**Supplementary material for ’GPT API can function as hihgly reliable second screeners’**

FIGURE S1. *The impact of inclusion probability thresholds on performance metrics.*

|  |
| --- |
| A |
| B |

*Note*: The inclusion probability threshold on the x-axis indicates the required number of times the given title and abstract record needed to be included over the 10 repeated requests in order to be coded as relevant. For example an inclusion probability threshold of 0.1 means that the record was coded as relevant if the GPT model included it in 1 or more out of 10 requests.

TEXTBOX S1: *Multi-prompts used for screening in Experiment 3*

PROMPT 1: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*We want to include studies with quantitative measures. For each study, we would like you to assess:*

*1) Does the study report quantitative measures?”*

PROMPT 2: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*Only investigations performed in a school setting on children or students (ages 4-18 years old) are relevant for this review. This means that experiments performed in laboratories must be excluded, because we are only interested in real school settings and educational systems. For each study, we would like you to assess:*

*1) Does the intervention take place within a school setting?”*

PROMPT 3: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*We only want to include studies that investigate children or students attending either primary or secondary school, this means from kindergarten until grade 12. In other words, we are looking for studies where the participants are students 4-18 years old. For each study, we would like you to assess:*

*1) Are the participants in the study children or students attending either primary or secondary school, this means from kindergarten until grade 12.”*

PROMPT 4: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*The study must entail testing students or children. The testing can be standardized and non-standardized tests as well as formative assessments and summative tests, and high-stakes and low-stakes exams. This also include repeated testing, interim assessment testing, class quizzes, multiple choice testing, progress monitoring assessments or measures, curriculum-based measurement or assessments, retrieval practice measures or assessments, etc. For each study, we would like you to assess:*

*1) Does the study report on tests or testing of students or children?”*

TEXTBOX S1 (Continued)

PROMPT 5: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*We like to include randomized controlled trials (RCT), field experiments, quasi-experimental studies (QES), or observational studies, which use a control/comparison research design to examine effects. This means that the study must compare at least two groups of students or children. Such studies can have many labels and the different designs can have different notations. The most common sub-categories of randomised controlled trials and quasi-experimental studies are: individual randomised assignment, cluster randomised assignment, stratified/blocked random assignment, pseudo-randomisation, matching cohort studies, difference-in-differences, regression-discontinuity designs, instrumental variable designs, propensity score matching, case-control studies, etc. Studies employing a within-subject design are also eligible for inclusion. For each study, we would like you to assess:*

*1) Is the study a randomized controlled trial (RCT), a field experiment, a quasi-experimental study, an observational study, or a study employing a within-subject design?”*

PROMPT 6: *“We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*In the review, we would like to include studies that measure students' academic achievement. In this review, we do not restrict measures of academic achievement to specific subjects. For each study, we would like you to assess:*

*1) Does the study report on measures of academic achievement or academic skills?”*

Textbox S1 presents all the prompts we engineered and used to conduct the third classifier experiment. When added to the AIscreenR, each of the above six prompts was pasted together with the text present in Textbox 2 in the main paper.

TEXTBOX S2: *Single-prompts used for screening in Experiment 3*

*We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.*

*We want to include studies with quantitative measures. Only investigations performed in a school setting on children or students (ages 4-18 years old) are relevant for this review. This means that experiments performed in laboratories must be excluded, because we are only interested in real school settings and educational systems. We only want to include studies that investigate children or students attending either primary or secondary school, this means from kindergarten until grade 12. In other words, we are looking for studies where the participants are students 4-18 years old. The study must entail testing students or children. The testing can be standardized and non-standardized tests as well as formative assessments and summative tests, and high-stakes and low-stakes exams. This also include repeated testing, interim assessment testing, class quizzes, multiple choice testing, progress monitoring assessments or measures, curriculum-based measurement or assessments, retrieval practice measures or assessments, etc. We like to include randomized controlled trials (RCT), fields experiments, quasi-experimental studies (QES), or observational studies, which use a control/comparison research design to examine effects. This means that the study must compare at least two groups of students or children. Such studies can have many labels and the different designs can have different notations. The most common sub-categories of randomised controlled trials and quasi-experimental studies are: individual randomised assignment, cluster randomised assignment, stratified/blocked random assignment, pseudo-randomisation, matching cohort studies, difference-in-differences, regression-discontinuity designs, instrumental variable designs, propensity score matching, case-control studies, etc. Studies employing a within-subject design are also eligible for inclusion. In the review, we would like to include studies that measures students' academic achievement. In this review, we do not restrict measures of academic achievement to specific subjects.*

*For each study, we would like you to assess:*

*1) Does the study report quantitative measures?*

*2) Does the intervention take place within a school setting?*

*3) Are the participants in the study children or students attending either primary or secondary school, this means from kindergarten until grade 12.*

*4) Does the study report on tests or testing of students or children?*

*5) Is the study a randomized controlled trial (RCT), a field experiment, a quasi-experimental study, an observational study, or a study employing a within-subject design?*

*6) Does the study report on measures of academic achievement or academic skills?*

FIGURE S2: *Hierarchical screening*

Prompt 1:   
With inclusion/exclusion criertion 1

*If included, the record moves to next prompt, otherwise the record is excluded*

Prompt 2:   
With inclusion/exclusion criertion 2

*If included, the record moves to next prompt, otherwise the record is excluded*

Prompt 3:   
With inclusion/exclusion criertion 3

*Continue as long as necessary*

Prompt x:   
With inclusion/exclusion criertion x