Framing and Signaling
An LLM-Based Approach to Information Design

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 framing—how information is contextually presented—is also a critical factor in decision-making. This paper formalizes a language-based notion of framing and bridges this to a popular model of Bayesian signaling: Persuasion. We model framing as a possibly non-Bayesian, linguistic way to influence a receiver's prior belief, while signaling further updates this belief in the classical Bayesian way. We analyze the computational complexity of jointly optimizing framing and signaling, as well as optimizing framing under a fixed signaling scheme. Our theoretical results also outline the settings where even minimal framing effects can significantly enhance a sender's utility, as compared to signaling alone. A key challenge in this optimization problem is the vast space of possible framings and the difficulty of predicting their effects on receivers. We explore the potential of Large Language Models (LLMs) to address these challenges and empirically validate an LLM-augmented optimization framework for framing and signaling. By formally integrating framing with signaling into a comprehensive model, we allow a systematic way to harness insights from Psychology and AI-augmented decision-making to Information Design.

1 Introduction

Information design is a core concept in microeconomics and decision theory, and considers the strategic communication of information from one party (i.e., sender) to shape the decisions of others (i.e., receivers) [1]. While it has been extensively studied through different lenses, the *persuasion* model, pioneered by Kamenica and Gentzkow [19], considers the decision-maker's belief influenced through a Bayesian approach and involves carefully engineering a signal that is correlated with an observable world state. These quantitative signal-to-state correlations are all that matter in such models. Crucially, they are agnostic to *how* signals or related metadata about the instance are presented. Under this strict Bayesian persuasion framework, a Nike advertisement credibly conveying the quality of a shoe (the signal) would be equally effective whether accompanied by the

*Correspondence to shossain@g.harvard.edu

slogan "Wear Nike" or "Just Do It." Similarly, a travel website describing a discounted trip as a "blissful vacation" would, in theory, yield the same decision-making outcome as labeling it "time off."

This perspective stands in strong contrast to established findings in behavioral economics and psychology, which emphasize that *framing* — the linguistic, visual, and contextual presentation of information — can significantly shape perception and choice, even when the underlying information remains unchanged [29]. Framing can take several forms when considering it in conjunction with signaling: (i) tangential information given before the signal; (ii) text or phrasing used to represent or describe the signal; (iii) non-textual cues such as color, background images, fonts, etc, used when transmitting the signal. Tversky and Kahneman [29] argue that such information combines with societal norms to affect the perception of "acts, outcomes, and contingencies" in the decision-maker. They compare it with how the same visual scene changes with the choice of vantage point. Mullainathan et al. [23] define this phenomenon formally as a process called "Coarse Thinking" (as opposed to "Bayesian Thinking"). Bordalo et al. [2], on the other hand, give a Bayesian perspective of framing. These models are all supported by ample evidence that suggests framing plays a crucial role in practically persuading decision-makers [9].

This paper formally combines the classic Bayesian signaling model popularized in the persuasion literature with the aforementioned framing perspective, which has been extensively studied in behavioral economics and psychology [29, 2, 13, 11]. Aligning ourselves with the perspective that framing influences belief formation [13, 11] (rather than changing the perception about the payoffs), our model assumes that the *sender* can use framing, alongside signaling, to influence the decision-maker toward more favorable beliefs. For a given problem instance, we consider a possible framing space C, where each framing $c \in C$ leads to a receiver belief μ_c, with this mapping defined by societal norms. In other words, we are agnostic to the process through which framing influences beliefs, allowing for both Bayesian and non-Bayesian perspectives. Signaling, on the other hand, is an information revelation strategy committed by the sender that correlates signals to the observed world state. The belief update induced by signaling is Bayesian, mirroring the persuasion literature [19, 12]. This model gives rise to three possible solution combinations for the sender:

(a) The framing (and thus the receiver's initial belief) is fixed and cannot be altered; the sender may only optimize the signaling scheme.

(b) The sender's signaling scheme is fixed, and they may only optimize over the framing.

(c) The sender may jointly optimize over both framing and their signaling scheme.

The first setting can be seen as facing a receiver with a fixed but possibly distinct (from the sender) prior belief. This is essentially captured by the large literature on Bayesian Persuasion. The key thrust of this work is to study, both theoretically and empirically, settings (b) and (c). Setting (b) is relevant where the sender must abide by an information revelation scheme they have committed to in the past (e.g., abiding to 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.

Observing that framing is often conveyed through natural language, we consider a linguistic framing space. Any quantitative treatment of such a setting must reconcile two key challenges: (a) how to discern what belief a given framing may induce — i.e. the mapping from natural language framing c to a mathematical distribution μ_c over the world states, and (b) how to systematically

search over this large framing space that includes all instance-relevant language expressions. Until recently, these challenges were a major obstacle to formally modeling and optimizing framing as an alternative information design choice. However, recent breakthroughs in Large Language Models (LLMs) offer powerful tools for addressing these two challenges, including understanding beliefs induced by a natural language framing and searching over the language space to produce a good framing. Indeed, as argued by influential recent works in economics (Horton [18], Brand et al. [3], Dillion et al. [10]) 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 [16]. Therefore, it is an appropriate time to revisit and systematically study the use of framing within information design.

1.1 Our Contributions

We propose a systematic, optimization-based perspective to information design that leverages both *framing and signaling*, connecting perspectives from behavioral economics and psychology with a classic model of Bayesian signaling: persuasion. This framework is formalized in Section 2, with two possible design choices outlined for the sender — optimizing only framing and jointly optimizing framing and signaling. We further argue that LLMs can play a key role within this model. In our theoretical investigations for these optimization problems, we abstract the LLM as a *belief oracle* that noisily maps a framing to its induced belief. Considering the framing-only strategy in Section 3, we illustrate the discontinuous nature of sender utility due to framing changes, highlighting that slight changes in framing can have a major impact. This highlights the sensitivity of We also formally show the optimization problem to be NP-Hard, even to approximate. We answer the same set of questions for the joint framing-signaling strategy design in Section 4. The sender utility now becomes continuous in the framing space. We give a Quasi-Polynomial-Time Approximate Scheme (QPTAS) to solve the general optimization problem. Lastly, we flesh out the promises of LLMs for this setting in Section 5. Using a real-estate case study, we demonstrate the ability of LLMs to act as a belief oracle, mapping framing to beliefs, and propose an end-to-end approach that combines LLMs and mathematical solvers to search the joint framing-signaling space to return strong candidates. Implications and open directions stemming from our work are discussed in Section 6.

1.2 Additional Related Works

While our study of information design using framing from a theoretical and LLM-based perspective is novel, our work does connect to three different lines of research. The first is the algorithmic study of information design, which has attracted significant recent interest. This literature starts from the complexity-theoretic study initiated by Dughmi and Xu [12], and lately has integrated many aspects of machine learning to address unknowns in the setting [5, 15]. The only work that we are aware of on the interface of persuasion and LLMs is [17], which studies a learning-theoretic question about learning optimal sender signaling scheme from simulation feedback generated by LLMs. This differs from our aim of introducing a new dimension, i.e., framing, to information design. The second line of work is the economic research on framing, starting from the extremely influential work of [29]. To the best of our knowledge, the main focus of this literature is to analyze how the framing (i.e., description) of a decision making problem or a game description will affect the

decisions or play by agents. Particularly relevant to us are a few influential findings showing that the framing of an economic decision problem or persuasion can affect agents' play through influencing their beliefs [26, 13, 11]. For instance, Ellingsen et al. [13] show that naming the standard prisoner's dilemma game different (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". Nelson and Oxley [26] observe similar belief influence by framing in persuasion problems. These behavioral studies motivate our principled study of using framing in information design. Thirdly, and more loosely, our work subscribes to the recent studies of using LLMs as proxies of human agents to understand the social norms they carry [18, 3, 20, 10]. Our work subscribes to this general theme, but is different from these works in research questions.

2 Model

Preliminaries: Consider a standard persuasion model with two Bayesian rational players[1]: a *sender* (persuader, with she/her pronouns) and a *receiver* (decision maker, with he/him pronouns). Let $\omega \in \Omega$ be a state of the world which will be known to the sender but not the receiver. The sender has a prior belief $\mu_0 \in \Delta(\Omega)$ over these world states, where $\Delta(\Omega)$ denotes the set of all distributions over Ω. Assume that the prior probability $\mu_0(\omega) > 0$ for every state $\omega \in \Omega$. The receiver chooses an action a from some finite action set $\mathcal{A}$, which, along with the realized state ω, jointly determine the utilities of both players[2]. Formally, the sender's utility function is $u : \mathcal{A} \times \Omega \to \mathbb{R}$ and the receiver's utility function is $v : \mathcal{A} \times \Omega \to \mathbb{R}$. Without loss of generality, we assume that 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 optimal for the receiver (i.e., $\mathbb{E}_{\omega\sim\mu}[v(\omega, a)] > \mathbb{E}_{\omega\sim\mu}[v(\omega, a')]$ for all $a' \in \mathcal{A} \setminus \{a\}$); indeed if this were not so, such an action can be safely ignored since the receiver will not take it under any circumstance. 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 within $[0, 1]$.[3]

Signaling: We propose a generalization of information design where the sender has two possible levers through which she may influence the receiver's actions. First, she may design and commit to a signaling scheme to partially reveal the realized state ω. Formally, for a signaling space $\mathcal{S}$, the sender commits to a policy $\pi : \Omega \to \Delta(\mathcal{S})$, 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 scheme, updates their belief over Ω and takes the expected utility maximizing action.

Framing: Novel to our work, and motivated by extensive behavioral studies, we posit that the receiver's belief can also be shaped through *framing*. Framing can most naturally be thought of as 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,

[1] We model framing c as updating the receiver to some belief μ_c; this approach allows the framing induced update to be possibly non-Bayesian

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

[3] If utilities are instead bounded within $[-A, B]$, then it can be normalized to our $[0, 1]$ setting without loss of generality.

consider a function $\ell : C \to \Delta(\Omega)$ (also called a belief oracle) that directly maps a framing c to the corresponding belief it induces in the receiver μ_c, with the set $B = \{\ell(c) : c \in C\}$ denoting 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 ω, which was 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 by our model.

Our theoretical analysis considers oracle access to a possibly noisy/imperfect belief oracle ℓ_ε, which maps each framing to within ε of the true belief; formally, $\forall c \in C$, $||\ell(c) - \ell_\varepsilon(c)||_2 \leq \varepsilon$. Under this setting, we formally consider the optimization problem of selecting the best c and the effect of error ε in Sections 3 and 4. In Section 5, we empirically show that LLMs can indeed be used to robustly approximate the framing-to-belief mapping function ℓ, serving as a justification for this model primitive.

The framing space C (and thus the corresponding inducible belief set B) can be thought of as either discrete or continuous. A discrete space is natural when considering framing in terms of phrases or language. Conversely, the continuous perspective can be informative from a theoretical perspective. In the continuous model, we assume the inducible beliefs form a convex subset of the simplex $\Delta(\Omega)$. While this may not immediately map to practice since linguistic framing space is still discrete and not every belief is necessarily inducible through language, optimizing directly over this continuous region is ostensibly a richer problem than optimizing over a large discrete set. Thus, computational and structural results proved under this model are generally more insightful. Our work considers both discrete and continuous framing spaces in the theoretical results, with the experiments focusing on the former. 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: To map the model so far to our running example of advertising, the advertiser can choose a slogan or description (the framing) that will accompany their product, regardless of its hidden features/states (the state) or discount/buy recommendation (the signal). Consistent with prior literature on persuasion, the sender chooses their strategy (framing and signaling scheme) before state realization; thereafter, the receiver takes their best response action. This outlines a leader-follower game:

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

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

In Proposition 1, we show that randomizing over framing does not increase sender utility under any variant of our problem. Hence, it is without loss of generality to consider them choosing a

single/fixed framing. Since the receiver will best-respond to the sender's choice of information design strategy, the sender's goal is to choose an optimal strategy that maximizes their ex-ante utility under this best-responding behaviour. Formally, this corresponds to a Stackelberg equilibrium.

We study two variants of the sender's design problem, corresponding to the solution combinations mentioned in the introduction. First, the sender may only optimize the framing c under a given/fixed signaling scheme. We denote this as the *framing-only strategy* and define such an instance by $\mathcal{I} = (\mu_0, u, v, \pi)$. Secondly, the sender may optimize both elements of the information design tuple (c, π) with such a *joint strategy* instance defined by $\mathcal{I} = (\mu_0, u, v)$. Note that the third possible combination, optimizing only signaling for a fixed framing (i.e. 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 tuple* (c, π) *is:*

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

where $a^*_{c,s}$ *is defined in Eq.* (1)*. In the framing-only strategy space, the sender's equilibrium strategy is* $\arg\max_{c\in C} \mathbb{E}_{\omega,s}[u(a^*_{c,s}, \omega)]$ *and under the joint strategy space, it is* $\arg\max_{c\in C,\pi} \mathbb{E}_{\omega,s}[u(a^*_{c,s}, \omega)]$*.*

3 Framing-Only Strategy

Consider choosing a framing when the signaling scheme is fixed/given. Formally, given an information design instance $\mathcal{I} = (\mu_0, u, v, \pi)$, the sender may only choose a framing $c \in C$. The optimal framing c^* corresponds to the Stackelberg equilibrium since choosing π is not part of the strategy space. While we do not make any assumptions on π, we highlight an important observation. In the classic literature on persuasion, it is without loss of generality to consider the signaling space $|S|$ to be as large as the action space $|A|$ by leveraging a revelation-principle style argument [12]. This rests on combining signals that lead to the same action into a single "action-recommending" signal. In the framing-only optimization setting, this ceases to be true since π is fixed in advance, and the mapping from signals to receiver actions can be different under different receiver beliefs. As such, we make no assumptions about the size of the signal space or its interpretation in this section.

Why do we consider choosing a single framing $c \in C$, and not a distribution over framing? The result below shows that randomizing over framings does not increase utility for the sender under *any* scheme π. This means that it is without loss of generality to consider deterministic framing both here and in the forthcoming section that studies jointly optimizing π and c (see Section 4).

Proposition 1. *For any instance* $\mathcal{I}$*, the optimal sender utility is not improved by choosing a distribution over framings* $\Delta(C)$*.*

3.1 The Effect of Framing on Sender Utility

How much can the sender improve their utility by manipulating framing when the signaling scheme π is given? Does meaningful improvement require a framing whose induced belief is substantially

far away from the underlying sender's prior? How much does the optimal sender utility suffer if they have a noisy mapping ℓ_ε? To build intuition for these questions, consider the judge-lawyer example Kamenica and Gentzkow [19] used to motivate canonical Bayesian Persuasion. A defendant may either be *innocent* or *guilty* (the 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 lawyer (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 lawyer, when innocent, always recommends acquittal, and when guilty, recommends either action with probability 0.5. If the lawyer'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 lawyer achieves utility 0. If, however, the lawyer can use framing (style of argument/language) to slightly alter the judge's belief to $[\frac{2}{3}, \frac{1}{3}]$, then the lawyer's utility under this very same signaling jumps to 0.66! This is due to the sender's utility as a function of the receiver's belief in this instance not being continuous for the given scheme. This discontinuity highlights the power of leveraging framings: for a fixed signaling, slightly altering the receiver's belief by framing can have a major impact on the sender's expected utility.

To formalize this, let $U_\pi(\mu)$ denote the sender's expected utility for fixed signaling scheme π when the receiver's prior belief is μ. In the judge-lawyer 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. Note that schemes not satisfying this condition are very revealing: upon observing any signal s, the receiver can rule out certain state(s) with full confidence. The proof of this result is presented in Appendix 7.

Proposition 2 (Discontinuous sender utility)**.** *Suppose for signaling scheme π 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, if u and v are sampled from any continuous distribution over utility values, then $U_\pi(\mu)$ 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. 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 $||U_\pi(\hat{\mu}) - U_\pi(\mu^*)|| \leq \lambda\varepsilon$. We next show that this discontinuity also makes finding the optimal c computationally challenging, even with a perfect oracle.

3.2 Computing the Optimal Framing

3.2.1 Discrete Framing Space

How can we find the optimal framing for a given instance. For any framing c and given signaling π and belief oracle ℓ, the corresponding sender utility can be computed in time poly$(|S|, |\mathcal{A}|, |\Omega|)$: use

the belief oracle to determine the induced belief μ_c and then for each signal $s \in S$, the receiver's best response $a^*_{c,s}$ can be calculated as in Eq. (1), and then the corresponding sender utility as in Eq. (2). When the framing space is discrete, this yields a simple enumeration algorithm to determine the optimal framing: enumerate all $|C|$ possible framings for the given problem instance, and choose the one with the highest utility. This terminates in $|C| \cdot \text{poly}(|S|, |\mathcal{A}|, |\Omega|)$ time and is only reasonable when the framing space is small. In our running examples, where framing encompasses contextually relevant natural language expressions, this is unlikely to be the case. This begs the question: are there algorithms that scale more gracefully in $|C|$, while preserving the polynomial dependence on the other instance parameters (Ω, $\mathcal{A}$, and S)?

3.2.2 Continuous Framing Space

To improve the dependency on $|C|$, we need to leverage additional structure in the framing space (i.e. how framings relate to the induced beliefs). By considering the continuous framing space where any belief in some convex subset $B \subseteq \Delta(\Omega)$ is inducible, we can leverage this structure. Indeed, 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 (Ω, $\mathcal{A}$, 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 hardness of an algorithm that receives 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:

$$\begin{aligned}\underset{\mu_c \in B = \Delta(\Omega)}{\text{maximize}}\, U_\pi(\mu_c) \;=\; & \underset{\mu_c \in B = \Delta(\Omega)}{\text{maximize}} \sum_s \sum_\omega \mu_0(\omega)\pi(s|\omega)u(a^*(\mu_c, s), \omega) \\ & \text{s.t.} \quad a^*(\mu_c, s) = \arg\max_{a \in \mathcal{A}} \sum_{\omega'} \mu_c(\omega')\pi(s|\omega')v(a, \omega')\end{aligned}$$

Computing the optimal objective value of this optimization problem is NP-Hard. We reduce this from the Bayesian Stackelberg game where a leader faces a follower of unknown type. Conitzer and Sandholm [6] formally show the following:

Lemma 1 (Conitzer and Sandholm [6])**.** *The* Bayesian Stackelberg Game *(BSG) consists of a leader with action space $\mathcal{A}_\ell$ and a follower of unknown type θ (from known distribution $P(\theta) \in \Delta(|\Theta|)$) with action space $\mathcal{A}_f$. With leader utility $u_\ell(a_\ell, a_f)$ and type-dependent follower utility $u^\theta_f(a_\ell, a_f)$, the leader must commit to a mixed strategy $x(a_\ell)$, noting that the receiver will best respond. The leader's optimal utility is given by:*

$$\begin{aligned}&\underset{x \in \Delta(\mathcal{A}_\ell)}{\text{maximize}} \sum_\theta P(\theta) \sum_{a_\ell} x(a_\ell) u_\ell(a_\ell, a^*_f(\theta, x)) \\ &\text{s.t.} \quad a^*_f(\theta, x) = \arg\max_{a_f \in \mathcal{A}_f} \sum_{a_\ell} x(a_\ell) u^\theta_f(a_\ell, a_f)\end{aligned}$$

It is NP-Hard to compute this optimal leader utility even when the follower's action space is binary.

We show that any BSG can be cast into an optimal framing-only instance $\mathcal{I}$ with $|\Omega| = |\mathcal{A}_\ell| + |\Theta| + 1$ states, $\mathcal{A} = |\Theta||\mathcal{A}_f| + 2$ actions, $|S| = |\Theta|$ signals, and a continuous framing space $C = \Delta(\Omega)$. The proof is technical and formally given in Appendix 7; but we sketch the high-level intuition below.

Theorem 1. *For a framing-only instance $\mathcal{I} = (\mu_0, u, v, \pi)$ with continuous framing-induced belief space $B = \Delta(\Omega)$, it is NP-Hard to select the optimal framing belief μ_c^*. Indeed it is NP-Hard to solve this even with an additive approximation error.*

Proof Sketch. 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_i^θ 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_0^θ, a_1^θ, which should directly correspond to follower θ's utility in taking action 0 or 1 in the BSG instance. 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) the persuasion sender places sufficient weight on the optimal μ_c states corresponding to ω_θ to ensure 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 Conitzer and Sandholm [6]. □

4 Joint Framing-Signaling Strategy

We now consider the joint strategy setting wherein the sender can choose both elements of the information design tuple. Formally, given an instance $\mathcal{I} = (\mu_0, u, v)$, the sender may select a framing $c \in C$ and a signaling scheme π. This can be viewed as a generalization of the standard Bayesian Persuasion model, which only designs the scheme π. As in Section 3, there is no benefit in randomizing over framings since Proposition 1 shows that for *any* signaling π, and thus the optimal scheme π^*, randomization can never increase utility.

The ability to choose π offers the sender more freedom as compared to the framing-only strategy. Indeed, a key limitation of that restricted strategy was an inability to apply a revelation-principle style argument: for a fixed π, the mapping from receiver action to observed signal could change depending on the receiver's belief. It turns out that the revelation principle is restored in the design of joint strategy, as we show below (proof in Appendix 8).

Proposition 3. *When jointly optimizing over the information design tuple $(c \in C, \pi : \Omega \to \Delta(\mathcal{S}))$, it suffices to consider signaling scheme π with a direct signal space, i.e., $\mathcal{S} = \mathcal{A}$.*

This allows us to restrict our attention to such direct signaling schemes without loss of generality. As in the preceding section, we focus on 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 computability of the optimal joint strategy with this oracle. Our results here highlight key differences from the framing-only strategy.

4.1 The Effect of Framing on Sender Utility

Given that signaling is part of the strategy space, for any induced receiver belief μ, the sender can use the optimal signaling scheme to accompany this belief. We denote this signaling scheme as π_μ^*

and the resulting sender utility as $U^*(\mu)$. Due to Proposition 3, $U^*(\mu)$ can be efficiently computed for any receiver belief μ using the following linear program (with the constraints referred to as the incentive-compatibility or IC constraints):

$$U^*(\mu) = \underset{\pi:\Omega\to\Delta(\mathcal{A})}{\text{maximize}} \sum_{\omega}\sum_{a} \mu_0(\omega)\pi(a|\omega)u(a,\omega) \tag{3}$$

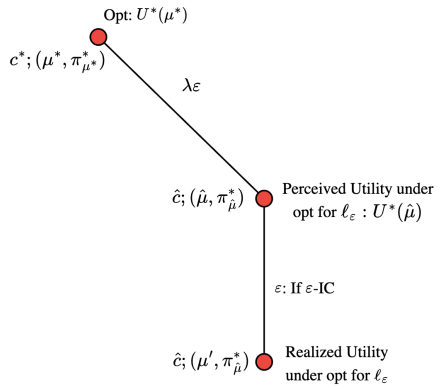
$$\text{s.t. } \forall a, a' : \sum_{\omega} \mu(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0. \tag{4}$$

We can compare this function $U^*(\mu)$, which tracks the sender utility as a function of the receiver belief under the corresponding optimal scheme, to the earlier function $U_\pi(\mu)$ which tracked the same quantity but with a fixed scheme π. Proposition 2 illustrated that $U_\pi(\mu)$ is generally discontinuous, allowing arbitrary changes in sender utility, even when framing only slightly changes the receiver's belief. However, when π is part of the strategy space and can be optimized, this is no longer the case. Theorem 2 below shows that $U^*(\mu)$ is continuous within the interior of the simplex (proof in Appendix 8). The result implies that if two different framings lead to slightly different beliefs, then the corresponding sender utility also changes slightly.

Theorem 2. *The sender's utility $U^*(\mu)$, as defined in Eq. (3), 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 defined in (3)-(4), 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 IC constraint (4) 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 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 for μ', $U^*(\mu')$, cannot be too much worse than $U^*(\mu)$. See details in Appendix 8.2. □

What does this imply about the loss in utility due to a noisy belief oracle ℓ_ε? Let $(\mu^*, \pi^*_{\mu^*})$ denote the optimal 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 $\ell_\varepsilon(c^*)$ in the ball $B_\varepsilon(\mu^*) = \{\mu \in \Delta(\Omega) : \|\mu - \mu^*\| \leq \varepsilon\}$, and due to Theorem 2, the perceived utility will be within $O(\varepsilon)$ of the optimal: $|U^*(\ell_\varepsilon(c^*)) - U^*(\mu^*)| \leq O(\varepsilon)$. Then, by choosing the optimal framing $\hat{c} \in C$ according to the noisy oracle, with corresponding belief $\ell_\varepsilon(\hat{c}) = \hat{\mu}$, we obtain $U^*(\hat{\mu}) \geq U^*(\ell_\varepsilon(c^*)) \geq U^*(\mu^*) - O(\varepsilon)$. When using framing $\hat{c}$ and signaling scheme $\pi^*_{\hat{\mu}}$ in practice, however, the actual belief induced by $\hat{c}$ is $\mu' = \ell(c) \neq \hat{\mu}$ satisfying $||\mu' - \hat{\mu}|| \leq \varepsilon$. The signaling scheme $\pi^*_{\hat{\mu}}$ is IC for $\hat{\mu}$ but not necessarily IC for μ'. One way to resolve this non-IC issue is to modify the signaling scheme $\pi^*_{\hat{\mu}}$ slightly to make it IC for μ'; this is feasible because μ' is close to $\hat{\mu}$, but will cause an additional $O(\varepsilon)$ loss to the sender's utility. Another way is to relax the IC notion, as follows.



We introduce the notion of ε-approximate incentive compatibility (ε-IC). Formally, an action a is ε-IC for a receiver with belief μ if for all actions $a' \in \mathcal{A}$, $\sum_\omega \mu(\omega)\pi(a|\omega)[v(a,\omega) - v(a',\omega)] \geq -\varepsilon$. In maintaining a fixed signaling $\pi^*_{\hat{\mu}}$ and shifting the induced belief from $\hat{\mu}$ to μ', it is evident that the IC constraints are violated by at most ε, and the utility perturbed at most ε as well. We visualize this on the left, and by the triangle inequality, the following is evident:

Corollary 1. *The realized utility in facing a ε-IC receiver under a joint optimal strategy based on an ℓ_ε noisy oracle is at most $O(\varepsilon)$ away from the optimal utility for an exactly IC receiver.*

The notion of ε-IC will feature prominently throughout the results in this section. As mentioned above, it enables more tractable computational and characterization results compared to the earlier section on context-only optimization. It is important to note that this relaxation is conceptually reasonable in *direct* signaling schemes where signals can be interpreted as action recommendations. While it is without loss of generality to consider such schemes when jointly optimizing framing and signaling (Proposition 3), when only optimizing framing as in Section 3, this is not true, making ε-IC ill-defined in that setting.

4.2 Computing the Optimal Joint Strategy

4.2.1 Discrete Framing Space:

In the discrete framing space, for any framing c with induced belief μ_c, we can compute the optimal signaling $\pi^*_{\mu_c}$ and its resulting utility $U^*(\mu_c)$ using the linear program specified in (3), which is computable in $\text{poly}(|\Omega|, |\mathcal{A}|)$. As in Section 3, this yields a simple enumeration algorithm to find the optimal framing that runs in time $|C| \cdot \text{poly}(|\Omega|, |\mathcal{A}|)$. While this approach may be similarly impractical for large $|C|$, unlike Section 3, the function $U^*(\mu)$ here is Lipschitz continuous in the interior of the simplex. Thus, from a practical perspective, it may be attractive to enumerate over a smaller set of framings that yield sufficiently different induced beliefs and obtain an efficient approximation. The exact mechanics of this depend on the properties of the mapping function ℓ and the corresponding density of the belief space B. Studying the continuous framing space provides more insights into these questions.

4.2.2 Continuous Framing Space:

Recall that in the continuous framing space model, we directly consider inducing a belief within some convex region $B \in \Delta(\Omega)$. The optimization problem can then be expressed as maximizing a

linear objective subject to bi-linear constraints:

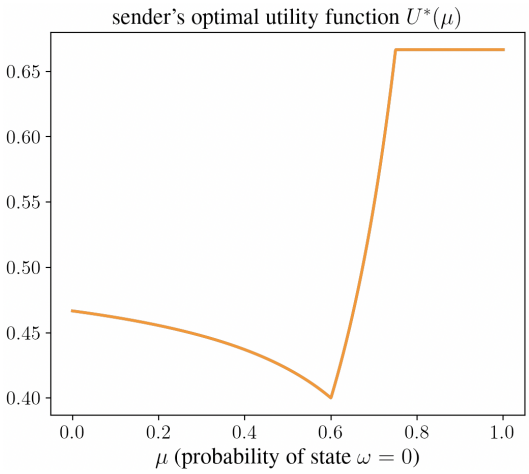
$$\underset{\mu \in B}{\text{maximize}}\, U^*(\mu) \;=\; \underset{\mu \in B,\, \pi:\Omega \to \Delta(\mathcal{A})}{\text{maximize}} \;\; \sum_{\omega} \sum_{a} \mu_0(\omega)\pi(a|\omega)u(a,\omega) \tag{5}$$
$$\text{s.t. } \forall\, a, a' \in \mathcal{A} \times \mathcal{A} : \sum_{\omega} \mu(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0.$$

Bi-linear optimization problems do not generally admit efficient solutions. To get a sense of the challenge for our specific problem, we observe that $U^*(\mu)$ is neither concave nor quasi-concave even for simple instances. We illustrate such an example below.

Example 1. *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.



This suggests that the underlying problem may be hard in the general case. Indeed, we saw that the analogous problem in the framing-only strategy setting is NP-Hard to approximate even when the induced belief space covers the entire simplex. Results in the joint optimization case, however, are not as pessimistic, mirroring other observations from this section. Specifically, under the same condition of inducible belief space B equaling $\Delta(\Omega)$, there exists a poly-time algorithm achieving $\left(1 - \frac{1}{|\Omega|}\right)$ approximation when the receiver is ε-IC. Further, for state-independent sender utility (i.e. sender utility only depends on the receiver action as in the judge-lawyer example), the exact optimal for an exactly IC receiver can be computed in poly-time. We formalize these results below:

Theorem 3. *For a joint strategy instance $\mathcal{I} = (\mu_0, u, v)$, and framing-induced 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;*
- *If the sender 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:

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

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

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

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 $[0, 1]$.

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 is inducible in the receiver - that is, for any action $a \in \mathcal{A}$, there exists some belief μ_a wherein taking that action is optimal for the receiver. For the sender optimal action a^u, let μ_{a^u} denote the belief where it 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 $B = \Delta(\Omega)$, choose the framing c that 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. Using this joint strategy means the receiver 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. □

These results rely on the inducible belief space being the entire simplex. We may also be interested in results that hold when B is any convex subset of the belief simplex. We now show that a Quasi-Polynomial Time Approximate Scheme (QPTAS) exists for this general problem (Theorem 4). Importantly, the algorithm computes a joint strategy with at least as much utility as the optimal exact-IC solution, but with ε-IC guarantees. While we do not formally prove hardness for the exact problem (we regard it as an important open direction), the existence of a QPTAS suggests that the general version of this problem is easier than the framing-only variant since Independent-Set, which has no known QPTAS, reduces to the latter.

Theorem 4. *For any instance $\mathcal{I}$ and any $\varepsilon > 0$, there exists a $\text{poly}(|\Omega|^{\frac{\log|\mathcal{A}|}{\varepsilon^2}})$-time algorithm that can compute an ε-IC joint strategy with at least as much utility as the optimal joint strategy under exact IC.*

Proof. Let $\mu^* \in B$ and π^* be the optimal induced belief and signaling scheme for this instance under exact incentive compatibility 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-IC 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-IC 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 only depends on π^*.

Now consider the following algorithm: enumerate over all n-uniform distributions in B, and for each, solve the optimal signaling linear program (3)-(4) but with a relaxed ε-IC constraint. Return the solution with the best sender utility. This solution must be weakly better than the ε-IC 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 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 $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ 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 $|\Omega| - 1$. It is thus clearly upper bounded by $O(|\Omega|^n)$. Similarily, for a fixed n and as $|\Omega|$ grows large, observe that $\binom{n+|\Omega|-1}{|\Omega|-1} = \binom{n+|\Omega|-1}{n} \approx \frac{|\Omega|^n}{n!} = O(|\Omega|^n)$, where we use Stirling's approximation. 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 studies of framing and signaling are worth noting: (a) our model rests on direct access to the framing-to-belief mapping $\ell : c \to \mu_c$; (b) our theoretical studies highlight the computational challenges of optimizing over the framing space C, even with access to such an oracle. In this section, we turn to empirical studies and harness the power of Large Language Models (LLMs) to address these issues. Specifically, we empirically show that LLMs not only can be used to approximately uncover the mapping ℓ, but also can help to search over the framing space efficiently to find good framing candidates.

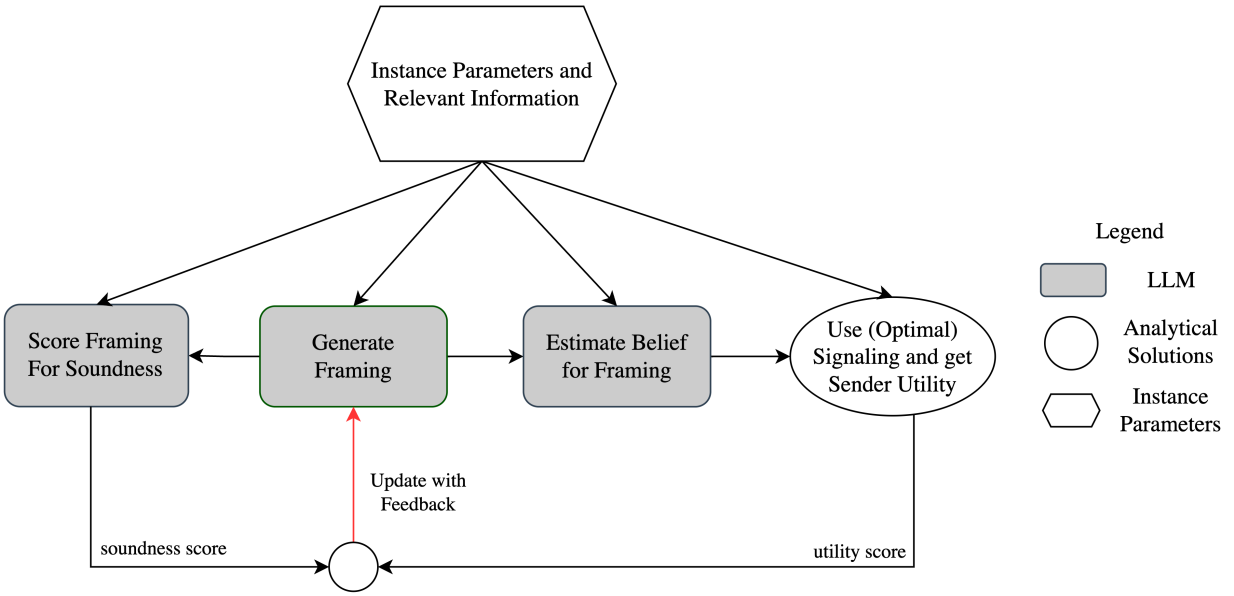


Figure 1: Diagram of our proposed framework for optimizing framing and signaling. It includes LLMs searching the framing space, verifying it for correctness, and generating framing-induced beliefs. It also includes poly-time analytical solvers to compute optimal signaling for a given belief.

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 them based on feedback, and estimating the belief induced by a receiver under any framing. These are complemented with an additional LLM module that verifies the soundness of any framing, and analytical solvers that compute the sender utility given a quantitative receiver belief. 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 in economics argues that LLMs can approximate this agent in various settings [18, 22], including as a statistical proxy to model beliefs of a human agent [25]. This approach involves endowing the LLM with information about the agent it is modeling - in our case, this is 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. [8] 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). In contrast, they show that eliciting numerical probabilities in a chat-style prompt results in better outcomes. We use this in our framework and comment on experimental observations about this approach in Section 5.3.

Validation of a 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 [21, 4]. 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 Equation 3. 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" [28], we propose 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 Searching the Framing Space: 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 over 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 9; 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.

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 considerd 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 [27] for the LLM portions of the framework, and SciPy [30] for the analytical parts. The results for each instance are presented below. 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 Equation 5 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.

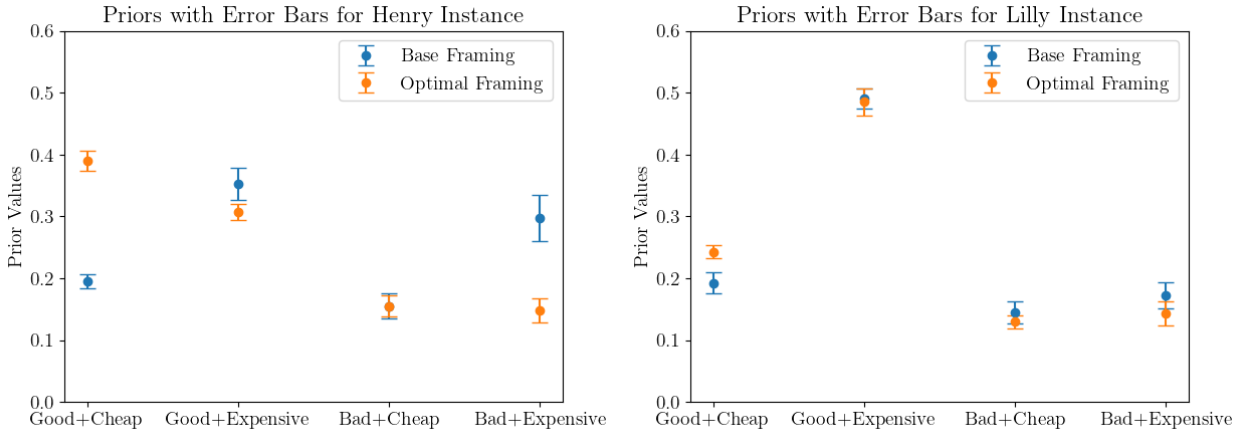
Framing for the "Henry" Instance	**Utility**
No framing - receiver prior equal to sender prior	0.28
Realtor Jeremy Profile from the Instance Description	0.30
Best LLM Framing: *Meet Jeremy Hammond, a dedicated realtor with over 8 years of experience, specializing in finding the perfect homes for outdoor 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 when $B = \Delta(\Omega)$)	0.41

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

5.3 Estimating Receiver Beliefs under Different Framing: A Real-Estate Case Study

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 results presented above. When decision making has been proxied to LLMs as argued in setting like [16], the belief LLMs generate for a given framing can essentially be considered ground truth. When the decision-maker is human, however, we use LLMs to approximate this agent. For both scenarios, it is instructive to question the consistency of these LLM beliefs. We consider consistency along two angles: variance and rationality. Variance refers to how different the beliefs returned by the LLM are across multiple runs[4]. Rationality refers to whether when asked to make decisions, the resulting action is consistent with the belief.

[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.



(a) Belief distribution for Henry

(b) Belief distribution for Lilly

Figure 2: Mean (across 20 runs) LLM generated beliefs with 90% confidence intervals at default temperature

In Figure 3, we plot the belief distributions for both the Henry and Lilly instances. It contains beliefs corresponding to both the initial realtor description defined in the instance (denoted "base framing"), and the optimal one generated by the LLM. We note that while there is variance between the runs, the beliefs generated by the LLM are reasonably consistent. These plots also highlight the effect of optimal framing in increasing the belief in "good+cheap" state and decreasing the belief in "bad+expensive" state.

To verify the "rationality" of these beliefs we consider the following. First, we prompt the LLM with the receiver's 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 *pre-belief 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 the optimal decision for each generated belief (denoted by "% actions are rational"). As we see in the presented results below, post-belief decisions are always consistent with the optimal action for each prior. The instinctive pre-belief decisions are slightly less so for Henry, but perfectly matches for Lilly.

Instance	**Pre-Belief Action**	**Post-Belief Action**	**% Actions that are Rational**
Henry	17 "buy" and 3 "not buy"	20 "buy" and 0 "not buy"	100%
Lilly	20 "buys" and 0 "not buy"	20 "buy" and 0 "not buy"	100%

Table 1: Results of 20 independent runs of the consistency experiment for the optimal framing generated for each instance.

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 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 that 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, persuasion is but one model and it would be instructive to combine the rich perspective of framing with other important signaling models such as cheap-talk [7, 14] or mediation [24].

References

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

[2] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience theory of choice under risk. *The Quarterly journal of economics*, 127(3):1243–1285, 2012.

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

[4] 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.

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

[6] 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.

[7] Vincent P. Crawford and Joel Sobel. Strategic Information Transmission. *Econometrica*, 50 (6):1431, November 1982. ISSN 00129682. doi: 10.2307/1913390. URL `https://www.jstor.org/stable/1913390?origin=crossref`.

[8] 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`.

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

[10] 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.

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

[12] 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. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897583. URL `https://dl.acm.org/doi/10.1145/2897518.2897583`.

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

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

[15] 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.

[16] 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.

[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] 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.

[22] 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.

[23] Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. Coarse thinking and persuasion. *The Quarterly journal of economics*, 123(2):577–619, 2008.

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

[25] 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.

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

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

[28] 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.

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

[30] 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. doi: 10.1038/s41592-019-0686-2.

[31] 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. ISBN 978-1-4503-8554-1. doi: 10.1145/3465456.3467593. URL `https://dl.acm.org/doi/10.1145/3465456.3467593`.

7 Section 3 Appendix

7.1 Proof of Proposition 1

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

$$\begin{aligned}\mathbb{E}_{c,\omega,s}[u(a^*_{c,s},\omega)] &= \sum_\omega \mu_0(\omega) \sum_{s\in S} \pi(s|\omega) \sum_{c\in C} p_C(c) u(a^*_{c,s},\omega) \\ &= \sum_{c\in C} p_C(c) \sum_\omega \mu_0(\omega) \sum_{s\in S} \pi(s|\omega) u(a^*_{c,s},\omega)\end{aligned}$$

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

$$\mathbb{E}_{c,\omega,s}[u(a^*_{c,s}, \omega)] \leq \sum_{c \in C} p_C(c) \sum_\omega \sum_s \mu_0(\omega)\pi(s|\omega)u(a^*_{c_{max},s}, \omega) = \sum_\omega \sum_s \mu_0(\omega)\pi(s|\omega)u(a^*_{c_{max},s}, \omega)$$

□

7.2 Proof of Proposition 2

Proof. 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 $\tilde{\mu}$. □

7.3 Proof of Theorem 1

Proof. 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 [6] 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 $\text{BS}(\mathcal{I}_{BS}, x^*)$, we will give a poly-time construction of an OF problem instance $\mathcal{I}'_{OF}$ whose optimal solution μ^*_c is such that $\text{OF}(\mathcal{I}'_{OF}, \mu^*) = \text{BS}(\mathcal{I}_{BS}, x^*)$.

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: $\mathcal{A} = \{a_0^\theta\}_{\forall \theta \in \Theta} \cup \{a_1^\theta\}_{\forall \theta \in \Theta} \cup \widetilde{a}_1 \cup \widetilde{a}_2$
- 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 $u(a, \omega)$:
  - $\forall a_\ell, u(a_*^\theta, \omega_{a_\ell}) = u_\ell(a_*)$[5]
  - $\forall \theta, u(a_*^{\theta'}, \omega_\theta) = -L$ 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:

$$\underset{\mu_c}{\text{maximize}} \underbrace{(1-\varepsilon)\sum_\theta P(\theta) u(a^*(\mu_c, s_\theta))}_{\text{state } \widetilde{\omega}} + \underbrace{\frac{\varepsilon}{|\Theta|}\sum_\theta u(a^*(\mu_c, s_\theta), \omega_\theta)}_{\text{for states } \omega_\theta \text{ where } \pi(s_\theta|\omega_\theta) = 1} \tag{6}$$

$$\text{s.t} \quad a^*(\mu_c, s_\theta) = \arg\max_{a \in \mathcal{A}} \left[\mu_c(\omega_\theta) v(a, \omega_\theta) + P(\theta)\mu_c(\widetilde{\omega}) v(a, \widetilde{\omega}) + \frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c(\omega_{a_\ell}) v(a, \omega_{a_\ell}) \right] \tag{7}$$

[5] where $* \in \{0, 1\}$

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 2. *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 incentive-compatible for the receiver to take action a_0^θ as opposed to $\widetilde{a}_2$ on receiving a signal s_θ:

$$\mu_c(\omega_\theta)[v(a_0^\theta, \omega_\theta) - v(\widetilde{a}_2, \omega_\theta)] + \frac{1}{|\Theta|}\sum_{a_\ell} \mu_c(\omega_{a_\ell})[v(a_0^\theta, \omega_{a_\ell}) - v(\widetilde{a}_2, \omega_{a_\ell})] \tag{8}$$

$$= -\mu_c(\omega_\theta)K + \frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c(\omega_{a_\ell})v(a_0^\theta, \omega_{a_\ell}) = \frac{1}{|\Theta|}\left(1 - \frac{v^{max}}{M}\right) - \frac{Kv^{max}}{|\Theta|M} \geq 0 \tag{9}$$

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 $v^{max} \leq \frac{M}{1+K}$[6]. 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 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^*, 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) $\exists$ *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):

$$\underbrace{-N(1-\varepsilon)P(\theta)}_{\text{on } \widetilde{\omega} \text{ and signal } \theta} + \underbrace{(1-\varepsilon)}_{\text{on } \widetilde{\omega} \text{ and other signals}} - \underbrace{\frac{N\varepsilon}{|\Theta|}}_{\text{on } \omega_\theta \text{ and } s_\theta} \leq -N(1-\varepsilon)P(\theta) + 1 < \frac{-P(\theta)}{P_{min}} + 1 \leq 0 \tag{10}$$

[6]We assume ties break in favour of a^θ actions

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. (10) simply becomes $c\left[\frac{P(\theta)}{P_{min}}+1\right] \leq 0$ for the same reason as above.

(2) $\exists$ *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:

$$-K(1-\varepsilon)P(\theta) + (1-\varepsilon) - \frac{K\varepsilon}{|\Theta|} \leq -K(1-\varepsilon)P(\theta) + 1 \leq \frac{-P(\theta)}{P_{min}} + 1 \leq 0 \tag{11}$$

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

$$(1-\varepsilon)\underbrace{\sum_{s_\theta} P(\theta)u(a^*(\mu_c, s_\theta))}_{\text{at most 1}} + \frac{\varepsilon}{|\Theta|}\underbrace{u(a_*^{\theta'}, \omega_\theta)}_{-L} + \frac{\varepsilon}{|\Theta|}\sum_{s_{\hat{\theta}}}\underbrace{u(a^*(\mu_c, s_{\hat{\theta}}), \omega_{\hat{\theta}})}_{\text{at most 0}} \tag{12}$$

$$\leq (1-\varepsilon) - \frac{L\varepsilon}{|\Theta|} \leq 1 - \frac{L\varepsilon}{|\Theta|} < 0 \tag{13}$$

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 13 simply becomes $c\left[(1-\varepsilon) - \frac{L\varepsilon}{|\Theta|}\right] < 0$ which still becomes negative due to the choice of L. □

Lemma 4. *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 incentive compatibility for a_θ implies:

$$-\mu_c(\widetilde{\omega})P(\theta)v(a', \widetilde{\omega}) - \mu_c(\omega_\theta)v(a', \omega_\theta) + \frac{1}{|\Theta|}\sum_{a_\ell}\mu_c(\omega)[v(a^\theta, \omega_{a_\ell}) - v(a', \omega_{a_\ell})] \geq 0 \ \forall a' \tag{14}$$

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. (14) 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. (14), 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 $BS(\mathcal{I}_{BS}, x^*) = \text{OF}(\mathcal{I}'_{OF}, \mu_c^*)$.

$\implies$ 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 incentive compatibility with respect to a_0^θ (the only other action possible since claim 3 disavows all others) implies:

$$\frac{1}{|\Theta|}\sum_{\omega_{a_\ell}} \mu_c^*(\omega_{a_\ell})[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] + \mu_c^*(\omega_\theta)\underbrace{[v(a_1^\theta, \omega_\theta) - v(a_0^\theta, \omega_\theta)]}_{0} \geq 0 \tag{15}$$

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:

$$0 \leq \sum_{a_\ell} x(a_\ell)[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] = \sum_{a_\ell} x(a_\ell)[u_f^\theta(a_1, a_\ell) - u_f^\theta(a_0, a_\ell)] \tag{16}$$

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:

$$\text{OF}(\mu_c^*) = (1-\varepsilon)\sum_\theta P(\theta)u(a^*(\mu_c^*, s_\theta)) = \sum_\theta P(\theta)u_\ell(a_f^*(x', \theta)) = BS(x') \tag{17}$$

$\impliedby$ Suppose we have an optimal solution to the x^* to the BSG instance $\mathcal{I}_{BS}$. Then by definition, the incentive compatibility 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:

$$\sum_{a_\ell} x^*(a_\ell)[u_f^\theta(a_1, a_\ell) - u_f^\theta(a_0, a_\ell)] \geq 0 \tag{18}$$

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. (18) holds for $x(a_\ell)$, and μ_c' is simply a rescaling of $x(a_\ell)$ on the ω_{a_ℓ} states, and $u_f^\theta(a_*, a_\ell) = v(a_*^\theta, \omega_{a_\ell})$:

$$\sum_{\omega_{a_\ell}} \mu_c'(\omega_{a_\ell})[v(a_1^\theta, \omega_{a_\ell}) - v(a_0^\theta, \omega_{a_\ell})] \geq 0 \tag{19}$$

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:

$$\text{BS}(x^*, \mathcal{I}_{BS}) = \sum_\theta P(\theta) u_\ell(a_f^*(x, \theta)) = (1-\varepsilon) \sum_\theta P(\theta) \frac{1}{(1-\varepsilon)} u_\ell(a_f^*(x, \theta)) \tag{20}$$

$$= (1-\varepsilon) \sum_\theta P(\theta) u(a^*(\mu_c', s_\theta)) = \text{OF}(\mu', \mathcal{I}'_{OF}) \tag{21}$$

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 [6] 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 [6] 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. □

8 Section 4 Appendix

8.1 Proof of Proposition 3

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:

$$\sum_{\omega} \sum_{a} u(a, \omega) \sum_{s \in m(a)} \pi^*(s|\omega) \tag{22}$$

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. (22) in unchanged by using this direct scheme with signal space S equal A as action recommendations. □

8.2 Proof of Theorem 2

Proof. 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 Eq. (3). 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 5 (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 5. 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,

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

For the first term above,

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

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

$$\frac{\pi(s|\omega)}{\pi(s)} = \frac{\pi(s|\omega)}{\sum_{\omega' \in \Omega} \mu(\omega')\pi(s|\omega')} \leq \frac{\pi(s|\omega)}{p_0 \sum_{\omega' \in \Omega} \pi(s|\omega')} \leq \frac{1}{p_0}. \tag{23}$$

$$\implies \sum_{\omega \in \Omega} \left| \frac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \frac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} \right| \leq \sum_{\omega \in \Omega} \frac{1}{p_0} |\mu(\omega) - \mu'(\omega)| = \frac{1}{p_0} \|\mu - \mu'\|_1. \tag{24}$$

For the second term,

$$\begin{aligned}
\sum_{\omega \in \Omega} \left| \frac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \frac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)} \right| &= \sum_{\omega \in \Omega} \mu'(\omega)\pi(s|\omega) \left| \frac{\pi'(s) - \pi(s)}{\pi(s)\pi'(s)} \right| \\
&= \sum_{\omega \in \Omega} \mu'(\omega)\pi(s|\omega) \left| \frac{\sum_{\omega' \in \Omega} (\mu'(\omega') - \mu(\omega'))\pi(s|\omega')}{\pi(s)\pi'(s)} \right| \\
&\leq \sum_{\omega \in \Omega} \mu'(\omega)\pi(s|\omega) \frac{\sum_{\omega' \in \Omega} |\mu'(\omega') - \mu(\omega')| \cdot \max_{\omega' \in \Omega} \pi(s|\omega')}{\pi(s)\pi'(s)} \\
&= \|\mu' - \mu\|_1 \sum_{\omega \in \Omega} \frac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)} \frac{\max_{\omega' \in \Omega} \pi(s|\omega')}{\pi(s)} \\
\text{by (23)} \quad &\leq \|\mu' - \mu\|_1 \sum_{\omega \in \Omega} \frac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)} \frac{1}{p_0} = \frac{1}{p_0} \|\mu' - \mu\|_1.
\end{aligned}$$

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:

$$\mathbb{E}_{\xi_a}[v(a,\omega) - v(a',\omega)] = (1-\delta)\mathbb{E}_{\hat{\mu}_a}[v(a,\omega) - v(a',\omega)] + \delta\mathbb{E}_{\eta_a}[v(a,\omega) - v(a',\omega)] \geq \delta D. \tag{25}$$

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 $\chi \in \Delta(\Omega)$:

$$\mu \;=\; (1-y)\xi + y\chi \;=\; \sum_{a\in A}(1-y)\pi^*(a)\xi_a \;+\; y\chi. \tag{26}$$

Lemma 6 (Proposition 1 of [31])**.** *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* (26)*.*

Since (26) 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 5), the posteriors induced by signal a under prior μ and μ' satisfy

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

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

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

Together with (25), we get

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

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,

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

where in the last line we used $y \le \frac{\delta}{p_0}$ from Lemma 6. □

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

9 Experimental Setup

9.1 Instance parameters

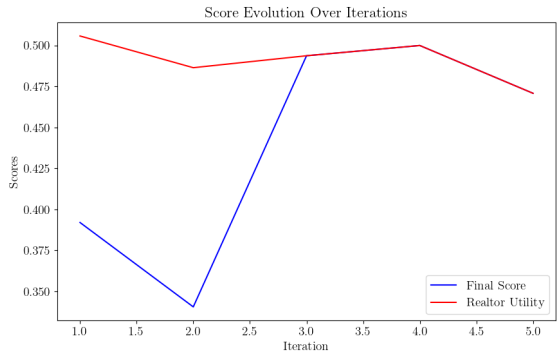
Here we include the detailed setup of the real-estate example we used to verify our framework experimentally. Note that the utilities here are not within the range $[0, 1]$, but can be normalized to be so without loss of generality. Indeed, the utility values highlighted in Section 5 are normalized utilities according to the $[0, 1]$ scale. There are 4 possible states for each instance: $(good, cheap), (good, expensive), (bad, cheap), (bad, expensive)$. 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): $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0.75 & -0.25 & 0.25 & -3 \end{bmatrix}$

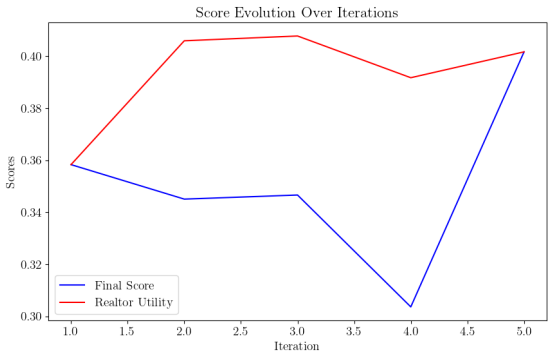
9.2 Intermediate Results

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.

- 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*



(a) Score over iterations on the Henry Instance



(b) Scores over iterations on the Lilly Instance

Figure 3: Scores over multiple iterations of the LLM generating framing. Final score is the product of the utility and the correctness score.

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.

9.3 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
}
```

9.4 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.

9.5 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`