A Model of ChatGPT

Tom Cunningham

2025-06-14

This note gives a simple model of human and AI ability to answer questions. Each question q is a high-dimensional vector of bits, with a true scalar answer a. Each agent tries to estimate the answer by interpolating among previous questions $(q^i, a^i)_{i=1,...,n}$ that they have encountered. This is an extension of an earlier model I wrote for a different purpose.

The model gives a series of general implications:

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

2. **The quality of answers will increase with the size of the training set.** Precisely, the expected error will decrease linearly with the number of linearly-independent examples in the training set.

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

Implications for ChatGPT

Interpreted as a model of ChatGPT, this gives a set of predictions. We can interpret this model as one of ChatGPT use, there are three basic ingredients:

1. p represents the dimensionality

2. Public questions – these are the training questions that ChatGPT has observed. Roughly speaking we can say these consist of all the questions and answers on the public internet. ChatGPT's full training process is of course more complicated, we discuss below.
3. Private questions – these are the questions that the human has themselves encountered (and observed the answer for).

A human will invoke ChatGPT if and only if the expected improvement in the answer exceeds some cost.

ChatGPT will be adopted for questions which contain novel components (outside of private question space)

Corollaries:

1. ChatGPT will not be used for questions that the user has encountered before.
2. ChatGPT will be more likely to be used for domains with higher *latent* dimensionality (p).
3. ChatGPT will be more likely be used for domains with lower surface dimensionality – because it takes more time to specify the question.
3. ChatGPT will be more likely to be used for humans with less experience in a domain (n_{private}).

occupation	dimensionality
junior contact center worker	
senior contact center worker	
physician	

task/question

Additional things we'd like to add:

1. ChatGPT will be more likely to be used for domains where humans have tacit knowledge.

Model

The world is characterized by a set of p weights, w. All agents have Gaussian priors over those weights:

$$w \sim N(0, \sigma^2 I_p)$$

Each agent observes a matrix Q of questions, each question has p binary parameters:

$$\begin{aligned}
Q &\in \{-1, 1\}^{n \times p} && (n \text{ questions, each has } p \text{ binary parameters}) \\
w &\sim N(0, \Sigma) && (p \times 1 \text{ vector of true parameters of the world}) \\
\underset{n \times 1}{\underset{\sim}{a}} &= \underset{n \times p}{\underset{\sim}{Q}} \; \underset{p \times 1}{\underset{\sim}{w}} && (\text{answers provided by the world})
\end{aligned}$$

We can also write this out in matrix form:

$$\begin{aligned}
Q &= \begin{bmatrix} q_1^1 & \dots & q_p^1 \\ & \ddots & \\ q_1^n & \dots & q_p^n \end{bmatrix} && (\text{matrix of } n \text{ questions, each with } p \text{ parameters}) \\
w' &= \begin{bmatrix} w_1 \dots w_p \end{bmatrix} && (\text{vector of } p \text{ unobserved weights}) \\
a &= \begin{bmatrix} a^1 \\ \vdots \\ a^n \end{bmatrix} = \begin{bmatrix} q_1^1 w_1 + \dots q_p^1 w_p \\ \vdots \\ q_1^n w_1 + \dots q_p^n w_p \end{bmatrix} && (\text{vector of } n \text{ observed answers})
\end{aligned}$$

Training Data. Each agent i has access to a set of observations, or "training data," which consists of a set of questions Q_i and their corresponding answers a_i.

$$\mathcal{D}_i = \{(Q_i, a_i)\}$$

Propositions

Proposition 1 (Posterior for a given question). The agent's posterior mean and variance will be:

$$\begin{aligned}
\hat{w} &= \Sigma Q^\top (Q \Sigma Q^\top)^{-1} a \\
\Sigma_{|a} &= \Sigma - \Sigma Q^\top (Q \Sigma Q^\top)^{-1} Q \Sigma
\end{aligned}$$