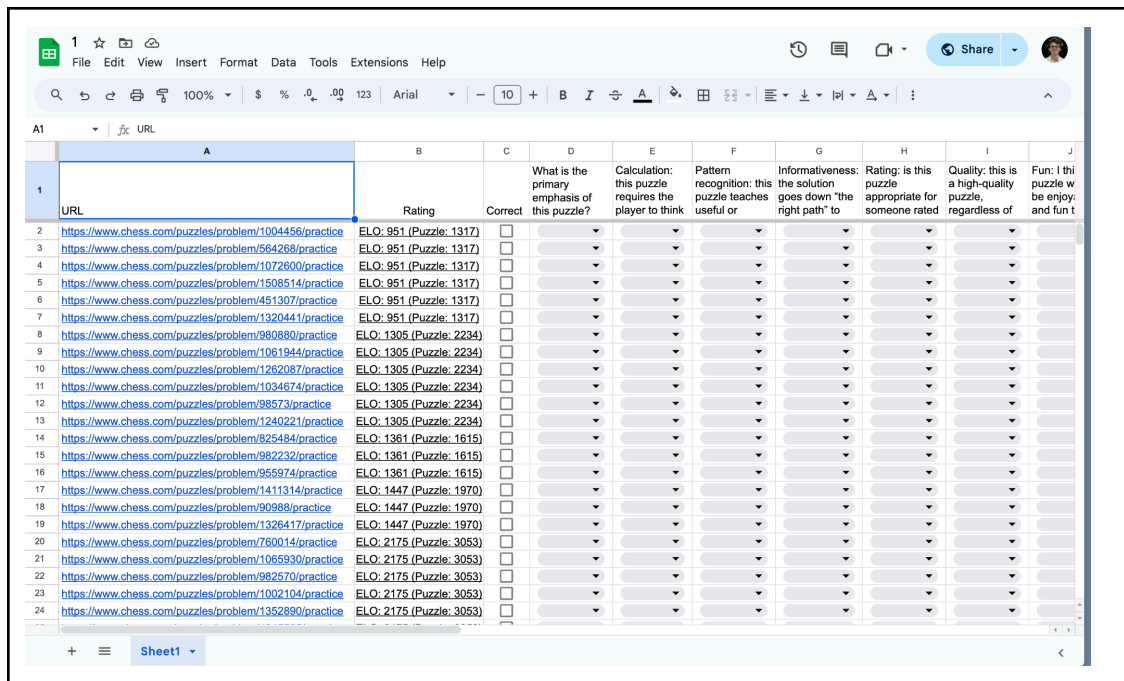# Annotator Guidelines

Thank you for agreeing to participate in our research study! Please read this entire document before beginning the annotation task.

We built a chess puzzle recommendation system and are measuring the extent to which puzzles recommended by our system are better or worse than those recommended by existing tactics trainers on websites like Chess.com. *In the following task, you will be presented with a series of chess puzzles and asked to judge their quality along a number of dimensions*. You should have received a spreadsheet of puzzle URLs like the one below:

Note: you will need a Chess.com Premium membership (Gold tier or higher) to participate in this study – if you do not currently have one, we will reimburse one month of Gold membership.

You will be reviewing puzzles recommended to users at a variety of rating levels. Before annotating the quality of each puzzle, please attempt to solve it first, and mark in the spreadsheet whether you got it right or wrong. Then you will score each puzzle on its overall quality, as well as six other criteria, such as whether the puzzle is rating-appropriate. The following pages will provide descriptions of each of these rating criteria. Finally, you will be asked to provide a written description of the puzzle and a justification for your ratings (minimum length: one sentence).

We anticipate it will take about 1.5 hours to solve and annotate 30 puzzles, including time to read these annotation guidelines.

Calculation or Pattern Recognition?

Some puzzles are primarily designed to train a player's ability to recognize common tactics or motifs, such as a smothered mate; other puzzles focus on testing a player's calculation skills. For each puzzle, please first describe whether it is primarily intended to focus on calculation, pattern recognition, both, or neither.

Based on your answer, please fill out the calculation and pattern recognition subproblems (described below). If you mark a puzzle as primarily focused on calculation, then please mark N/A on the pattern recognition subproblem (and vice versa).

Calculation

The first criteria measures the extent to which a puzzle trains a player's ability to calculate and is graded on a 4-point scale. Some puzzles require the player to reason about very long sequences of moves, or sequences with high branching factors, and would receive a high rating under this criteria. Other puzzles are either very short or can be solved by simply placing the only forcing move at each step of the puzzle, without seeing the full solution ahead of time. For instance, in the below position, Black is forced to take on g6 and play Ng7, which solves the puzzle.

(This puzzle would be a 1)

Pattern Recognition:

The second criteria measures the patterns or tactical motifs that a puzzle is testing, and is graded on a 3-point scale. Some puzzle motifs can be repetitive or feel 'overdone' -- e.g., a classic smothered mate. Others can test more unusual and unordinary patterns. The highest score, a '3', should be given to puzzles with such interesting motifs; alternatively, puzzles may also be ranked as a '3' if the motifs are banal, but the puzzle puts a unique spin on them.

Informativeness:

Sometimes, the solution given should not be the 'main line'. For instance, in the below position, after we take on h3, the solution follows 2. Qg2 instead of taking on f5 (the most challenging line, which the player must have calculated before taking on h3).

(This puzzle would be a 1)

If the solution goes down the path of the best move for the opponent and offers the most challenging line, the puzzle should be scored highly. If, after seeing the solution, the puzzle makes sense, then it should be ranked highly.

Rating:

*Next to each puzzle, you will be shown the **chess.com ELO** of the player the puzzle was recommended to, along with their chess.com puzzle rating.*

This criteria is ranked on a 5-point scale. Given the ELO of the player this puzzle is recommended for, how appropriate is the puzzle? A 3 would be a 'just right' puzzle; as the number decreases, the puzzle is easier; as the number increases, the puzzle is more difficult. There is a bit more subjectivity with this criteria, but try to put yourself in the mind of a player of such ELO and/or consider if you would offer this puzzle to a student of such ELO.

Fun:

Rated on a 5-point scale, this criteria measures how enjoyable this puzzle would be for a player of the target ELO.

Quality:

Regardless of the difficulty level, take the puzzle as a whole and determine if it is a high-quality puzzle. The earlier criteria should help and inform, but they may not necessarily decide your final judgment.