JSONL Cheat Sheet

jsonlines.org

JSON Lines (Newline-Delimited JSON) — The AI Training Format

WHAT IS JSONL?

JSONL (JSON Lines) is a text file format where every line contains exactly one valid JSON value. The standard fine-tuning format for GPT, Llama, Mistral, and virtually every major LLM.

THE 3 OFFICIAL RULES (JSONLINES.ORG)

01 UTF-8 Encoding
Required. No BOM (byte order mark) allowed.

02 Each line = valid JSON
Object, array, string, number, bool, or null.

03 Line terminator = \n
\r\n also accepted. Trailing \n recommended.

FORMAT QUICK REFERENCE

Extension	.jsonl
MIME type	application/jsonl
Encoding	UTF-8 (required)
Line sep.	\n or \r\n
Compressed	.jsonl.gz / .jsonl.bz2
Also called	ndjson, nd-json

training.jsonl

```
# Each line = one training example
{"messages": [
  {"role":"system",
   "content":"You are helpful."},
  {"role":"user",
   "content":"What is JSON?"},
  {"role":"assistant",
   "content":"JSON is..."}
]}
# ← entire above is ONE line in the .jsonl file
```

VALID VS INVALID JSONL

✓ `{"id":1,"text":"Hello"}`
Valid — JSON object on one line

✓ `{"id":2,"text":"World"}`
Valid — second independent record

✗ `[{"id":1},{"id":2}]`
Invalid — JSON array wrapper

✗ `{"id":1},`
Invalid — trailing comma

✗ `(empty line)`
Invalid — blank line

✓ `null`
Valid — any JSON value is ok

PLATFORM FORMAT DIFFERENCES

OpenAI / Azure
`{"messages":[{"role":"...","content":"..."}]}`

Google Vertex
`{"messages":[{"author":"...","content":"..."}]}`

Mistral AI
`{"messages":[{"role":"...","content":"..."}]}`

Llama / Axolotl
`{"instruction":"...","output":"..."}`

HuggingFace
`{"text":"..."}` or custom schema

jsonl_utils.py

```python
import json

# Read JSONL
with open("data.jsonl") as f:
    for line in f:
        obj = json.loads(line.strip())

# Write JSONL
with open("out.jsonl", "w") as f:
    for r in records:
        f.write(json.dumps(r) + "\n")

# Convert JSON array → JSONL
data = json.load(f)
for r in data:
    out.write(json.dumps(r)+"\n")
```

COMMON MISTAKES

JSON array wrapper
Remove [ ] — no outer structure in JSONL

Blank lines
Every line must be a valid JSON value

Trailing commas
End each line with } or] — no comma

Wrong encoding
Must be UTF-8. Check: file -i data.jsonl

Multi-line JSON
Each object must fit on ONE single line

json.dump() not dumps()
Use dumps(r) + '\n', not dump(r, f)

Missing final \n
Add trailing newline after last record

JSON VS JSONL

Property	JSON	JSONL
Structure	Single document	One record per line
Streaming	Needs full parse	Line-by-line ✓
Append	Requires rewrite	Just append a line ✓
Memory	Loads whole file	One line at a time ✓
LLM training	Not standard	Universal standard ✓