Class 21: Sprint 2 Planning

- LLM Randomness
- Standup Meeting
- Sprint 2 Planning
- Studio

Copyright 2026, Kyle J. Harms

LLM Randomness

Why do responses change for the same prompt?

LLMs generate responses one token at a time.

1. Calculate a probability for every possible next token.
2. **Randomly sample** one token using those probabilities.
3. Add it to the response and repeat.

Step 2 can produce different responses to the same prompt.

`temperature`

Temperature (0.0 to 2.0) changes the odds of which token is selected.

Lower `temperature` is more predictable. (Closer to 0)

Higher `temperature` can be more “creative”. (Closer to 2)

`temperature: 0` always responds with the most likely token.

Example: `temperature`

Prompt: “The best pizza topping is…”

Next token	`0`	`0.5`	`1`	`2`
pepperoni	100%	73%	50%	35%
mushrooms	0%	18%	25%	25%
sausage	0%	7%	15%	19%
onions	0%	1%	7%	13%
pineapple	0%	< 1%	3%	9%

WebLLM: `temperature`

```
await llmEngine.generate(messages.value, {
  temperature: 0.2,
  onResponse: (text) => (response.value = text),
});
```

Model Defaults:

`Llama-3.2-1B-Instruct-q4f16_1-MLC`: `temperature` 0.6

Example: Consistency, not Creativity

```
const messages = [
  {
    role: "system",
    content: `Given a review, assign a rating from 1 to 5,
              where 1 is the worst and 5 is the best.
              Output the rating as only JSON. Example: {"rating": 4}`,
  },
  { role: "user", content: "The food was okay, but the service was slow." },
];

const reply = await llmEngine.generate(messages, { temperature: 0, max_tokens: 20 });
```

Gotcha: Consistent does not guarantee **correctness**.

Project

Activity: Standup Meeting

Respect your team members by giving them your **full** attention; put away laptops and phones.

Stand Up.

- What did I work on since the last time my group met?
- What am I working on today?
- What issues are blocking me?

Time limit: 5 minutes (Standup meeting should be short!)

Sprint 2

Core Functionality

Review: Sprint Planning

1. What is our goal for this sprint?
2. Which backlog items can we complete in this sprint?
3. How will the work get done, and who will do it?

Gotcha: Once planning is complete, the sprint plan is locked.

Activity: Sprint 2 Planning

Working with your team, answer the two sprint planning questions on the handout provided.

Suggestion: Focus on your team's **greatest source of uncertainty**. (This is probably your "backend" type stuff – the core functionality. Getting the LLM working well is more important right now than the fun "frontend" user interface stuff.)

Studio: Project

Use this time to work on sprint 1 or 2.

Today: Sprint 1 due

Next Thursday: Sprint 2 due