A Model of ChatGPT

Tom Cunningham

2025-06-14

Abstract

We present a simple Bayesian model of how large language models such as ChatGPT and humans answer factual questions. Each question is represented as a high-dimensional binary vector and its true answer is a linear form in an unknown weight vector. Agents interpolate from previously-answered questions drawn from private (human) or public (AI) data. We derive closed-form expressions for posterior beliefs, prediction error, the value of additional examples, and the incremental value of consulting ChatGPT before making a decision. The framework yields testable implications about which users will use ChatGPT for which tasks, specifically ChatGPT will be used when encountering a question logically distinct from the questions in a user's own prior experience, and that novel component is represented in the space of questions in the public dataset (e.g on the internet).

I first describe a simple model of an agent answering a question, so we can formalize the sense in which the answer will be better when the question is more similar to questions that the agent already knows the answer to. We can then also talk about what types of questions it will be worthwhile asking for advice from another agent (ChatGPT). This gives us a simple model of which types of questions will be worthwhile asking ChatGPT demand, based on the nature of the question, the prior experience of the agent, and the prior experience of ChatGPT (AKA its training data).

Formally, each $\boldsymbol{q}$ is a vector of binary attributes, with a true scalar answer a, that is linear in the attributes. An agent's estimate of the answer will be an interpolation based on previously-seen questions and answers $(\boldsymbol{q}^i, a^i)_{i=1,\ldots,n}$. This framework extends an earlier model I developed for a different purpose.

This framework can be interpreted as a model of an agent, "the user," who must provide an estimate for the answer to a question $\boldsymbol{q}$ and can choose whether to consult ChatGPT. The key statistics of the setting will determine the incremental value of consulting ChatGPT:

1. **The dimensionality of the question (p).** A higher-dimensional problem may be more costly to enter into the AI, but it also increases the potential benefit.

2. **The public information set.** These are the training questions that the AI has observed, which we can conceptualize as the corpus of public knowledge (e.g., the internet).

3. **The private information set.** These are the questions that the user has personally encountered and for which they have observed the true answer.

A user will consult the AI if and only if the expected improvement in their answer exceeds the associated cost. The model predicts that an AI will be most useful for questions with components that are novel to the user but contained within the AI's public training data.

This leads to several corollaries:

1. An AI will not be used for questions the user has encountered before.
2. An AI is more likely to be used for domains with higher *latent* dimensionality (p).
3. An AI is more likely to be used for domains with lower *surface* dimensionality, as this reduces the cost of specifying the question.
4. An AI is more likely to be used by humans with less experience in a domain (i.e., smaller n_{private}).

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

Occupation	**Predicted ChatGPT use**	**Reason**
software engineer	high	many novel discrete problems, similar to those on public internet
software engineer - idiosyncratic language	low	many novel discrete problems, not similar to those on public internet
physician	high	many novel discrete problems, similar to those on public internet
contact center worker	low	novel problems, but not similar to those on the internet
architect	low	novel problems, not discrete, not text-based
manual worker	low	not not text-based

Table 1: Conjectures about adoption by occupation

Additional things to add:

1. **High-dimensional answers.** Our model assumes *scalar* answers. In fact ChatGPT gives high-dimensional outputs. I think we can say some nice things here.
2. **Tacit knowledge.** ChatGPT will be more likely to be used for domains where humans have tacit knowledge.

1 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:

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

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

Table 2: Conjectures about adoption by task

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, $\hat{a}_{\text{new}} = \boldsymbol{q}'_{\text{new}} \mathbb{E}[\boldsymbol{w} \mid \mathcal{D}_i]$.

2 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 $\boldsymbol{Q}$:

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

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

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

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

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

□

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:*

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

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*

$$\begin{aligned}
\mu_2 &= \boldsymbol{q}'\hat{\boldsymbol{w}}_2, & \sigma_2^2 &= \sigma^2\, \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{q}, && (1) \\
\mu_{2,1} &= \boldsymbol{q}'\boldsymbol{P}_1 \hat{\boldsymbol{w}}_2, & \sigma_{21} &= \sigma^2\, \boldsymbol{q}'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q}, && (2) \\
\sigma_{1|2}^2 &= \sigma^2\, \boldsymbol{q}'\boldsymbol{P}_1(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{P}_1^\top \boldsymbol{q}, & \kappa &= \frac{\sigma_{21}}{\sigma_{1|2}^2}. && (3)
\end{aligned}$$

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.

This formalizes the intuitive notion that collaboration is valuable when agents have *complementary* rather than identical or completely unrelated knowledge. If agent 1's expertise is orthogonal to the unfamiliar aspects of the question for agent 2, then agent 1's opinion provides no useful information. Conversely, if there is overlap between agent 2's knowledge gaps and agent 1's strengths, then consultation becomes valuable.

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)
3. The user's existing knowledge and the AI's training data are complementary rather than redundant

3 References