the realtor, what are the probabilities (across the 4 categories as defined by `buyer_name` *preferences) for houses this realtor might be familiar with/used to showing. Explain your reasoning but please give a precise probability vector (of size 4) for the 4 states a house listed/shown/specialized in by this realtor can have. To give context, we wish to determine what this realtor is usually showing/familiar with see if that matches* `buyer_name` *preferences. Please pay attention to the tangible aspects of this realtor's description and background (ignore fluff like excellent customer service) and how they relate to* `buyer_name`*. Lastly, recall that a probability vector must sum to 1. Provide your response in the following JSON format:*

```
{
    "probabilities": {
        "good_cheap": float,
        "good_expensive": float,
        "bad_cheap": float,
        "bad_expensive": float
    },
    "reasoning": string
}
```

C.4 Prompts to Search over the Framing Space

To search over the framing space, we use the following prompt template. The key-words `buyer_name` and `buyer_desc` and `realtor_desc` correspond to the instance parameters mentioned in Section 5.2. Any generated framing and corresponding feedback is appended to this prompt for the next iteration:

You will be asked to generate a short description/bio of a realtor (in json format) to make them appeal to a specific buyer.For each description you generate, quantitative feedback will be provided on the generated, which you will use to improve what you generate.

TASK DESC: You will be given a REALTOR_PROFILE that outlines features and attributes of a realtor. You will be given BUYER_DESC that outlines properties of a house buyer we wish to target. Your task is to generate at most 100 words REALTOR_DESC string that will be shown to this buyer. Given this profile you generate, the buyers perception of the type of houses the realtor can show them will be measured (quantitatively). Please see BUYER_DESC fow how we partition possible houses into 4 states - it is the buyer's belief over these states that we measure. Using this perceived prior, we will signal the buyer (think Bayesian Persuasion) to influence their actions (which are buy or not buy).We will compute all of this and give you numerical feedback (see FEEDBACK_DESC). Please use this feedback to improve the REALTOR_DESC you generate. Note that the realtor profile you generate directly influences how the buyer perceives this realtor and their possible expertise and offerings which is captured in the perceived prior. This will directly influence the utility we can derive after signaling. To give a simple example, if you can generate a *prior such that the buyer with high probability thinks that the houses this realtor can show/offer is good and cheap, they may be more inclined to buying.*

FEEDBACK_DESC: Whenever you generate a REALTOR_DESC, you will be given feedback as follows: correctness_score: [0,1] specifies whether you wrote something blatantly incorrect given realtor_desc.
prior_generated: The generated prior and the reasoning given by the buyer for this belief.
realtor_utility: The expected utility the realtor could get with the given prior.
*final_score: realtor_utility*correctness_score. This is what we are trying to maximize.*

REALTOR_PROFILE: `realtor_desc`

BUYER DESCRIPTION: `buyer_desc`

Please generate a REALTOR_DESC in json form with key REALTOR_DESC and we will give you feedback. Avoid just re-stating the preferences of the buyer.

C.5 Prompts to Score Soundness

We use the following prompt to score for soundness, which we refer to correctness in the prompt:

You will be given a profile of a realtor (labelled REALTOR_PROFILE) agent which includes various information about them. You will also be givea short natural language bio/description (labelled REALTOR_DESC) about them that is meant to be shown to a prospective buyer who may wish to work with them.

Your goal is to score the REALTOR_DESC (give a number between 0 and 1 with 0 being bad and 1 being good) on correctness.

Correctness refers to whether the REALTOR_DESC mentions something that is clearly in contradiction/factually incorrect given the information in the profile. For example, the REALTOR_DESC mentioning the realtor has 2 kids, when the REALTOR_PROFILE explicitly states that he has no children. For blatant incorrectness like this, give 0. For this same example, however, if the REALTOR_DESC mentions the realtor has 2 kids and the REALTOR_PROFILE did not explicitly mention anything about kids, then it DOES NOT violate correctness (and should have score 1). I.e. not mentioning information does not violate correctness. Note that platitutdes about their skills or abilities or general flowery descriptions also do not violate correctness. But making leaps about their work/professional capabilities can be a violation. If it is something plausible about their expertise but not directly in the profile give it between 0.4 and 0.6 score. If there is placeholder text or any text that is not presentable to the buyer, give it 0. For given instance, return a correctness_score. Please see some example scoring below. This is an example, not the real instance.

REALTOR_PROFILE: Richard Clarkson is Male, 42 years old, Worked with the our firm for 2 years, Worked previously as a realtor for 6 years, and a contractor before that. Lives with his wife and 3 kids and a dog and a cat in Downtown Boston. Hobbies include playing the drums, spending time with kids, hiking, and backyard gardening Active member of his Home Owner's Association.

REALTOR_DESC (1): Richard is dedicated and highly experienced real estate agent specializing in the Denver area. Proven success in navigating complex negotiations and market trends to provide exceptional client experiences. Known for personalized attention and exceeding client. - correctness_score: 0 (Since it mentiones Jeremy as working in Denver when in reality they are in Boston)

REALTOR_DESC (2): Richard is an seasoned realtor with 8 years of experience in real-estate. He loves to spend time in the great outdoors and is an avid hiker. - correctness_score: 1 (2 years with this company and 6 with an earlier one is 8 years)

REALTOR_DESC (3): If you want a spacious house look no further than Richard, he lives in big mansion with his wife and kids. - correctness_score: 0.2 (Makes a somewhat unplausible leap *that Richard lives in a mansion when the profile does not say anything of that sort)*

REALTOR_DESC (4): Richard is dedicated and highly experienced real estate agent specializing in the Boston area. He can navigate complex settings and work to ensure his clients get the best deal possible. You will get attention to detail, perseverance and exception skill with Richard. - correctness_score: 1 (Does no mention anything factually incorrect)

REALTOR_DESC (5): Richard is a Boston realtor. The realtor James enjoys biking and finds houses close to nature. - correctness_score: 0 (Statement about Richard is correct. But mentions another realtor James, which is not part of the REALTOR_DESC)

REALTOR_DESC (6): Richard is a Boston realtor. He specializes in [SPECIALIZATIONS]. - correctness_score: 0 (Statement about Richard is correct. But includes placeholder text or text that is not proper to show a buyer)

REALTOR_DESC (7): Richard the realtor focuses on commercial and lakeside properties in the Boston and offers relocation services. - correctness_score: 0.2 (While nothing that is an explicit contradiction, it does make many suppositions which may not be accurate.)

REALTOR_DESC (7): Richard sepcializes in fixer-uppers that are below market price. - correctness_score: 0.3 (The realtor profile does not mention anything like specific like this)

For the given instance of REALTOR_PROFILE and REALTOR_DESC please explain your reasoning first before scoring REALTOR_DESC on the correctness_score. Return a JSON object with the key "reasoning", which is a natural language description of why you chose the score. Then use keys "correctness_score" whose value is a number between 0 and 1 to give your score.

REALTOR_PROFILE: `realtor_desc`

REALTOR_DESC: `framing`

D Experimental Setup: Advertising

D.1 Prompts to Estimating Beliefs from Framing

To estimate the belief for a given framing, we use the following prompt template. The strings `Target Demographic` and `Brand Description` are mentioned in the main paper body.

You will be used as a proxy for a target demographic to assess shopping inclinations for market research. You will be given a description of the demographic (their preferences, etc) and the motto and description of a clothing line. You will be asked to provide your responses in a JSON format specified in the prompt

GENERAL PROBLEM DESCRIPTION: You are taking the role of someone in the given demographic. You can imagine they categorize clothing into the following categories: (trendy, more durable), (trendy, less durable), (not trendy, more durable), (not trendy, less durable). Please see BUYER DESC for what this buyer values.

BUYER DESC: `Target Demographic`

BRAND DESC: `Brand Description`

TASK: Your role is to act as a member of the described demographic and evaluate how you would interpret the products from a new outerwear line based solely on the brand's motto and product description (See BRAND DESC). Your goal is to determine how you (as an average style-aware mall-shopper) would categorize the product line into the following four quadrants: (Trendy, More Durable), (Trendy, Less Durable), (Not Trendy, More Durable) and (Not Trendy, Less Durable) Please return what probabilities (recall they sum to 1) this average users from this demographic would assign to each of these categories for products from this line. Note that this is not about what the demographic cares about or prioritizes in purchases. Instead, focus on how they would interpret the messaging — what assumptions they would make about the clothing's fashionability and *durability from the language, tone, and emphasis in the brand's description and motto. IMPORTANT: DO NOT MAKE FAR REACHING ASSUMPTIONS OR TRY TO BE UNJUSTIFIABILY OPTIMISTIC. Provide your response in the following JSON format:*

```
{
    "reasoning": string,
    "probabilities": {
        "trendy_more_durable": float,
        "trendy_less_durable": float,
        "not_trendy_more_durable": float,
        "not_trendy_less_durable": float
    },
}
```

D.2 Prompts to Search over the Framing Space

To search over the framing space, we use the following prompt template. The key-words `buyer_name` and `buyer_desc` and `realtor_desc` correspond to the instance parameters mentioned in Section 5.2. Any generated framing and corresponding feedback is appended to this prompt for the next iteration:

You will be asked to generate a brand motto and description for one of its product lines. For each motto, description you generate, quantitative feedback will be provided on the generated, which you will use to improve what you generate.

TASK DESC: You will be given a BRAND DESC that describes the clothing brand 'Himalaya' and a new product line they are trying to launch. You will be given DEMOGRAPHIC DESC that outlines the features of the demographic they are targeting for this new product line. Your task is to generate a BRAND MOTTO (atmost 15 words) and PRODUCT LINE DESC of their new product line (at most 100 words). The motto and description will be shown to members in the target demographic. Their perception of how products from this new line fit into the 4 possible categories this demographic cares about will be measures (quantitatively). Please see BUYER DESC fow how they partition clothes into 4 possible states - it is their belief over these states that we measure. Feel free to USE OR NOT USE any information in the provided BRAND DESC to sway the target demographic. Not revealing information can sometimes be helpful. Using this perceived prior, we will signal the buyer (using Bayesian Persuasion) to influence their actions (which buy-on-sale, buy-regular-price, not buy). We will compute all of this and give you the numerical utility the company achieves when using your generated motto and description. See FEEDBACK DESC on how the feedback will be structured. Please use this feedback to improve the BRAND MOTTO and PRODUCT DESC you generate. Note that your generated motto and description directly influences how the buyer perceives this Himalaya's new product line. This is captured in their prior, which will directly influence the utility we can derive after signaling. Lastly, feel free to navigate this space to see what works and what doesn't.

FEEDBACK DESC: Whenever you generate a candidate, you will be given feedback as follows: prior generated: The generated prior and the reasoning given by the target demographic for this belief. brand utility: The expected utility the brand could get with the given prior and using Bayesian Persuasion signaling. This is what we are trying to maximize.

Please generate a BRAND MOTTO and PRODUCT LINE DESC in json form with those as keys. Avoid just re-stating the preferences of the buyer.