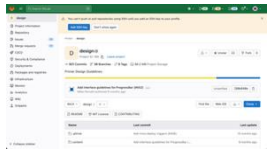
(1) Instruction & Length

Keep: Long instruction style



"Show all information in content."

Remove: Short instruction (<3 tokens) , item-style

"Delete." & "Bin."

Ground truth bbox
Success click
Fail click

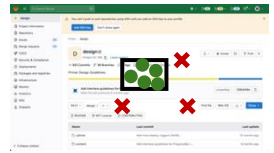
(2) Cross-domain

Remove: Web-like sample in non-web domain

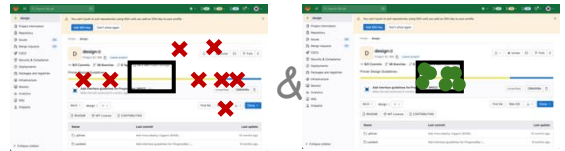
OCR detects: http://www....

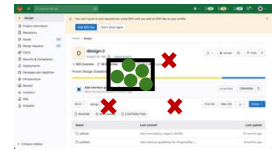
(3) 8-click success rate

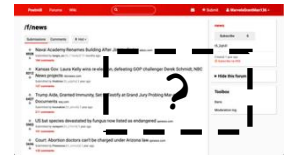
Keep: Mixed success (5/8)



Remove: All success (8/8) or All fail (0/8)







</>

(4) Difficulty scoring

8-click success rate w_1

Target size w_2

Prediction dispersion w_3

Localization error w_4

Parsing failures w_5

Raw Score (Weighted Sum)

$$\tilde{q}_i = \sum_{k=1}^{5} w_k \, g_{i,k}$$

Platform-wise percentile normalization

0 — 1

$q_i \in [0,1]$