A Model of ChatGPT

Tom Cunningham

June 14, 2025

This paper develops a simple model of human and AI ability to answer questions. Each question $\boldsymbol{q}$ is a high-dimensional vector, with a true scalar answer a. An agent's estimate of the answer is an interpolation based on previously-seen questions and answers $(\boldsymbol{q}^i, a^i)_{i=1,\ldots,n}$. This framework extends an https://tecunningham.github.io/posts/2023-09-05-model-of-ai-imitation.htmlearlier model developed for a different purpose.

The model yields several implications:

1. **The quality of an answer to a new question depends on its distance from the training set.** For a new question $\boldsymbol{q}$, the expected error is a function of the distance between $\boldsymbol{q}$ and the training set $\boldsymbol{Q}$.

2. **The quality of answers increases with the size of the training set.** The expected error decreases linearly with the number of linearly-independent examples in the training set.

3. **The value of advice from another agent depends on the distance between their training sets.**

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 an AI model like ChatGPT. The key components of the model are:

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:

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 text-based

We can make some conjectures about adoption by task:

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 on the internet

Additional things to add:

1. **High-dimensional answers.** Our model assumes *scalar* answers. In fact ChatGPT gives high-dimensional outputs. We discuss extensions below.

2. **Tacit knowledge.** ChatGPT will be more likely to be used for domains where humans have tacit knowledge.

1 Model

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. Write this information as $(\boldsymbol{Q}_i, \boldsymbol{a}_i)$ with

$$\boldsymbol{Q}_i = \begin{bmatrix} \boldsymbol{q}'_{i,1} \\ \vdots \\ \boldsymbol{q}'_{i,n_i} \end{bmatrix}, \qquad \boldsymbol{a}_i = \boldsymbol{Q}_i \boldsymbol{w}.$$

Beliefs

All agents share a common prior belief that $\boldsymbol{w}$ is drawn from

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

A common assumption is an isotropic prior, $\Sigma = \sigma^2 \boldsymbol{I}_p$. Given their information, agent i forms a posterior for $\boldsymbol{w}$ and hence an estimate for a new question $\hat{a}_{\text{new}} = \boldsymbol{q}'_{\text{new}} \mathbb{E}[\boldsymbol{w} \mid \mathcal{D}_i]$.

2 Propositions

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

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

Proof. See Appendix for a full derivation. □

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 $\Sigma = \sigma^2 \boldsymbol{I}$*, this simplifies to*

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

where $\boldsymbol{P}_{\boldsymbol{Q}}$ *projects onto the row-span of* $\boldsymbol{Q}$*.*

Proposition 3 (Error decreases with more independent questions)**.** *With an isotropic prior and new questions uniformly random on* $\{-1, 1\}^p$,

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

Proposition 4 (Posterior in two-stage estimation)**.** *Consider a computer (C) with data* $(\boldsymbol{Q}_C, \boldsymbol{a}_C)$ *and a human (H) with* $(\boldsymbol{Q}_H, \boldsymbol{a}_H)$*. The human observes the computer's estimate* $\hat{a}_C$ *for a new question* $\boldsymbol{q}$ *and updates their belief. Various informational assumptions about the computer yield weights that nest oracle trust and total skepticism; see text for details.*