Supplementary Materials for

Productivity gains of artificial intelligence outweigh energy costs

Antoine Mandel, Giovanni D'Elia

June 2026

Contents

1 Comparison between our ratings and those of Elonundou et al., 2024

Here we present the degree of agreement between the task-level exposure ratings I obtained with the Open AI API of GPT-5, with those obtained by [1] using GPT-4, available in `https://github.com/openai/GPTs-are-GPTs/blob/main/data`.

Table 1: Share of matching ratings

Originally E0	Originally E1	Originally E2	Total
0.94	0.52	0.57	0.73

As it can be seen in table 1, the ratings obtained by me with GPT-5 and those obtained by the authors with GPT-4 coincide only for 72% of tasks.

A statistics frequently used to test accordance between two different classifications is Cohen' s Kappa ([2]). Taking two classifications made of the same categories, Cohen' s Kappa is computed comparing the frequency of matching cases, with the probability of matching if the two classification were independent. The Cohen's Kappa between the original classification and the one obtained through my replication exercise is equal to 0.555. This means that the level of accordance between the two classification is about halfway between a

perfect matching and a situation in which the two are completely independent. According to [3], this level of Cohen's Kappa could be seen as representing a "moderate" accordance.

Figure 1 shows graphically the comparison between the two classifications with a Sankey plot on the left and a confusion matrix on the right. As it can be seen, the category remaining more integer is $E0$ but still, with a considerable portion being classified as $E2$ in the replicated classifications. Being conceptually $E2$ the medium category between $E0$ and $E1$, it makes sense that it is the category which creates more fuzziness in the correspondence. Indeed, both for tasks that were originally $E0$, and for those who were originally $E1$, the majority of switching happens for tasks becoming $E2$ in the new classification.

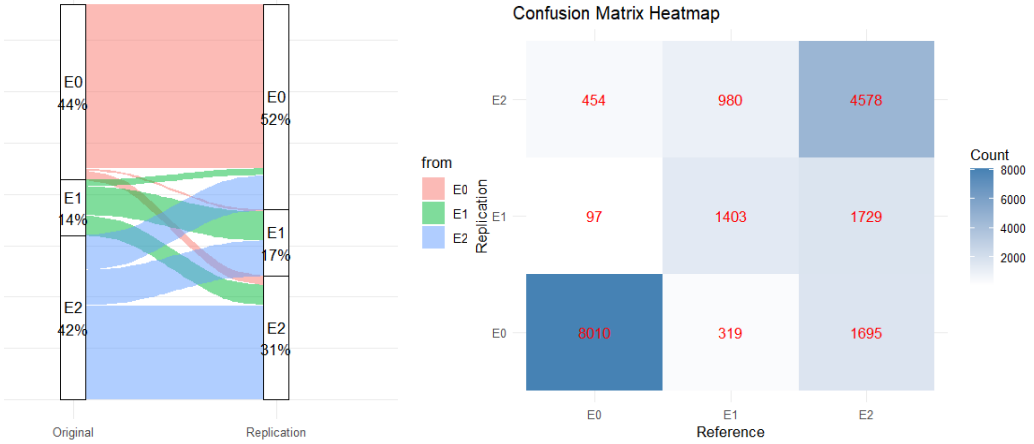

Figure 1: Comparison of classifications, full sample

The fact that category $E1$ is the one most responsible for discrepancies is evident in Table 2. Sensitivity measures the share of tasks which were originally of one specific category being confirmed in that category. Only 52% of original $E1$ tasks were confirmed $E1$ in my replication.$E1$ remains however highly specific, meaning that most of tasks which were not in that category in the original classification, remained "not $E1$" also in my replication. On the contrary it has a very low Positive Predictive Value: when my replication labelled a task $E1$, the task was originally $E1$ only in 43% of cases. The statistics for the other two categories appear instead much better.

The discrepancies between my ratings and those reported in [1] may stem from several factors. First, this study relies on GPT-5, a substantially more advanced model than the GPT-4 version employed in the original analysis. Second, the evaluations are conducted approximately two years after the publication of the original paper, during a period characterized by rapid improvements in AI capabilities. However, if these discrepancies were driven solely by technological progress, one would expect a relatively monotonic pattern of transitions, with tasks previously classified as $E0$ moving toward $E1$ or $E2$, and tasks initially

Table 2: Statistics by class from confusion matrix

	Class: E0	Class: E1	Class: E2
Sensitivity	0.94	0.52	0.57
Specificity	0.81	0.89	0.87
Pos Pred Value	0.80	0.43	0.76
Neg Pred Value	0.94	0.92	0.74
Prevalence	0.44	0.14	0.42
Detection Rate	0.42	0.07	0.24
Detection Prevalence	0.52	0.17	0.31
Balanced Accuracy	0.87	0.70	0.72

labelled as $E2$ shifting toward $E1$ as standalone LLM capabilities improve. The Sankey diagram instead reveals a more complex reclassification structure. In particular, a considerable share of tasks originally categorized as $E1$ are now classified as $E2$, while a non-negligible fraction of tasks initially labelled as $E2$ are reclassified as $E0$.

Most likely, these discrepancies are instead attributable to the substantial variance associated with LLM-based estimations already documented by [4]. The validity of the present analysis would therefore benefit considerably from robustness checks based on alternative frontier models performing the same exposure-rating and time-savings evaluations, as explicitly recommended by the authors.

2 GPT-5 prompt to produce ratings and time savings estimates

Classify the task "i" performed within the occupation "j" according to the taxonomy reported below:

{Original exposure-classification prompt from Eloundou et al., 2024}

You should also provide the percentage time savings if it were executed having access to the LLM only (if it is E1 or E0) or also to the additional software as specified in the Taxonomy (if it is E2). Note that for tasks not labelled E0, this percentage time savings should not be less then 0.50, otherwise the Task should be labelled E0. Of course, if the task execution does not benefit at all from LLMM usage, time savings should be 0.00

Return only: (i) the exposure category and (ii) the estimated percentage time savings.

The original exposure-classification prompt from [1] is available in the Sup-

plementary Materials of that article, downloadable at `https://www.science.org/doi/10.1126/science.adj0998`

References

[1] Tyna Eloundou et al. "GPTs are GPTs: Labor market impact potential of LLMs". In: *Science* 384.6702 (2024), pp. 1306–1308.

[2] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[3] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[4] Michelle Yin, Hoa Vu, and Claudia Persico. *How (un) Stable Are LLM Occupational Exposure Scores? Evidence from Multi-Model Replication*. Tech. rep. National Bureau of Economic Research, 2026.