A Knowledge-Sharing Model of Chatbots

Tom Cunningham

Sept 24 2025

Abstract

This note formalizes a common intuitive model of chatbots as sharing existing knowledge. People will consult a chatbot when they encounter a question for which they do not know the answer, but they expect the answer exists somewhere in the public domain (and therefore the chatbot's training set). The model gives strong predictions about chatbot adoption across demographics, occupations, and tasks. Insofar as wages are driven by differences in knowledge then we should expect chatbots to flatten those differences, and cause (1) higher average consumption; (2) lower wages for the highest-paid; (3) lower trade, i.e. lower GDP (because more tasks move into home production).

1 Introduction

This note treats chatbots as lowering the cost of access to existing public knowledge. A chatbot is essentially a database of questions for which the answer already exists in the public domain (and hence in the LLM training data). It implies that people will consult a chatbot when they encounter a question for which (i) they do not know the answer, and (ii) they expect that someone else does know the answer, and has documented it in the public domain.

The model expresses a very common view of LLMs, but has not been stated as explicitly before as far as I am aware. LLM chatbots clearly do many things which would not normally be described as answering questions (drawing pictures, drafting text, editing text, writing code), however the majority share of chatbot queries appear to be asking for information (**?**). The core predictions can be expressed assuming a discrete set of questions, but I also derive predictions from a fuller model which allows for the user and chatbot to interpolate between answers among different related questions.

This model gives us very basic predictions about when chatbots will be used:

1. **Chatbots will be consulted more by novices more than experts in a domain.** This is because novices are more likely to encounter a question that they have not encountered before. A corollary is that chatbots will be used where you are a *relative* novice, e.g. a doctor will be more likely to use a chatbot for a legal question, and a lawyer will be relatively more likely to use a chatbot for a medical question.
2. **Chatbots will be used more in occupations with high returns to experience.** In domains where workers frequently encounter problems that are neither highly repet-

itive, nor entirely idiosyncratic, then we would expect high returns to experience, and also high chatbot use.

3. **Chatbots will be used more in occupations with high mobility.** If the same questions tend to recur across jobs within the same occupation then we expect high mobility between jobs, and also high chatbot use.

4. **Chatbots will be consulted more for well-documented domains.** Chatbots will be used more for questions about technologies that are public than technologies that are proprietary (which do not appear in the training material), and prevalence of questions about a technology should be convex in the popularity of that technology.

The model also gives us basic predictions about the equilibrium effect of chatbots:

1. **Chatbots raise welfare through better decisions.** We should also expect increases in welfare if agents can adjust on other margins, especially reducing the use of other information-sources (books, web search), and through expanding the set of questions that a person attempts to answer.

2. **Chatbots flatten comparative advantage.** If comparative advantage is due to differences in knowledge (the set of questions for which you know the answer) then chatbots will flatten those differences. In a class of trade models where productivity differences are entirely driven by knowledge gaps, this is formally equivalent to "catch up" technology growth, where the catch-up implies: (1) higher average wages; (2) lower wages for the highest-paid; (3) lower trade, i.e. lower GDP (**?**, **?**).

Existing models of AI adoption. We can compare this to a few existing models of AI:

- **AI performs a task autonomously. ?** models automation as allowing capital to perform some task instead of labor. We then expect workers who previously specialized in those tasks to see declines in their wages, e.g. **?**, **?**.

- **AI reduces the time required to perform a task. ?** implicitly assumes that LLMs reduce the time required to complete a task by a fixed proportion.[UNFINISHED]

- **AI supplies knowledge. ?** and **?** [UNFINISHED]. Important point about Garicano/Ide/Talamas model: (1) whether the knowledge required for a problem is known ex ante or ex post; (2) whether you can rank knowledge on a scale.

  Note: **?** says "Our theory suggests that the impact of communication technology during the last years of the 1990s and early 2000s should lead organizations toward increasing centralization, an equalization of routine jobs, and a reduction of self-employment as hierarchies get larger through larger spans of control and more layers ... Our interpretation is that the underlying source of variation is the costs of acquiring and communicating information: improvements in the cost of accessing knowledge, mostly concentrated in the 1980s and early 1990s but present throughout the period, followed by improvements in communication technology in the late 1990s."

Vector model. We also describe a model in which each question is a vector, based on an earlier model of LLMs written for a different purpose, **?**.

Suppose each question $\boldsymbol{q}$ is a vector of binary characteristics, and the true answer is a scalar, a, determined by a set of unobserved weights, $\boldsymbol{w}$, with $a = \boldsymbol{wq}$. The human guesses the answer to the new question by interpolating among previously-seen questions and answers $((\boldsymbol{q}^i, a^i)_{i=1,\ldots,n})$. They can also consult a chatbot which answers questions in the same way, but with a different set of previously-seen questions (i.e. the chatbot's training data). We can then give a crisp closed-form expression for the expected benefit of consulting a chatbot, based on the relationship between the question q, the human's experience Q_1, and the chatbot's experience, Q_2. We expect a chatbot to be consulted when you encounter a question that has components which fall outside the space formed from by your knowledge set (questions for which you know the answer), but which fall inside the space formed by the chatbot's knowledge set.

1.1 Conjectures About Adoption

We can make some conjectures about adoption by occupation and by task:

Occupation	**Predicted ChatGPT use**	**Reason**
software engineer - python	high	problems novel, discrete, similar to those on internet
software engineer - proprietary	low	problems novel, discrete, not similar to those on internet
physician	high	problems novel, discrete, similar to those on internet
contact center worker	low	problems novel, discrete, not similar to those on internet
architect	low	problems novel, not discrete, not text-based
manual worker	low	problems not text-based

Table 1: Conjectures about adoption by occupation

Task	Predicted chatbot use	Reason
Intellectual curiosity	high	novel, discrete, similar to those on the internet
Diagnosing medical problems	high	novel, discrete, similar to those on the internet
Problems with widely-adopted systems (car, house, computer)	high	novel, discrete, similar to those on the internet
Problems with idiosyncratic systems (custom setups)	low	novel, discrete, *not* similar to those on the internet

Table 2: Conjectures about adoption by task

1.2 Additional issues.

1. **Why are LLMs used as advisors, not deputies?** It's notable that LLMs are relatively rarely given autonomy to make a decision without human oversight. ChatGPT is mostly built as an advisor, it can't take actions on your behalf, but it's worth asking why that is. This could be accommodated by this model if most questions have large private components, such that ChatGPT's answer is worse than the human's, but the ChatGPT-augmented answer is better than the human's.

2. **Relationship to time-savings.** This model quantifies the benefit from ChatGPT as the *accuracy* of an answer to a question. Much other literature on LLMs measures the value in time-savings, e.g. most RCTs of LLM augmentation (both in the laboratory and the field), and many self-reports. We could convert the accuracy increase to a time-savings if we say that people can spend more time to increase accuracy.

3. **How does ChatGPT differ from Google?** If we interpret ChatGPT's training set as the content of the public internet then the same model could be applied to a search engine. We could distinguish ChatGPT in two ways: (1) the cost of consulting ChatGPT is significantly lower because it will give an answer immediately, instead of directing the human to another page where they have to skim the text; (2) ChatGPT can *interpolate* questions in its training data, e.g. it will give an answer to your question even if nobody has answered that question before, but the answer can be predicted from the answer to other questions.

4. **Dimensionality of the domain.** We might be able to extend the model to distinguish between two types of dimensionality: (1) surface dimensionality (how many letters it takes to express the question/prompt); (2) the latent dimensionality of the domain. In general the returns to experience will depend on the latent dimensionality of the domain: if the latent dimensionality is low then a few examples is enough to learn the patterns, if the latent dimensionality is high then error continuously decreases with experience (there's a nice closed-form expression for this).

5. **High-dimensional answers.** Our model assumes *scalar* answers. In fact ChatGPT gives high-dimensional outputs. I think we can say some nice things here.

6. **Tacit knowledge.** For some ChatGPT prompts the user can *recognize* a correct answer, but cannot produce a correct answer themselves (AKA a generation-validation gap). E.g. asking for a picture, asking for a poem. The generation-validation asymmetry can be due either to (1) computational difficulty; (2) tacit knowledge.

7. **Routine questions.** For some ChatGPT prompts the user can clearly do the task themselves, without any extra information but it's time-consuming: e.g. doing a simple mathematical operation, alphabetizing a list, typing out boilerplate code. These types of queries don't fit our model so well.

2 Simple Model

Here we define the simplest model: given some question you will consult the chatbot if and only if (1) you do not know the answer to this question; (2) the chatbot does know the answer to this question.

Suppose you are confronted by a question ($q \in \mathcal{Q}$), and you have to supply an answer, $\hat{a} \in \mathcal{R}$.

The chatbot's set of prior questions-observed is $\boldsymbol{Q}_1 \subseteq \mathcal{Q}$, the user's set is $\boldsymbol{Q}_2 \subseteq \mathcal{Q}$, with the composition of both sets public knowledge (i.e. you know *whether* the chatbot knows the answer to each question, without you knowing what that answer is).

It is clear that the user will consult the chatbot if and only if $q \in Q_1$ and $q \not\in Q_2$. In the vector model below, this discrete rule is replaced by a continuous analogue: you benefit from consulting the chatbot when the components of the question that lie outside your own experience overlap with the subspace spanned by the chatbot's experience.

3 Model

The State of the World and Questions. The state of the world is defined by a vector of p unobserved parameters, $\boldsymbol{w} \in \mathbb{R}^p$. A question is a vector of p binary features, $\boldsymbol{q} \in \{-1, 1\}^p$. The true answer to a question $\boldsymbol{q}$ is a scalar a determined by the linear relationship:

$$a = \boldsymbol{q}'\boldsymbol{w} = \sum_{k=1}^{p} q_k w_k$$

Agents and Information. There is a set of agents, indexed by $i \in \mathcal{I}$. Each agent i possesses an information set $\mathcal{D}_i$, which consists of n_i questions they have previously encountered, along with their true answers. We can represent this information as a pair $(\boldsymbol{Q}_i, \boldsymbol{a}_i)$:

- $\boldsymbol{Q}_i$ is an $n_i \times p$ matrix where each row is a question vector. Let the j-th question for

agent i be $\boldsymbol{q}'_{i,j}$, so that:

$$\boldsymbol{Q}_i = \begin{bmatrix} \boldsymbol{q}'_{i,1} \\ \vdots \\ \boldsymbol{q}'_{i,n_i} \end{bmatrix} = \begin{bmatrix} q_{i,1,1} & \cdots & q_{i,1,p} \\ \vdots & \ddots & \vdots \\ q_{i,n_i,1} & \cdots & q_{i,n_i,p} \end{bmatrix}$$

- $\boldsymbol{a}_i$ is an $n_i \times 1$ vector of the corresponding answers. The answers are generated according to the true model:

$$\boldsymbol{a}_i = \boldsymbol{Q}_i \boldsymbol{w}$$

Beliefs. All agents share a common prior belief about the state of the world, assuming the weights $\boldsymbol{w}$ are drawn from a multivariate Gaussian distribution:

$$\boldsymbol{w} \sim N(\boldsymbol{0}, \Sigma)$$

where Σ is a $p \times p$ positive-semidefinite covariance matrix. A common assumption we will use is an isotropic prior, where $\Sigma = \sigma^2 \boldsymbol{I}_p$ for some scalar $\sigma^2 > 0$. This implies that, a priori, the weights are uncorrelated and have equal variance.

Given their information set $\mathcal{D}_i$, agent i forms a posterior belief about $\boldsymbol{w}$. When a new question $\boldsymbol{q}_{\text{new}}$ arises, the agent uses their posterior distribution to form an estimate of the answer, $\hat{a}_{\text{new}} = \boldsymbol{q}'_{\text{new}} \mathbb{E}[\boldsymbol{w} \mid \mathcal{D}_i]$.

Throughout the analysis below we make two simplifying assumptions. First, observations are noiseless: when an agent has seen a question before, they observe its exact true answer, so that $\boldsymbol{a}_i = \boldsymbol{Q}_i \boldsymbol{w}$. Second, the matrices $\boldsymbol{Q}_i \Sigma \boldsymbol{Q}_i^\top$ (and, under an isotropic prior, $\boldsymbol{Q}_i \boldsymbol{Q}_i^\top$) are invertible, so that the posterior expressions are well defined. Both assumptions can be relaxed (for example by allowing noisy answers), at the cost of slightly more involved algebra but with the same basic geometry: what matters is how a new question projects onto the subspaces spanned by past questions.

4 Propositions

Proposition 1 (Posterior over $\boldsymbol{w}$ given $\boldsymbol{Q}$ and $\boldsymbol{a}$)**.** *The agent's posterior mean and variance will be:*

$$\hat{\boldsymbol{w}} = \Sigma \boldsymbol{Q}^\top (\boldsymbol{Q} \Sigma \boldsymbol{Q}^\top)^{-1} \boldsymbol{a}$$
$$\Sigma_{|a} = \Sigma - \Sigma \boldsymbol{Q}^\top (\boldsymbol{Q} \Sigma \boldsymbol{Q}^\top)^{-1} \boldsymbol{Q} \Sigma.$$

Proof. The derivation follows from the standard formula for conditional Gaussian distributions. We begin by defining the joint distribution of the weights $\boldsymbol{w}$ and the answers $\boldsymbol{a}$. The weights and answers are jointly Gaussian:

$$\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{a} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma \boldsymbol{Q}' \\ \boldsymbol{Q}\Sigma & \boldsymbol{Q}\Sigma\boldsymbol{Q}' \end{pmatrix} \right)$$

where the covariance terms are derived as follows:

- $Cov(\boldsymbol{w}, \boldsymbol{w}) = \Sigma$ (prior covariance)
- $Cov(\boldsymbol{a}, \boldsymbol{a}) = Cov(\boldsymbol{Q}\boldsymbol{w}, \boldsymbol{Q}\boldsymbol{w}) = \boldsymbol{Q}Cov(\boldsymbol{w}, \boldsymbol{w})\boldsymbol{Q}' = \boldsymbol{Q}\Sigma\boldsymbol{Q}'$
- $Cov(\boldsymbol{w}, \boldsymbol{a}) = Cov(\boldsymbol{w}, \boldsymbol{Q}\boldsymbol{w}) = Cov(\boldsymbol{w}, \boldsymbol{w})\boldsymbol{Q}' = \Sigma\boldsymbol{Q}'$

The conditional mean $E[\boldsymbol{w}|\boldsymbol{a}]$ is given by the formula:

$$E[\boldsymbol{w}|\boldsymbol{a}] = E[\boldsymbol{w}] + Cov(\boldsymbol{w}, \boldsymbol{a})Var(\boldsymbol{a})^{-1}(\boldsymbol{a} - E[\boldsymbol{a}])$$

Substituting the values from our model ($E[\boldsymbol{w}] = \mathbf{0}$, $E[\boldsymbol{a}] = \mathbf{0}$):

$$\hat{\boldsymbol{w}} = \mathbf{0} + (\Sigma\boldsymbol{Q}')(\boldsymbol{Q}\Sigma\boldsymbol{Q}')^{-1}(\boldsymbol{a} - \mathbf{0}) = \Sigma\boldsymbol{Q}'(\boldsymbol{Q}\Sigma\boldsymbol{Q}')^{-1}\boldsymbol{a}$$

This gives us the posterior mean of the weights. The posterior covariance is given by:

$$Var(\boldsymbol{w}|\boldsymbol{a}) = Var(\boldsymbol{w}) - Cov(\boldsymbol{w}, \boldsymbol{a})Var(\boldsymbol{a})^{-1}Cov(\boldsymbol{a}, \boldsymbol{w}) = \Sigma - \Sigma\boldsymbol{Q}'(\boldsymbol{Q}\Sigma\boldsymbol{Q}')^{-1}\boldsymbol{Q}\Sigma.$$

□

Proposition 2 (Expected error for a given question)**.** *The expected squared error for a new question $\boldsymbol{q}$ is:*

$$\mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] = \boldsymbol{q}'\Sigma_{|a}\boldsymbol{q}$$

For an isotropic prior where $\Sigma = \sigma^2\boldsymbol{I}$, the error is proportional to the squared distance of $\boldsymbol{q}$ from the subspace spanned by the previously seen questions in $\boldsymbol{Q}$:

$$\mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] = \sigma^2\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}\|^2$$

where $\boldsymbol{P}_{\boldsymbol{Q}}$ is the projection matrix onto the row-span of $\boldsymbol{Q}$.

Proof. The prediction error is $\boldsymbol{q}'\boldsymbol{w} - \boldsymbol{q}'\hat{\boldsymbol{w}} = \boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}})$. The expected squared error is the variance of this prediction error.

$$\begin{aligned}\mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] &= \mathbb{E}[\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}})(\boldsymbol{w} - \hat{\boldsymbol{w}})'\boldsymbol{q}] \\ &= \boldsymbol{q}'\mathbb{E}[(\boldsymbol{w} - \hat{\boldsymbol{w}})(\boldsymbol{w} - \hat{\boldsymbol{w}})']\boldsymbol{q} \\ &= \boldsymbol{q}'Var(\boldsymbol{w} \mid \boldsymbol{a})\boldsymbol{q} = \boldsymbol{q}'\Sigma_{|a}\boldsymbol{q}\end{aligned}$$

This proves the first part of the proposition. For the second part, we assume an isotropic prior $\Sigma = \sigma^2\boldsymbol{I}$. Substituting this into the expression for $\Sigma_{|a}$ from Proposition 1:

$$\begin{aligned}\Sigma_{|a} &= \sigma^2\boldsymbol{I} - (\sigma^2\boldsymbol{I})\boldsymbol{Q}'(\boldsymbol{Q}(\sigma^2\boldsymbol{I})\boldsymbol{Q}')^{-1}\boldsymbol{Q}(\sigma^2\boldsymbol{I}) \\ &= \sigma^2\boldsymbol{I} - \sigma^4\boldsymbol{Q}'(\sigma^2\boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{Q} \\ &= \sigma^2\boldsymbol{I} - \sigma^4(\sigma^2)^{-1}\boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{Q} \\ &= \sigma^2(\boldsymbol{I} - \boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{Q})\end{aligned}$$

Let $\boldsymbol{P}_{\boldsymbol{Q}} = \boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{Q}$, which is the projection matrix onto the row space of $\boldsymbol{Q}$. Then $\Sigma_{|a} = \sigma^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})$. The expected squared error is:

$$\mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] = \boldsymbol{q}'\sigma^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q} = \sigma^2\boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}$$

Since $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}}$ is an idempotent projection matrix, $\boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q} = \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q} = \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}\|^2$. Thus,

$$\mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] = \sigma^2\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}\|^2$$

□

Proposition 3 (Error decreases with more independent questions)**.** *The average expected squared error over all possible new questions $\boldsymbol{q}$ decreases linearly with the number of linearly independent questions in the training set $\boldsymbol{Q}$. Specifically, with an isotropic prior $\Sigma = \sigma^2\boldsymbol{I}$, the average error is:*

$$\mathbb{E}_{\boldsymbol{q}}[\mathit{error}(\boldsymbol{q})] = \sigma^2(p - \operatorname{rank}(\boldsymbol{Q}))$$

where the expectation is taken over new questions $\boldsymbol{q}$ with i.i.d. components drawn uniformly from $\{-1, 1\}$.

Proof. The proof proceeds in two steps. First, we write the expression for the error for a given new question $\boldsymbol{q}$. Second, we average this error over the distribution of all possible questions.

1. **Predictive error for a fixed $\boldsymbol{q}$.** From Proposition 2, the expected squared error for a specific new question $\boldsymbol{q}$, given an isotropic prior $\Sigma = \sigma^2 \boldsymbol{I}$, is:

$$\text{error}(\boldsymbol{q}) = \mathbb{E}[(\boldsymbol{q}'(\boldsymbol{w} - \hat{\boldsymbol{w}}))^2] = \sigma^2 \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}$$

   where $\boldsymbol{P}_{\boldsymbol{Q}} = \boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{Q}$ is the projection matrix onto the row-span of $\boldsymbol{Q}$.

2. **Average over random new questions.** We now take the expectation of this error over the distribution of new questions $\boldsymbol{q}$. The components of $\boldsymbol{q}$ are i.i.d. uniform on $\{-1, 1\}$, which implies that $\mathbb{E}[\boldsymbol{q}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{q}\boldsymbol{q}'] = \boldsymbol{I}_p$. The average error is:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}[\text{error}(\boldsymbol{q})] &= \mathbb{E}_{\boldsymbol{q}}[\sigma^2 \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{q}}[\text{tr}(\boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q})] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{q}}[\text{tr}((\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{q}\boldsymbol{q}')] \\
&= \sigma^2 \,\text{tr}((\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\mathbb{E}_{\boldsymbol{q}}[\boldsymbol{q}\boldsymbol{q}']) \\
&= \sigma^2 \,\text{tr}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}}) \\
&= \sigma^2(\text{tr}(\boldsymbol{I}) - \text{tr}(\boldsymbol{P}_{\boldsymbol{Q}}))
\end{aligned}$$

   The trace of the identity matrix is p. The trace of a projection matrix is the dimension of the subspace it projects onto, so $\text{tr}(\boldsymbol{P}_{\boldsymbol{Q}}) = \text{rank}(\boldsymbol{Q})$. Thus, the average error is:

$$\mathbb{E}_{\boldsymbol{q}}[\text{error}(\boldsymbol{q})] = \sigma^2(p - \text{rank}(\boldsymbol{Q}))$$

   Since the rank of $\boldsymbol{Q}$ increases with each linearly independent question added, the average error decreases linearly until $\text{rank}(\boldsymbol{Q}) = p$, at which point it becomes zero.

□

Proposition 4 (Two–stage updating with agents 1 and 2)**.** *Consider two agents who share an isotropic prior $\boldsymbol{w} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$.*

- *Agent 1 observes data $(\boldsymbol{Q}_1, \boldsymbol{a}_1)$ and forms the posterior mean*

$$\hat{\boldsymbol{w}}_1 = \boldsymbol{Q}_1^\top (\boldsymbol{Q}_1 \boldsymbol{Q}_1^\top)^{-1} \boldsymbol{a}_1, \qquad \boldsymbol{P}_1 := \boldsymbol{Q}_1^\top (\boldsymbol{Q}_1 \boldsymbol{Q}_1^\top)^{-1} \boldsymbol{Q}_1.$$

- *Agent 2 observes data $(\boldsymbol{Q}_2, \boldsymbol{a}_2)$ and forms the posterior mean*

$$\hat{\boldsymbol{w}}_2 = \boldsymbol{Q}_2^\top (\boldsymbol{Q}_2 \boldsymbol{Q}_2^\top)^{-1} \boldsymbol{a}_2, \qquad \boldsymbol{P}_2 := \boldsymbol{Q}_2^\top (\boldsymbol{Q}_2 \boldsymbol{Q}_2^\top)^{-1} \boldsymbol{Q}_2.$$

- *A new question $\boldsymbol{q} \in \{-1, 1\}^p$ arrives. Agent 1 announces the estimate $\hat{a}_1 = \boldsymbol{q}' \hat{\boldsymbol{w}}_1$.*

Let

$$\mu_2 = \boldsymbol{q}' \hat{\boldsymbol{w}}_2, \qquad \sigma_2^2 = \sigma^2 \, \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{q}, \tag{1}$$

$$\mu_{2,1} = \boldsymbol{q}' \boldsymbol{P}_1 \hat{\boldsymbol{w}}_2, \qquad \sigma_{21} = \sigma^2 \, \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q}, \tag{2}$$

$$\sigma_{1|2}^2 = \sigma^2 \, \boldsymbol{q}' \boldsymbol{P}_1 (\boldsymbol{I} - \boldsymbol{P}_2) \boldsymbol{P}_1^\top \boldsymbol{q}, \qquad \kappa = \frac{\sigma_{21}}{\sigma_{1|2}^2}. \tag{3}$$

Then, the posterior distribution of the true answer $a = \boldsymbol{q}'\boldsymbol{w}$ for agent 2 before seeing $\hat{a}_1$ is $N(\mu_2, \sigma_2^2)$, and after observing $\hat{a}_1$ it is

$$a \mid \hat{a}_1, \boldsymbol{a}_2 \;\sim\; N\big(\mu_{2|1},\, \sigma_{2|1}^2\big), \qquad \mu_{2|1} = \mu_2 + \kappa\big(\hat{a}_1 - \mu_{2,1}\big), \quad \sigma_{2|1}^2 = \sigma_2^2 - \kappa\,\sigma_{21}.$$

Proof. The estimate of agent 1 is a linear function of the true weights $\boldsymbol{w}$, since $\boldsymbol{a}_1 = \boldsymbol{Q}_1\boldsymbol{w}$, so $\hat{a}_1 = \boldsymbol{q}'\hat{\boldsymbol{w}}_1 = \boldsymbol{q}'\boldsymbol{Q}_1^\top(\boldsymbol{Q}_1\boldsymbol{Q}_1^\top)^{-1}\boldsymbol{a}_1 = \boldsymbol{q}'\boldsymbol{P}_1\boldsymbol{w}$.
Conditioning on agent 2's data, the posterior for $\boldsymbol{w}$ is $N(\hat{\boldsymbol{w}}_2, \Sigma_2)$ with $\Sigma_2 = \sigma^2(\boldsymbol{I} - \boldsymbol{P}_2)$. The pair $(a, \hat{a}_1)$ is therefore jointly Gaussian, since both are linear functions of $\boldsymbol{w}$. Their joint distribution conditional on agent 2's data has:

- $E[a|\boldsymbol{a}_2] = \boldsymbol{q}'\hat{\boldsymbol{w}}_2 = \mu_2$
- $E[\hat{a}_1|\boldsymbol{a}_2] = \boldsymbol{q}'\boldsymbol{P}_1\hat{\boldsymbol{w}}_2 = \mu_{2,1}$
- $\mathrm{Var}(a|\boldsymbol{a}_2) = \boldsymbol{q}'\sigma^2(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{q} = \sigma_2^2$
- $\mathrm{Var}(\hat{a}_1|\boldsymbol{a}_2) = \boldsymbol{q}'\boldsymbol{P}_1\sigma^2(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top\boldsymbol{q} = \sigma_{1|2}^2$
- $\mathrm{Cov}(a, \hat{a}_1|\boldsymbol{a}_2) = \boldsymbol{q}'\sigma^2(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top\boldsymbol{q} = \sigma_{21}$

So the covariance matrix of $(a, \hat{a}_1)$ conditional on agent 2's data is:

$$\begin{pmatrix} \sigma_2^2 & \sigma_{21} \\ \sigma_{21} & \sigma_{1|2}^2 \end{pmatrix}$$

For any joint Gaussian vector, the conditional distribution of the first component given the second is again Gaussian with

$$\mu_{2|1} = \mu_2 + \frac{\sigma_{21}}{\sigma_{1|2}^2}\big(\hat{a}_1 - \mu_{2,1}\big), \qquad \sigma_{2|1}^2 = \sigma_2^2 - \frac{\sigma_{21}^2}{\sigma_{1|2}^2}.$$

Identifying $\kappa = \sigma_{21}/\sigma_{1|2}^2$ yields the stated result. □

Proposition 5 (Conditions for valuable two-stage updating)**.** *In the setting of Proposition 4, consulting agent 1 provides value to agent 2 if and only if:*

$$\boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top\boldsymbol{q} \neq 0$$

When this condition holds:

- *The posterior mean changes: $\mu_{2|1} \neq \mu_2$*
- *The posterior variance decreases: $\sigma_{2|1}^2 < \sigma_2^2$*

When this condition fails, consulting agent 1 provides no additional information: $\mu_{2|1} = \mu_2$ and $\sigma_{2|1}^2 = \sigma_2^2$.

Proof. From Proposition 4, the change in the posterior mean is $\mu_{2|1} - \mu_2 = \kappa(\hat{a}_1 - \mu_{2,1})$, and the change in posterior variance is $\sigma_2^2 - \sigma_{2|1}^2 = \kappa\sigma_{21}$, where $\kappa = \frac{\sigma_{21}}{\sigma_{1|2}^2}$ and $\sigma_{21} = \sigma^2 \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q}$.
If $\sigma_{21} = 0$, then $\kappa = 0$, so $\mu_{2|1} = \mu_2$ and $\sigma_{2|1}^2 = \sigma_2^2$.
If $\sigma_{21} \neq 0$, then $\kappa \neq 0$ (since $\sigma_{1|2}^2 \geq 0$ with equality only when $\boldsymbol{P}_1(\boldsymbol{I} - \boldsymbol{P}_2) = \boldsymbol{0}$, which implies $\sigma_{21} = 0$). In this case, both the mean and variance will generally change unless $\hat{a}_1 = \mu_{2,1}$, which occurs with probability zero.
Therefore, two-stage updating provides value if and only if $\sigma_{21} = \sigma^2 \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q} \neq 0$. □

The intuition behind Proposition 5 is straightforward: **it is worthwhile to consult another agent if and only if the component of the question that you don't understand overlaps with the other agent's area of expertise.**

More precisely:

- $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{q}$ represents the *residual* of the question after projecting it onto agent 2's own experience. This is the part of the question that agent 2 finds novel or unfamiliar.
- $\boldsymbol{P}_1^\top \boldsymbol{q}$ represents the component of the question that lies within agent 1's area of expertise (the row space of their experience matrix $\boldsymbol{Q}_1$).
- The condition $\boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q} \neq 0$ requires that these two components are not orthogonal—there must be some overlap between what agent 2 doesn't know and what agent 1 does know.

Put informally: consultation becomes valuable when there is overlap between agent 2's knowledge gaps and agent 1's strengths.

In the context of the ChatGPT model, this suggests that an AI assistant is most valuable for questions where:

1. The question contains elements that are novel to the human user (large $\|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{q}\|$)
2. These novel elements fall within the AI's training domain (non-zero projection onto the AI's knowledge space)

4.1 Human alone, chatbot alone, or consultation?

We can characterize the conditions for the three possible actions:

Suppose there is some small ε cost to delegating a question to the chatbot, and some slightly larger $\delta > \varepsilon$ cost to consulting the chatbot first, and then giving a human-adjusted answer. Then we can characterize the conditions for the three regions:

$$S_{1\to 2} = \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q} \quad \text{(chatbot has info the human lacks)}$$
$$S_{2\to 1} = \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{P}_2^\top \boldsymbol{q} \quad \text{(human has info the chatbot lacks)}$$

Policy	Condition	Interpretation
Human alone	$S_{1\to 2} = 0$ and $S_{2\to 1} \neq 0$	Chatbot has no relevant information beyond what the human already knows, while the human has some information the chatbot lacks.
Chatbot alone	$S_{2\to 1} = 0$ and $S_{1\to 2} \neq 0$	Human has no relevant information beyond what the chatbot already knows, while the chatbot has some information the human lacks.
Consultation	$S_{1\to 2} \neq 0$ and $S_{2\to 1} \neq 0$	Each has relevant information that the other lacks, so combining them is worthwhile despite the higher consultation cost $\delta > \varepsilon$.

Table 3: Information-structure conditions for preferring human alone, chatbot alone, or consultation.