| **Metric** | **Meaning** |
| --- | --- |
| **BLEU** | Measures **n-gram overlap** between the generated output and the reference. Higher = better accuracy. |
| **ROUGE** | Measures **recall**-based similarity (focuses on how much of the reference is captured). |
| **Pass@1** | Fraction of examples where the generated output **matches exactly** (often used in coding tasks, stricter). |
| **Avg Latency** | Average time (in milliseconds) the server took to respond to each prompt. Lower = faster. |