```
DEEP DATA BENCHMARK SONUÇLARI:
================================================================================
LIVECODEBENCH EVALUATION PIPELINE
Author: naholav
================================================================================
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Checkpoint directory: ./models
LiveCodeBench version: release_v5
Date range: 2408 - 2502
Platform filter: atcoder
Difficulty filter: easy
Model type filter: deep_instruction
Step filter: all
Include base model: False
Output directory: ./results/livecodebench
================================================================================

Discovered 4 checkpoints:
  deep_instruction: 4 checkpoints
✅ Downloaded: livecodebench/code_generation_lite/test.jsonl
[MAX_PROBLEMS] Using 41 problems (MAX_PROBLEMS=41)

Total problems to evaluate: 41

================================================================================
EVALUATING 41 PROBLEMS
================================================================================

================================================================================
EVALUATING: deep_instruction_checkpoint-step-100-epoch-1 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/deep_instruction/checkpoints/checkpoint-step-100-
epoch-1
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):   0% 0/41 [00:00<?,
?it/s]/usr/local/lib/python3.12/dist-
packages/transformers/generation/configuration_utils.py:537: UserWarning:
`do_sample` is set to `False`. However, `top_k` is set to `20` -- this flag is
```

```
only used in sample-based generation modes. You should set `do_sample=True` or
unset `top_k`.
  warnings.warn(
  [1] abc301_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):   2% 1/41 [00:17<11:57,
17.94s/it]  [2] abc301_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):   5% 2/41 [00:51<17:42,
27.24s/it]  [3] abc302_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):   7% 3/41 [01:34<21:48,
34.44s/it]  [4] abc303_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  10% 4/41 [01:56<18:14,
29.59s/it]  [5] abc303_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  12% 5/41 [02:46<22:04,
36.78s/it]  [6] abc304_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  15% 6/41 [03:14<19:39,
33.71s/it]  [7] abc304_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  17% 7/41 [03:33<16:26,
29.00s/it]  [8] abc305_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  20% 8/41 [04:06<16:37,
30.24s/it]  [9] abc305_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  22% 9/41 [04:50<18:29,
34.67s/it]  [10] abc306_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  24% 10/41
[04:57<13:32, 26.20s/it]  [11] abc306_b: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  27% 11/41
[05:19<12:27, 24.91s/it]  [12] abc307_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  29% 12/41
[05:45<12:05, 25.00s/it]  [13] abc307_b: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  32% 13/41
[06:18<12:51, 27.57s/it]  [14] abc308_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  34% 14/41
[06:53<13:22, 29.73s/it]  [15] abc308_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  37% 15/41
[07:26<13:20, 30.78s/it]  [16] abc309_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  39% 16/41
[08:01<13:22, 32.11s/it]  [17] abc309_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  41% 17/41
[08:57<15:39, 39.15s/it]  [18] abc310_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  44% 18/41
[09:12<12:16, 32.03s/it]  [19] abc310_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  46% 19/41
[09:51<12:26, 33.92s/it]  [20] abc311_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  49% 20/41
[10:24<11:49, 33.80s/it]  [21] abc311_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  51% 21/41
[10:58<11:17, 33.90s/it]  [22] abc312_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  54% 22/41
[11:09<08:32, 26.95s/it]  [23] abc312_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  56% 23/41
[12:08<10:58, 36.59s/it]  [24] abc313_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  59% 24/41
```

```
[12:38<09:50, 34.71s/it]  [25] abc314_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  61% 25/41
[12:53<07:38, 28.66s/it]  [26] abc314_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  63% 26/41
[13:20<07:02, 28.20s/it]  [27] abc315_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  66% 27/41
[13:27<05:06, 21.88s/it]  [28] abc315_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  68% 28/41
[14:09<06:00, 27.73s/it]  [29] abc318_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  71% 29/41
[14:59<06:54, 34.58s/it]  [30] abc318_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  73% 30/41
[15:29<06:05, 33.22s/it]  [31] abc319_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  76% 31/41
[16:05<05:40, 34.06s/it]  [32] abc320_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  78% 32/41
[16:31<04:43, 31.46s/it]  [33] abc320_b: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  80% 33/41
[17:14<04:41, 35.14s/it]  [34] abc321_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  83% 34/41
[17:34<03:34, 30.65s/it]  [35] abc321_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  85% 35/41
[18:19<03:29, 34.90s/it]  [36] abc322_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  88% 36/41
[18:47<02:43, 32.61s/it]  [37] abc322_b: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  90% 37/41
[18:58<01:44, 26.14s/it]  [38] abc323_a: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  93% 38/41
[19:10<01:06, 22.16s/it]  [39] abc323_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  95% 39/41
[19:45<00:51, 25.91s/it]  [40] abc324_a: PASS
Evaluating (deep_instruction_checkpoint-step-100-epoch-1):  98% 40/41
[19:54<00:20, 20.92s/it]  [41] abc324_b: FAIL
Evaluating (deep_instruction_checkpoint-step-100-epoch-1): 100% 41/41
[20:24<00:00, 29.86s/it]

============================================================
Results for deep_instruction_checkpoint-step-100-epoch-1 on 2408-2502_atcoder:
  Total: 41
  Passed: 11
  Failed: 30
  Errors: 0
  No tests: 0
  Pass@1: 26.83%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/deep_instruction_checkpoint-
step-100-epoch-1_2408-2502_atcoder.jsonl
  - Summary JSON: results/livecodebench/evaluations/deep_instruction_checkpoint-
```

```
step-100-epoch-1_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/deep_instruction_checkpoint-step-100-epoch-
1_2408-2502_atcoder.json

================================================================================
EVALUATING: deep_instruction_checkpoint-step-200-epoch-1 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/deep_instruction/checkpoints/checkpoint-step-200-
epoch-1
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):   2% 1/41 [00:23<15:58,
23.97s/it]  [2] abc301_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):   5% 2/41 [01:03<21:36,
33.24s/it]  [3] abc302_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):   7% 3/41 [01:47<24:12,
38.21s/it]  [4] abc303_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  10% 4/41 [02:06<18:49,
30.53s/it]  [5] abc303_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  12% 5/41 [03:09<25:24,
42.34s/it]  [6] abc304_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  15% 6/41 [03:35<21:20,
36.59s/it]  [7] abc304_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  17% 7/41 [03:51<16:53,
29.82s/it]  [8] abc305_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  20% 8/41 [04:25<17:07,
31.15s/it]  [9] abc305_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  22% 9/41 [05:02<17:40,
33.15s/it]  [10] abc306_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  24% 10/41
[05:10<13:01, 25.21s/it]  [11] abc306_b: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  27% 11/41
[05:32<12:13, 24.45s/it]  [12] abc307_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  29% 12/41
[05:57<11:48, 24.42s/it]  [13] abc307_b: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  32% 13/41
[06:27<12:13, 26.19s/it]  [14] abc308_a: PASS
```

```
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  34% 14/41
[06:41<10:06, 22.48s/it]  [15] abc308_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  37% 15/41
[07:15<11:12, 25.88s/it]  [16] abc309_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  39% 16/41
[07:26<08:55, 21.43s/it]  [17] abc309_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  41% 17/41
[07:48<08:42, 21.76s/it]  [18] abc310_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  44% 18/41
[08:08<08:04, 21.08s/it]  [19] abc310_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  46% 19/41
[08:53<10:21, 28.27s/it]  [20] abc311_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  49% 20/41
[09:38<11:41, 33.43s/it]  [21] abc311_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  51% 21/41
[10:12<11:11, 33.57s/it]  [22] abc312_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  54% 22/41
[10:24<08:31, 26.94s/it]  [23] abc312_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  56% 23/41
[11:17<10:25, 34.73s/it]  [24] abc313_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  59% 24/41
[11:25<07:38, 26.97s/it]  [25] abc314_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  61% 25/41
[11:40<06:13, 23.37s/it]  [26] abc314_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  63% 26/41
[12:08<06:09, 24.62s/it]  [27] abc315_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  66% 27/41
[12:16<04:35, 19.68s/it]  [28] abc315_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  68% 28/41
[12:54<05:28, 25.25s/it]  [29] abc318_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  71% 29/41
[13:35<05:56, 29.75s/it]  [30] abc318_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  73% 30/41
[14:29<06:47, 37.05s/it]  [31] abc319_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  76% 31/41
[15:06<06:12, 37.21s/it]  [32] abc320_a: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  78% 32/41
[15:27<04:51, 32.38s/it]  [33] abc320_b: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  80% 33/41
[16:04<04:30, 33.79s/it]  [34] abc321_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  83% 34/41
[16:19<03:15, 27.96s/it]  [35] abc321_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  85% 35/41
[16:54<03:01, 30.29s/it]  [36] abc322_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  88% 36/41
[17:18<02:20, 28.20s/it]  [37] abc322_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  90% 37/41
[17:28<01:31, 22.92s/it]  [38] abc323_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  93% 38/41
[17:40<00:58, 19.53s/it]  [39] abc323_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  95% 39/41
```

```
[18:02<00:40, 20.15s/it]  [40] abc324_a: PASS
Evaluating (deep_instruction_checkpoint-step-200-epoch-1):  98% 40/41
[18:11<00:16, 16.94s/it]  [41] abc324_b: FAIL
Evaluating (deep_instruction_checkpoint-step-200-epoch-1): 100% 41/41
[18:57<00:00, 27.73s/it]

============================================================
Results for deep_instruction_checkpoint-step-200-epoch-1 on 2408-2502_atcoder:
  Total: 41
  Passed: 13
  Failed: 28
  Errors: 0
  No tests: 0
  Pass@1: 31.71%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/deep_instruction_checkpoint-
step-200-epoch-1_2408-2502_atcoder.jsonl
  - Summary JSON: results/livecodebench/evaluations/deep_instruction_checkpoint-
step-200-epoch-1_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/deep_instruction_checkpoint-step-200-epoch-
1_2408-2502_atcoder.json

================================================================================
EVALUATING: deep_instruction_checkpoint-step-300-epoch-1 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/deep_instruction/checkpoints/checkpoint-step-300-
epoch-1
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):   2% 1/41 [00:13<09:02,
13.57s/it]  [2] abc301_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):   5% 2/41 [00:59<21:08,
32.53s/it]  [3] abc302_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):   7% 3/41 [01:46<24:55,
39.35s/it]  [4] abc303_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  10% 4/41 [02:04<19:05,
30.95s/it]  [5] abc303_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  12% 5/41 [02:55<22:49,
38.03s/it]  [6] abc304_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  15% 6/41 [03:12<17:57,
30.79s/it]  [7] abc304_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  17% 7/41 [03:31<15:21,
27.10s/it]  [8] abc305_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  20% 8/41 [04:09<16:44,
30.44s/it]  [9] abc305_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  22% 9/41 [04:40<16:23,
30.75s/it]  [10] abc306_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  24% 10/41
[04:50<12:30, 24.20s/it]  [11] abc306_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  27% 11/41
[05:18<12:38, 25.28s/it]  [12] abc307_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  29% 12/41
[05:44<12:21, 25.57s/it]  [13] abc307_b: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  32% 13/41
[06:17<13:05, 28.04s/it]  [14] abc308_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  34% 14/41
[06:28<10:17, 22.88s/it]  [15] abc308_b: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  37% 15/41
[07:08<12:08, 28.02s/it]  [16] abc309_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  39% 16/41
[07:19<09:27, 22.71s/it]  [17] abc309_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  41% 17/41
[08:00<11:20, 28.37s/it]  [18] abc310_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  44% 18/41
[08:25<10:28, 27.31s/it]  [19] abc310_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  46% 19/41
[09:08<11:45, 32.08s/it]  [20] abc311_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  49% 20/41
[10:06<13:52, 39.66s/it]  [21] abc311_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  51% 21/41
[10:36<12:19, 36.99s/it]  [22] abc312_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  54% 22/41
[10:47<09:14, 29.19s/it]  [23] abc312_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  56% 23/41
[11:46<11:26, 38.12s/it]  [24] abc313_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  59% 24/41
[12:11<09:40, 34.16s/it]  [25] abc314_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  61% 25/41
[22:00<53:29, 200.57s/it]  [26] abc314_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  63% 26/41
[22:29<37:16, 149.11s/it]  [27] abc315_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  66% 27/41
[22:36<24:51, 106.56s/it]  [28] abc315_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  68% 28/41
[23:08<18:13, 84.08s/it]   [29] abc318_a: PASS
```

```
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  71% 29/41
[23:44<13:56, 69.70s/it]  [30] abc318_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  73% 30/41
[24:25<11:10, 60.95s/it]  [31] abc319_b: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  76% 31/41
[24:59<08:50, 53.07s/it]  [32] abc320_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  78% 32/41
[25:11<06:05, 40.56s/it]  [33] abc320_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  80% 33/41
[25:48<05:16, 39.51s/it]  [34] abc321_a: PASS
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  83% 34/41
[26:02<03:43, 31.92s/it]  [35] abc321_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  85% 35/41
[26:53<03:45, 37.57s/it]  [36] abc322_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  88% 36/41
[27:23<02:57, 35.47s/it]  [37] abc322_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  90% 37/41
[27:38<01:56, 29.09s/it]  [38] abc323_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  93% 38/41
[27:48<01:10, 23.58s/it]  [39] abc323_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  95% 39/41
[28:07<00:44, 22.04s/it]  [40] abc324_a: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1):  98% 40/41
[28:17<00:18, 18.62s/it]  [41] abc324_b: FAIL
Evaluating (deep_instruction_checkpoint-step-300-epoch-1): 100% 41/41
[28:53<00:00, 42.27s/it]

============================================================
Results for deep_instruction_checkpoint-step-300-epoch-1 on 2408-2502_atcoder:
  Total: 41
  Passed: 11
  Failed: 30
  Errors: 0
  No tests: 0
  Pass@1: 26.83%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/deep_instruction_checkpoint-
step-300-epoch-1_2408-2502_atcoder.jsonl
  - Summary JSON: results/livecodebench/evaluations/deep_instruction_checkpoint-
step-300-epoch-1_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/deep_instruction_checkpoint-step-300-epoch-
1_2408-2502_atcoder.json

================================================================================
EVALUATING: deep_instruction_checkpoint-step-400-epoch-1 on 2408-2502_atcoder
problems
================================================================================
Loading model...

Loading LoRA checkpoint: models/deep_instruction/checkpoints/checkpoint-step-400-
epoch-1
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):   2% 1/41 [00:11<07:59,
11.99s/it]  [2] abc301_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):   5% 2/41 [01:12<26:19,
40.49s/it]  [3] abc302_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):   7% 3/41 [02:43<40:20,
63.69s/it]  [4] abc303_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  10% 4/41 [03:06<29:17,
47.49s/it]  [5] abc303_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  12% 5/41 [03:49<27:33,
45.92s/it]  [6] abc304_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  15% 6/41 [04:08<21:22,
36.64s/it]  [7] abc304_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  17% 7/41 [04:28<17:43,
31.27s/it]  [8] abc305_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  20% 8/41 [05:03<17:56,
32.61s/it]  [9] abc305_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  22% 9/41 [05:26<15:42,
29.47s/it]  [10] abc306_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  24% 10/41
[05:34<11:47, 22.83s/it]  [11] abc306_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  27% 11/41
[06:11<13:34, 27.15s/it]  [12] abc307_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  29% 12/41
[06:42<13:41, 28.32s/it]  [13] abc307_b: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  32% 13/41
[07:14<13:49, 29.64s/it]  [14] abc308_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  34% 14/41
[07:30<11:23, 25.33s/it]  [15] abc308_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  37% 15/41
[08:06<12:26, 28.72s/it]  [16] abc309_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  39% 16/41
[08:14<09:19, 22.37s/it]  [17] abc309_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  41% 17/41
[08:55<11:13, 28.04s/it]  [18] abc310_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  44% 18/41
```

```
[09:16<09:56, 25.92s/it]  [19] abc310_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  46% 19/41
[10:03<11:50, 32.28s/it]  [20] abc311_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  49% 20/41
[10:53<13:08, 37.56s/it]  [21] abc311_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  51% 21/41
[11:31<12:33, 37.68s/it]  [22] abc312_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  54% 22/41
[11:47<09:48, 30.97s/it]  [23] abc312_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  56% 23/41
[12:38<11:10, 37.26s/it]  [24] abc313_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  59% 24/41
[17:09<30:23, 107.24s/it]  [25] abc314_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  61% 25/41
[17:27<21:26, 80.43s/it]   [26] abc314_b: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  63% 26/41
[17:45<15:26, 61.77s/it]  [27] abc315_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  66% 27/41
[17:53<10:37, 45.56s/it]  [28] abc315_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  68% 28/41
[18:36<09:43, 44.85s/it]  [29] abc318_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  71% 29/41
[28:36<42:17, 211.43s/it]  [30] abc318_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  73% 30/41
[29:23<29:42, 162.07s/it]  [31] abc319_b: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  76% 31/41
[29:56<20:32, 123.23s/it]  [32] abc320_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  78% 32/41
[30:05<13:21, 89.01s/it]   [33] abc320_b: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  80% 33/41
[30:46<09:57, 74.69s/it]  [34] abc321_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  83% 34/41
[31:03<06:40, 57.27s/it]  [35] abc321_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  85% 35/41
[32:04<05:51, 58.55s/it]  [36] abc322_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  88% 36/41
[32:32<04:07, 49.41s/it]  [37] abc322_b: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  90% 37/41
[32:46<02:34, 38.75s/it]  [38] abc323_a: PASS
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  93% 38/41
[32:59<01:32, 31.00s/it]  [39] abc323_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  95% 39/41
[33:31<01:02, 31.24s/it]  [40] abc324_a: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1):  98% 40/41
[33:41<00:24, 24.83s/it]  [41] abc324_b: FAIL
Evaluating (deep_instruction_checkpoint-step-400-epoch-1): 100% 41/41
[34:17<00:00, 50.19s/it]

============================================================
Results for deep_instruction_checkpoint-step-400-epoch-1 on 2408-2502_atcoder:
  Total: 41
  Passed: 14
  Failed: 27
  Errors: 0
  No tests: 0
  Pass@1: 34.15%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/deep_instruction_checkpoint-step-400-epoch-1_2408-2502_atcoder.jsonl
  - Summary JSON: results/livecodebench/evaluations/deep_instruction_checkpoint-step-400-epoch-1_2408-2502_atcoder_results.json
  - LiveCodeBench format: results/livecodebench/generations/deep_instruction_checkpoint-step-400-epoch-1_2408-2502_atcoder.json



================================================================================
EVALUATION COMPLETE
================================================================================

Results saved to: ./results/livecodebench
Summary file: results/livecodebench/summary.json

Model                                              Pass@1     Problems
----------------------------------------------------------------------
deep_instruction_checkpoint-step-100-epoch-1       26.8%      41
deep_instruction_checkpoint-step-200-epoch-1       31.7%      41
deep_instruction_checkpoint-step-300-epoch-1       26.8%      41
deep_instruction_checkpoint-step-400-epoch-1       34.1%      41

====================================================================
DİVERSE DATA BENCHMARK SONUÇLARI:

/content/CodeGenBench
================================================================================
LIVECODEBENCH EVALUATION PIPELINE
Author: naholav
================================================================================
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Checkpoint directory: ./models
LiveCodeBench version: release_v5
Date range: 2408 - 2502
Platform filter: atcoder
Difficulty filter: easy
Model type filter: diverse_instruction
Step filter: all
Include base model: False
Output directory: ./results/livecodebench
```

```
================================================================================

Discovered 5 checkpoints:
  diverse_instruction: 5 checkpoints
✅ Downloaded: livecodebench/code_generation_lite/test.jsonl
[MAX_PROBLEMS] Using 41 problems (MAX_PROBLEMS=41)

Total problems to evaluate: 41

================================================================================
EVALUATING 41 PROBLEMS
================================================================================

================================================================================
EVALUATING: diverse_instruction_checkpoint-step-500-epoch-2 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/diverse_instruction/checkpoints/checkpoint-step-
500-epoch-2
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):   0% 0/41 [00:00<?,
?it/s]/usr/local/lib/python3.12/dist-
packages/transformers/generation/configuration_utils.py:537: UserWarning:
`do_sample` is set to `False`. However, `top_k` is set to `20` -- this flag is
only used in sample-based generation modes. You should set `do_sample=True` or
unset `top_k`.
  warnings.warn(
  [1] abc301_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):   2% 1/41
[00:27<18:03, 27.10s/it]  [2] abc301_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):   5% 2/41
[01:21<27:59, 43.06s/it]  [3] abc302_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):   7% 3/41
[01:42<20:51, 32.93s/it]  [4] abc303_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  10% 4/41
[02:25<22:54, 37.15s/it]  [5] abc303_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  12% 5/41
[12:10<2:20:49, 234.71s/it]  [6] abc304_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  15% 6/41
```

```
[12:49<1:38:00, 168.02s/it]  [7] abc304_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  17% 7/41
[13:38<1:13:13, 129.22s/it]  [8] abc305_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  20% 8/41
[14:08<53:38, 97.53s/it]     [9] abc305_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  22% 9/41
[14:54<43:25, 81.41s/it]  [10] abc306_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  24% 10/41
[15:04<30:38, 59.32s/it]  [11] abc306_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  27% 11/41
[15:32<24:50, 49.70s/it]  [12] abc307_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  29% 12/41
[15:57<20:23, 42.19s/it]  [13] abc307_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  32% 13/41
[16:34<19:02, 40.81s/it]  [14] abc308_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  34% 14/41
[17:02<16:39, 37.00s/it]  [15] abc308_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  37% 15/41
[17:33<15:14, 35.17s/it]  [16] abc309_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  39% 16/41
[27:12<1:22:49, 198.80s/it]  [17] abc309_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  41% 17/41
[28:19<1:03:36, 159.03s/it]  [18] abc310_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  44% 18/41
[28:41<45:14, 118.03s/it]    [19] abc310_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  46% 19/41
[29:39<36:39, 99.97s/it]   [20] abc311_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  49% 20/41
[30:33<30:11, 86.25s/it]  [21] abc311_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  51% 21/41
[31:01<22:54, 68.70s/it]  [22] abc312_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  54% 22/41
[31:16<16:39, 52.62s/it]  [23] abc312_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  56% 23/41
[32:05<15:28, 51.56s/it]  [24] abc313_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  59% 24/41
[32:33<12:36, 44.47s/it]  [25] abc314_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  61% 25/41
[33:03<10:39, 39.95s/it]  [26] abc314_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  63% 26/41
[33:35<09:24, 37.66s/it]  [27] abc315_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  66% 27/41
[33:44<06:45, 29.00s/it]  [28] abc315_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  68% 28/41
[34:21<06:49, 31.48s/it]  [29] abc318_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  71% 29/41
[35:01<06:46, 33.91s/it]  [30] abc318_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  73% 30/41
[35:54<07:16, 39.69s/it]  [31] abc319_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  76% 31/41
[36:40<06:55, 41.53s/it]  [32] abc320_a: PASS
```

```
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  78% 32/41
[37:07<05:34, 37.16s/it]  [33] abc320_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  80% 33/41
[37:53<05:19, 39.94s/it]  [34] abc321_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  83% 34/41
[38:14<03:58, 34.10s/it]  [35] abc321_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  85% 35/41
[39:15<04:13, 42.22s/it]  [36] abc322_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  88% 36/41
[39:41<03:06, 37.33s/it]  [37] abc322_b: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  90% 37/41
[39:55<02:01, 30.42s/it]  [38] abc323_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  93% 38/41
[40:05<01:12, 24.16s/it]  [39] abc323_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  95% 39/41
[40:45<00:58, 29.10s/it]  [40] abc324_a: PASS
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2):  98% 40/41
[40:55<00:23, 23.37s/it]  [41] abc324_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-500-epoch-2): 100% 41/41
[41:35<00:00, 60.87s/it]

============================================================
Results for diverse_instruction_checkpoint-step-500-epoch-2 on 2408-2502_atcoder:
  Total: 41
  Passed: 17
  Failed: 24
  Errors: 0
  No tests: 0
  Pass@1: 41.46%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/diverse_instruction_checkpoint-
step-500-epoch-2_2408-2502_atcoder.jsonl
  - Summary JSON:
results/livecodebench/evaluations/diverse_instruction_checkpoint-step-500-epoch-
2_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/diverse_instruction_checkpoint-step-500-epoch-
2_2408-2502_atcoder.json

================================================================================
EVALUATING: diverse_instruction_checkpoint-step-600-epoch-2 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/diverse_instruction/checkpoints/checkpoint-step-
600-epoch-2
```

```
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):   2% 1/41
[00:25<17:02, 25.56s/it]  [2] abc301_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):   5% 2/41
[01:09<23:28, 36.11s/it]  [3] abc302_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):   7% 3/41
[01:32<19:15, 30.40s/it]  [4] abc303_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  10% 4/41
[02:15<21:40, 35.16s/it]  [5] abc303_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  12% 5/41
[12:03<2:20:46, 234.63s/it]  [6] abc304_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  15% 6/41
[12:41<1:37:50, 167.72s/it]  [7] abc304_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  17% 7/41
[13:26<1:12:20, 127.68s/it]  [8] abc305_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  20% 8/41
[14:07<55:04, 100.12s/it]    [9] abc305_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  22% 9/41
[14:34<41:08, 77.14s/it]   [10] abc306_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  24% 10/41
[14:44<29:10, 56.48s/it]  [11] abc306_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  27% 11/41
[15:16<24:33, 49.11s/it]  [12] abc307_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  29% 12/41
[15:46<20:56, 43.33s/it]  [13] abc307_b: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  32% 13/41
[16:22<19:08, 41.02s/it]  [14] abc308_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  34% 14/41
[16:54<17:14, 38.32s/it]  [15] abc308_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  37% 15/41
[17:31<16:20, 37.73s/it]  [16] abc309_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  39% 16/41
[18:15<16:32, 39.68s/it]  [17] abc309_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  41% 17/41
[19:22<19:11, 47.97s/it]  [18] abc310_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  44% 18/41
[19:51<16:11, 42.23s/it]  [19] abc310_b: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  46% 19/41
[20:41<16:18, 44.46s/it]  [20] abc311_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  49% 20/41
[21:27<15:47, 45.13s/it]  [21] abc311_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  51% 21/41
```

```
[22:03<14:03, 42.20s/it]  [22] abc312_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  54% 22/41
[22:18<10:48, 34.11s/it]  [23] abc312_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  56% 23/41
[23:13<12:07, 40.43s/it]  [24] abc313_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  59% 24/41
[23:40<10:18, 36.36s/it]  [25] abc314_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  61% 25/41
[24:11<09:18, 34.91s/it]  [26] abc314_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  63% 26/41
[24:52<09:06, 36.46s/it]  [27] abc315_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  66% 27/41
[25:03<06:45, 28.97s/it]  [28] abc315_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  68% 28/41
[25:42<06:56, 32.01s/it]  [29] abc318_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  71% 29/41
[26:12<06:16, 31.39s/it]  [30] abc318_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  73% 30/41
[26:55<06:21, 34.71s/it]  [31] abc319_b: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  76% 31/41
[27:36<06:07, 36.78s/it]  [32] abc320_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  78% 32/41
[28:03<05:03, 33.68s/it]  [33] abc320_b: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  80% 33/41
[28:49<05:01, 37.65s/it]  [34] abc321_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  83% 34/41
[29:10<03:48, 32.59s/it]  [35] abc321_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  85% 35/41
[30:03<03:52, 38.69s/it]  [36] abc322_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  88% 36/41
[30:36<03:04, 36.93s/it]  [37] abc322_b: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  90% 37/41
[30:49<01:58, 29.72s/it]  [38] abc323_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  93% 38/41
[30:58<01:11, 23.67s/it]  [39] abc323_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  95% 39/41
[31:39<00:57, 28.61s/it]  [40] abc324_a: PASS
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2):  98% 40/41
[31:48<00:22, 22.99s/it]  [41] abc324_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-600-epoch-2): 100% 41/41
[32:19<00:00, 47.30s/it]

============================================================
Results for diverse_instruction_checkpoint-step-600-epoch-2 on 2408-2502_atcoder:
  Total: 41
  Passed: 16
  Failed: 25
  Errors: 0
  No tests: 0
  Pass@1: 39.02%
============================================================
Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/diverse_instruction_checkpoint-
step-600-epoch-2_2408-2502_atcoder.jsonl
  - Summary JSON:
results/livecodebench/evaluations/diverse_instruction_checkpoint-step-600-epoch-
2_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/diverse_instruction_checkpoint-step-600-epoch-
2_2408-2502_atcoder.json

================================================================================
EVALUATING: diverse_instruction_checkpoint-step-700-epoch-2 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/diverse_instruction/checkpoints/checkpoint-step-
700-epoch-2
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):   2% 1/41
[00:21<14:37, 21.94s/it]  [2] abc301_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):   5% 2/41
[01:09<23:59, 36.90s/it]  [3] abc302_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):   7% 3/41
[01:32<19:18, 30.48s/it]  [4] abc303_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  10% 4/41
[02:10<20:37, 33.44s/it]  [5] abc303_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  12% 5/41
[03:03<24:19, 40.54s/it]  [6] abc304_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  15% 6/41
[03:40<22:57, 39.34s/it]  [7] abc304_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  17% 7/41
[04:26<23:30, 41.48s/it]  [8] abc305_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  20% 8/41
[04:59<21:24, 38.91s/it]  [9] abc305_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  22% 9/41
[05:28<19:01, 35.67s/it]  [10] abc306_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  24% 10/41
[05:36<14:08, 27.36s/it]  [11] abc306_b: FAIL
```

```
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  27% 11/41
[06:06<13:58, 27.94s/it]  [12] abc307_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  29% 12/41
[06:30<12:59, 26.86s/it]  [13] abc307_b: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  32% 13/41
[07:06<13:50, 29.66s/it]  [14] abc308_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  34% 14/41
[07:35<13:14, 29.43s/it]  [15] abc308_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  37% 15/41
[08:12<13:41, 31.59s/it]  [16] abc309_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  39% 16/41
[08:43<13:07, 31.50s/it]  [17] abc309_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  41% 17/41
[09:08<11:50, 29.60s/it]  [18] abc310_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  44% 18/41
[09:32<10:44, 28.01s/it]  [19] abc310_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  46% 19/41
[10:26<13:05, 35.69s/it]  [20] abc311_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  49% 20/41
[11:22<14:34, 41.64s/it]  [21] abc311_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  51% 21/41
[11:51<12:42, 38.12s/it]  [22] abc312_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  54% 22/41
[12:04<09:39, 30.47s/it]  [23] abc312_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  56% 23/41
[13:20<13:15, 44.19s/it]  [24] abc313_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  59% 24/41
[13:44<10:46, 38.01s/it]  [25] abc314_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  61% 25/41
[14:16<09:41, 36.37s/it]  [26] abc314_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  63% 26/41
[14:58<09:27, 37.85s/it]  [27] abc315_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  66% 27/41
[15:07<06:49, 29.28s/it]  [28] abc315_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  68% 28/41
[15:47<07:02, 32.52s/it]  [29] abc318_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  71% 29/41
[16:34<07:23, 36.96s/it]  [30] abc318_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  73% 30/41
[17:14<06:54, 37.64s/it]  [31] abc319_b: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  76% 31/41
[17:56<06:31, 39.13s/it]  [32] abc320_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  78% 32/41
[18:06<04:32, 30.30s/it]  [33] abc320_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  80% 33/41
[18:58<04:54, 36.87s/it]  [34] abc321_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  83% 34/41
[19:21<03:49, 32.72s/it]  [35] abc321_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  85% 35/41
[19:59<03:26, 34.41s/it]  [36] abc322_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  88% 36/41
```

```
[20:28<02:42, 32.53s/it]  [37] abc322_b: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  90% 37/41
[20:42<01:47, 27.00s/it]  [38] abc323_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  93% 38/41
[20:56<01:09, 23.18s/it]  [39] abc323_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  95% 39/41
[21:38<00:57, 28.97s/it]  [40] abc324_a: PASS
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2):  98% 40/41
[21:48<00:23, 23.26s/it]  [41] abc324_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-700-epoch-2): 100% 41/41
[22:40<00:00, 33.19s/it]

============================================================
Results for diverse_instruction_checkpoint-step-700-epoch-2 on 2408-2502_atcoder:
  Total: 41
  Passed: 13
  Failed: 28
  Errors: 0
  No tests: 0
  Pass@1: 31.71%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/diverse_instruction_checkpoint-
step-700-epoch-2_2408-2502_atcoder.jsonl
  - Summary JSON:
results/livecodebench/evaluations/diverse_instruction_checkpoint-step-700-epoch-
2_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/diverse_instruction_checkpoint-step-700-epoch-
2_2408-2502_atcoder.json

================================================================================
EVALUATING: diverse_instruction_checkpoint-step-800-epoch-3 on 2408-2502_atcoder
problems
================================================================================

Loading model...

Loading LoRA checkpoint: models/diverse_instruction/checkpoints/checkpoint-step-
800-epoch-3
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
```

```
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):   2% 1/41
[00:22<15:13, 22.85s/it]  [2] abc301_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):   5% 2/41
[01:23<29:27, 45.32s/it]  [3] abc302_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):   7% 3/41
[01:46<22:10, 35.01s/it]  [4] abc303_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  10% 4/41
[02:22<21:52, 35.46s/it]  [5] abc303_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  12% 5/41
[03:14<24:41, 41.14s/it]  [6] abc304_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  15% 6/41
[03:51<23:21, 40.04s/it]  [7] abc304_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  17% 7/41
[04:55<27:05, 47.82s/it]  [8] abc305_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  20% 8/41
[05:32<24:16, 44.15s/it]  [9] abc305_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  22% 9/41
[06:06<21:57, 41.16s/it]  [10] abc306_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  24% 10/41
[06:16<16:17, 31.55s/it]  [11] abc306_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  27% 11/41
[06:44<15:15, 30.53s/it]  [12] abc307_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  29% 12/41
[07:13<14:28, 29.93s/it]  [13] abc307_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  32% 13/41
[07:48<14:44, 31.60s/it]  [14] abc308_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  34% 14/41
[08:17<13:48, 30.70s/it]  [15] abc308_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  37% 15/41
[08:49<13:24, 30.95s/it]  [16] abc309_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  39% 16/41
[09:19<12:50, 30.80s/it]  [17] abc309_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  41% 17/41
[09:59<13:24, 33.53s/it]  [18] abc310_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  44% 18/41
[10:32<12:50, 33.52s/it]  [19] abc310_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  46% 19/41
[11:42<16:17, 44.45s/it]  [20] abc311_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  49% 20/41
[12:41<17:06, 48.86s/it]  [21] abc311_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  51% 21/41
[13:19<15:10, 45.54s/it]  [22] abc312_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  54% 22/41
[13:37<11:46, 37.19s/it]  [23] abc312_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  56% 23/41
[15:11<16:14, 54.13s/it]  [24] abc313_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  59% 24/41
[15:35<12:49, 45.26s/it]  [25] abc314_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  61% 25/41
```

```
[16:21<12:09, 45.57s/it]  [26] abc314_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  63% 26/41
[17:03<11:05, 44.37s/it]  [27] abc315_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  66% 27/41
[17:14<08:00, 34.32s/it]  [28] abc315_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  68% 28/41
[17:59<08:06, 37.44s/it]  [29] abc318_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  71% 29/41
[18:31<07:12, 36.08s/it]  [30] abc318_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  73% 30/41
[19:14<06:56, 37.89s/it]  [31] abc319_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  76% 31/41
[19:58<06:39, 39.94s/it]  [32] abc320_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  78% 32/41
[20:08<04:37, 30.88s/it]  [33] abc320_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  80% 33/41
[20:51<04:35, 34.38s/it]  [34] abc321_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  83% 34/41
[21:14<03:36, 31.00s/it]  [35] abc321_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  85% 35/41
[22:02<03:36, 36.09s/it]  [36] abc322_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  88% 36/41
[22:29<02:47, 33.54s/it]  [37] abc322_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  90% 37/41
[22:42<01:49, 27.39s/it]  [38] abc323_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  93% 38/41
[22:58<01:11, 23.82s/it]  [39] abc323_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  95% 39/41
[23:40<00:58, 29.42s/it]  [40] abc324_a: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3):  98% 40/41
[23:51<00:23, 23.78s/it]  [41] abc324_b: PASS
Evaluating (diverse_instruction_checkpoint-step-800-epoch-3): 100% 41/41
[24:38<00:00, 36.06s/it]

============================================================
Results for diverse_instruction_checkpoint-step-800-epoch-3 on 2408-2502_atcoder:
  Total: 41
  Passed: 18
  Failed: 23
  Errors: 0
  No tests: 0
  Pass@1: 43.90%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/diverse_instruction_checkpoint-
step-800-epoch-3_2408-2502_atcoder.jsonl
  - Summary JSON:
results/livecodebench/evaluations/diverse_instruction_checkpoint-step-800-epoch-
3_2408-2502_atcoder_results.json
```

```
  - LiveCodeBench format:
results/livecodebench/generations/diverse_instruction_checkpoint-step-800-epoch-
3_2408-2502_atcoder.json

====================================================================
EVALUATING: diverse_instruction_checkpoint-step-852-epoch-3 on 2408-2502_atcoder
problems
====================================================================

Loading model...

Loading LoRA checkpoint: models/diverse_instruction/checkpoints/checkpoint-step-
852-epoch-3
Base model: Qwen/Qwen2.5-Coder-1.5B-Instruct
Special tokens have been added in the vocabulary, make sure the associated word
embeddings are fine-tuned or trained.
LoRA checkpoint loaded (torch.bfloat16)
Flash Attention 2 enabled

Generating and evaluating 41 problems...
(Skipping 0 already processed)
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):   0% 0/41 [00:00<?,
?it/s]  [1] abc301_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):   2% 1/41
[00:25<16:53, 25.33s/it]  [2] abc301_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):   5% 2/41
[01:20<27:58, 43.03s/it]  [3] abc302_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):   7% 3/41
[01:43<21:24, 33.80s/it]  [4] abc303_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  10% 4/41
[02:19<21:24, 34.72s/it]  [5] abc303_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  12% 5/41
[03:07<23:44, 39.58s/it]  [6] abc304_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  15% 6/41
[03:43<22:17, 38.22s/it]  [7] abc304_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  17% 7/41
[04:23<22:02, 38.89s/it]  [8] abc305_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  20% 8/41
[04:56<20:18, 36.93s/it]  [9] abc305_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  22% 9/41
[14:54<1:53:13, 212.29s/it]  [10] abc306_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  24% 10/41
[15:04<1:17:28, 149.94s/it]  [11] abc306_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  27% 11/41
[15:38<57:12, 114.42s/it]    [12] abc307_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  29% 12/41
[16:01<41:53, 86.67s/it]   [13] abc307_b: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  32% 13/41
[16:39<33:33, 71.91s/it]  [14] abc308_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  34% 14/41
```

```
[17:07<26:25, 58.71s/it]  [15] abc308_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  37% 15/41
[17:40<21:59, 50.76s/it]  [16] abc309_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  39% 16/41
[18:17<19:26, 46.67s/it]  [17] abc309_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  41% 17/41
[18:58<17:58, 44.93s/it]  [18] abc310_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  44% 18/41
[19:28<15:34, 40.65s/it]  [19] abc310_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  46% 19/41
[20:18<15:54, 43.39s/it]  [20] abc311_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  49% 20/41
[21:15<16:33, 47.32s/it]  [21] abc311_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  51% 21/41
[21:48<14:22, 43.14s/it]  [22] abc312_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  54% 22/41
[22:06<11:14, 35.50s/it]  [23] abc312_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  56% 23/41
[23:59<17:39, 58.86s/it]  [24] abc313_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  59% 24/41
[24:24<13:47, 48.70s/it]  [25] abc314_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  61% 25/41
[25:03<12:09, 45.61s/it]  [26] abc314_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  63% 26/41
[25:45<11:08, 44.60s/it]  [27] abc315_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  66% 27/41
[25:54<07:57, 34.11s/it]  [28] abc315_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  68% 28/41
[26:36<07:53, 36.40s/it]  [29] abc318_a: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  71% 29/41
[27:10<07:06, 35.56s/it]  [30] abc318_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  73% 30/41
[27:55<07:02, 38.40s/it]  [31] abc319_b: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  76% 31/41
[28:40<06:44, 40.44s/it]  [32] abc320_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  78% 32/41
[28:50<04:41, 31.24s/it]  [33] abc320_b: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  80% 33/41
[29:36<04:44, 35.62s/it]  [34] abc321_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  83% 34/41
[29:59<03:44, 32.08s/it]  [35] abc321_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  85% 35/41
[30:45<03:36, 36.16s/it]  [36] abc322_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  88% 36/41
[31:13<02:48, 33.62s/it]  [37] abc322_b: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  90% 37/41
[31:27<01:50, 27.71s/it]  [38] abc323_a: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  93% 38/41
[31:42<01:12, 24.04s/it]  [39] abc323_b: FAIL
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  95% 39/41
[32:12<00:51, 25.90s/it]  [40] abc324_a: PASS
```

```
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3):  98% 40/41
[32:23<00:21, 21.29s/it]  [41] abc324_b: PASS
Evaluating (diverse_instruction_checkpoint-step-852-epoch-3): 100% 41/41
[33:05<00:00, 48.42s/it]

============================================================
Results for diverse_instruction_checkpoint-step-852-epoch-3 on 2408-2502_atcoder:
  Total: 41
  Passed: 17
  Failed: 24
  Errors: 0
  No tests: 0
  Pass@1: 41.46%
============================================================

Results saved to:
  - Detailed JSONL: results/livecodebench/detailed/diverse_instruction_checkpoint-
step-852-epoch-3_2408-2502_atcoder.jsonl
  - Summary JSON:
results/livecodebench/evaluations/diverse_instruction_checkpoint-step-852-epoch-
3_2408-2502_atcoder_results.json
  - LiveCodeBench format:
results/livecodebench/generations/diverse_instruction_checkpoint-step-852-epoch-
3_2408-2502_atcoder.json



================================================================================
EVALUATION COMPLETE
================================================================================

Results saved to: ./results/livecodebench
Summary file: results/livecodebench/summary.json

Model                                              Pass@1     Problems
----------------------------------------------------------------------
diverse_instruction_checkpoint-step-500-epoch-2    41.5%      41
diverse_instruction_checkpoint-step-600-epoch-2    39.0%      41
diverse_instruction_checkpoint-step-700-epoch-2    31.7%      41
diverse_instruction_checkpoint-step-800-epoch-3    43.9%      41
diverse_instruction_checkpoint-step-852-epoch-3    41.5%      41

================================================================================
```