Consider a question whose correct answer is the n-digit sequence

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}, \tag{1}$$

where a_j is the correct digit for position j. We suppose the AI can compute the matrix

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,n} \end{bmatrix},$$

where $p_{i,j}$ is the probability that the student wrote the digit i in position j. Given the probability matrix $\mathbf{P}$, the AI guesses that the student wrote the answer

$$\mathbf{G} = \begin{bmatrix} g_1 & g_2 & \cdots & g_n \end{bmatrix},$$

where each g_j is the digit corresponding to the maximum probability in column j of $\mathbf{P}$. That is, each digit of $\mathbf{G}$ is the most probable digit for its position. Thus, the AI assesses a correct answer if and only if $\mathbf{A} = \mathbf{G}$.

In grading a single question, the AI performs two primary tasks: (i) *reading* the answer and (ii) *evaluating* the answer. Accordingly, there are at least two measures of the AI's confidence in its results:

- *Reading confidence*, which is the probability $P(R)$ that the answer read by the AI matches the answer written by the student; and
- *Evaluating confidence*, which is the probability $P(E)$ that the final classification of the answer (as correct or incorrect) is correct.

The more intuitive of these measures, and the one more directly useful to the grader, is the evaluating confidence $P(E)$. TODO: We should make sure this is what we want.

If we assume that the AI classifies each digit independently, we may compute these measures in terms of $\mathbf{A}$, $\mathbf{P}$, and $\mathbf{G}$. Indeed, $P(R)$ is the probability that each individual digit was read correctly, and thus

$$P(R) = p_{g_1,1} p_{g_2,2} \cdots p_{g_n,n} = \prod_{j=1}^{n} p_{g_j,j}.$$

Similarly, the probablity $P(C)$ that the student's entire answer is correct is given by

$$P(C) = p_{a_1,1} p_{a_2,2} \cdots p_{a_n,n} = \prod_{j=1}^{n} p_{a_j,j}.$$

Now, if $\mathbf{A} = \mathbf{G}$, the AI determines that the answer was correct, and in this case the evaluating confidence is simply $P(C)$. Otherwise, the AI determines that the answer was incorrect, and in this case the evaluating confidence is the

probability $1-P(C)$ that the answer was *not* correct. The evaluating confidence is therefore given by

$$P(E) = \begin{cases} P(C) & \text{if } \mathbf{A} = \mathbf{G} \\ 1 - P(C) & \text{otherwise.} \end{cases}$$