

Supplemental Material 2

Prompting Protocol and In-Context Instructional Material Used for Large Language Model Evaluation

General description

ChatGPT and Gemini were evaluated under two conditions: pre-instruction and post-instruction. No model weights were modified, and no fine-tuning or retraining was performed. The post-instruction condition instead involved in-context prompting based on ERS/ATS 2022 criteria and 200 independently and randomly selected spirometry cases that had previously been interpreted by pulmonologists.

The 200-case instructional set was randomly selected from a separate archival pool and was independent of the main evaluation cohort. For these instructional cases, spirometry images, numerical spirometric parameters, and pulmonologist-derived reference classifications were provided as contextual examples. Expert labels were provided only for the instructional cases and were not disclosed to the models for cases belonging to the evaluation cohort.

Identical instructional material was used for both models.

Before model evaluation, any system-generated interpretive text present on the original spirometry reports was removed. The models therefore received the anonymized numerical spirometric parameters and graphical spirometric information without pre-existing automated diagnostic labels.

Each evaluation case was assessed in a separate, newly initiated chat session, and pre-instruction and post-instruction assessments were performed in separate sessions.

A. Pre-instruction evaluation prompt

Examine the uploaded JPEG images of the pulmonary function tests (spirometry reports).

For each spirometry image, extract the following information and present it in a structured table:

- Age
- Height
- Weight
- Body mass index (BMI)
- FVC
- FVC percent predicted (FVC%)
- FVC Z-score
- FEV₁
- FEV₁ percent predicted (FEV₁%)

- FEV₁ Z-score
- FEV₁/FVC
- FEV₁/FVC Z-score
- PEF

In addition to the numerical parameters, evaluate all available spirometric information contained in the report, including flow-volume curves and other graphical information.

Interpret the spirometry according to the international ERS/ATS 2022 interpretive standards.

For each spirometry test, report the following:

1. Is the test interpretable? Yes / No
2. If interpretable, classify the ventilatory pattern as: Normal / Obstructive / Restrictive / Mixed

Present all extracted variables and interpretation results in a structured table.

B. In-context instructional phase

Before the post-instruction evaluation, the models were provided with 200 independent spirometry cases that had previously been interpreted by pulmonologists.

For each instructional case, the following information was provided:

- The spirometry report as a JPEG image
- Age
- Height
- Weight
- BMI
- FVC
- FVC% predicted
- FVC Z-score
- FEV₁
- FEV₁% predicted
- FEV₁ Z-score
- FEV₁/FVC
- FEV₁/FVC Z-score
- PEF
- The pulmonologists' reference assessment of test interpretability
- The pulmonologists' reference classification of the ventilatory pattern

The models were instructed to treat the pulmonologist assessments in the instructional cases as reference labels.

For ventilatory pattern classification, Z-scores and the expert reference classifications were prioritized. Classification based on FEV₁/FVC, FEV₁% predicted, and FVC% predicted was used only as a supplementary consistency check.

For the instructional cases:

- Pulmonologist assessment was considered the reference label for test interpretability.
- The pulmonologist classification incorporating Z-score information was considered the reference label for ventilatory pattern.
- If a test had been classified as non-interpretable by the pulmonologists, the ventilatory-pattern field was left blank.

This procedure did not modify the underlying model weights and constituted example-based in-context prompting rather than model training or fine-tuning.

C. Post-instruction evaluation prompt

Evaluate the provided JPEG images of the pulmonary function tests using the previously supplied ERS/ATS 2022-based instructions and the example assessments contained in the instructional cases as contextual guidance.

For each spirometry report, assess all available numerical and graphical spirometric information.

Present the following information for each image in a single structured table:

- Age
- Height
- Weight
- BMI
- FVC
- FVC% predicted
- FVC Z-score
- FEV₁
- FEV₁% predicted
- FEV₁ Z-score
- FEV₁/FVC
- FEV₁/FVC Z-score
- PEF
- Test interpretability (Yes/No)
- Ventilatory pattern (Normal/Obstructive/Restrictive/Mixed)

For ventilatory pattern classification, consider the Z-scores and the ERS/ATS 2022 interpretive criteria. FEV₁/FVC, FEV₁% predicted, and FVC% predicted should be used as supplementary interpretive parameters.

If a spirometry test is classified as non-interpretable, leave the ventilatory-pattern field blank.

Present the results in a single table following the order of the uploaded files.