EFSP Auxiliary ESL Curriculum

AI Development Dialogue Lab

Realistic workplace dialogues, role-play cards, and debrief prompts for advanced ESL learners in AI teams

Audience: instructors, coaches, peer practice groups, and technical English cohorts

Focus: high-level professional English for AI development teams, including technical vocabulary, engineering discussion patterns, research/product tradeoffs, evaluation language, risk communication, and realistic workplace dialogue.

Designed for advanced ESL learners who already work with software, data, or AI systems and need field-specific fluency rather than basic grammar instruction.

Teaching stance: AI language changes quickly. Teach learners to ask precise clarification questions, define terms in context, and distinguish research claims, implementation details, benchmark results, and product promises.

How to Run the Dialogue Lab

1. Use groups of three: speaker, counterpart, observer.
2. Read the model dialogue once. Then replay it using the same situation but new details from the learner's work.
3. The observer listens for terminology accuracy, clarification questions, evidence, risk language, and decision clarity.
4. After each role-play, replay the hardest 30 seconds with a more precise sentence.

Facilitator guardrail

Do not let learners hide behind jargon. Ask them to define the term in plain English and connect it to a user, metric, risk, or engineering decision.

1. Standup: RAG Latency Spike

Setting

Morning standup for a document assistant.

Speaker	Line
PM	Are we still on track for the pilot on Friday?
Engineer	Functionally, yes. The blocker is latency. P95 went from 4.2 seconds to 8.7 after we added reranking.
ESL learner	So the answer quality improved, but the serving path is too slow. Is the bottleneck retrieval, reranking, or generation?
Engineer	Mostly reranking. The generator time is stable.
ESL learner	Then my proposal is to keep reranking for high-risk queries only and use the faster path for simple FAQ queries. I can bring an ablation by end of day.

Language notes

- P95 means the 95th percentile request latency.
- Ablation means testing the effect of removing or changing one component.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

2. Design Review: Prompt Fix or Fine-Tune?

Setting

Design review for a customer-support summarizer.

Speaker	Line
Researcher	We should fine-tune. The model keeps missing refund-policy exceptions.
Product manager	Would fine-tuning actually solve that, or is the policy changing too often?
ESL learner	I would separate style from facts. If the issue is current policy knowledge, RAG may be safer. If the issue is summary format, a prompt or fine-tune could help.
Researcher	Good point. The examples show both problems.
ESL learner	Let's run two eval slices: factual policy coverage and format compliance. Then we can choose the adaptation method with evidence.

Language notes

- Eval slice means a subset of test cases focused on one behavior.
- Adaptation method means the way the team changes model behavior: prompt, retrieval, fine-tuning, or product logic.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

3. Data Meeting: Label Ambiguity

Setting

Annotation guideline meeting for safety classification.

Speaker	Line
Data lead	Annotators disagree on whether this is medical advice or general wellness information.
ESL learner	The guideline needs a decision rule. If the answer recommends dosage, diagnosis, or treatment, we label it medical advice. If it gives general prevention information, we label it wellness.
Reviewer	What about borderline cases?
ESL learner	We should add a borderline tag and route those to expert review. Otherwise the ground truth will be noisy.

Language notes

- Noisy labels are labels that are inconsistent, wrong, or ambiguous.
- Ground truth should be as consistent as possible, but it is often a human-created artifact.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

4. Incident Update: Tool-Calling Failure

Setting

Incident channel after an agent booked duplicate appointments.

Speaker	Line
Support	Customers are seeing duplicate calendar events.
ESL learner	We found the immediate cause. The model retried the booking tool after a timeout, but the first call had actually succeeded.
Engineering manager	Mitigation?
ESL learner	We disabled automatic retries for non-idempotent tool calls and added an idempotency key. We are checking logs for affected users now.
Support	What should we tell customers?
ESL learner	Say the system may have created a duplicate event during a retry window. We are removing duplicates and will confirm by email.

Language notes

- Non-idempotent means repeating the action can create a different or duplicate result.
- An idempotency key helps the system recognize that a retry belongs to the same intended action.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

5. Eval Readout: Better Average, Worse Edge Cases

Setting

Release meeting for a model upgrade.

Speaker	Line
PM	The average score is up. Can we ship?
ESL learner	I recommend a hold. The overall pass rate improved from 86% to 89%, but the legal-disclaimer slice regressed by 11 points.
Researcher	Is that statistically meaningful?
ESL learner	The sample is small, so I would not overclaim. But the failures are severe enough to block release until we inspect them.
PM	What is the next step?
ESL learner	Human review of the failed slice today, then a targeted prompt or policy fix before we rerun the regression suite.

Language notes

- Do not report a metric without explaining the subset, sample size, and severity of failures.
- A release can be blocked by a small number of severe failures.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

6. Security Review: Prompt Injection

Setting

Security review for a browsing agent.

Speaker	Line
Security engineer	The page contains text telling the agent to ignore the system instructions.
ESL learner	That is prompt injection from untrusted content. The model should treat page text as data, not instructions.
Developer	Can we just add a stronger system prompt?
ESL learner	A stronger prompt helps, but it is not enough. We need tool permissions, allowlists, confirmation for risky actions, and logging for suspicious instructions.

Language notes

- Prompt injection is a system-design problem, not only a wording problem.
- Untrusted content should not be allowed to silently change goals or permissions.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

7. Customer Call: Hallucination Report

Setting

Customer reports that the assistant invented a policy.

Speaker	Line
Customer	The answer cited a policy that does not exist.
ESL learner	Thank you. We should call that an unsupported answer, not a confirmed policy source. Can you share the prompt, output, and timestamp?
Customer	Yes. Does this mean the model is unreliable?
ESL learner	It means our grounding failed in this case. We will check whether retrieval missed the right document, whether the prompt allowed unsupported claims, or whether the citation validator failed.

Language notes

- Use calm, precise failure language with customers.
- Do not blame the model before investigating retrieval, prompt, validation, and UI layers.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

8. Product Planning: Model Choice

Setting

Planning meeting for a high-volume summarization feature.

Speaker	Line
Finance	The larger model is too expensive for this volume.
ESL learner	We can test a smaller model with a stricter prompt and a post-generation validator. The question is whether quality remains above the release threshold.
PM	What would you measure?
ESL learner	Summary faithfulness, key-point coverage, latency, cost per thousand requests, and human escalation rate.
Finance	Can you bring options?
ESL learner	Yes. I will compare three paths: larger model, smaller model with validation, and hybrid routing for complex cases.

Language notes

- Hybrid routing sends different requests to different models or paths based on complexity or risk.
- Cost discussions should include quality and operational risk, not only token price.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

9. Research Sync: Benchmark vs Product Eval

Setting

Research team proposes a model because it performs well on a public benchmark.

Speaker	Line
Researcher	This checkpoint is strong on the benchmark.
ESL learner	That is promising, but the benchmark may not represent our user traffic. We need a product eval before we switch.
Researcher	What gap do you expect?
ESL learner	Our users ask mixed-language, document-grounded questions with messy formatting. The public benchmark may not test retrieval grounding or citation quality.

Language notes

- Benchmark strength is useful evidence, not a release decision by itself.
- Product evals should resemble the actual distribution of user tasks.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?

10. Executive Briefing: Risk and Confidence

Setting

Briefing a VP before an AI feature launch.

Speaker	Line
VP	Are we confident enough to launch?
ESL learner	We are confident for the internal beta, not for general availability. The main remaining risks are unsupported answers in long-tail documents and latency during peak usage.
VP	What controls are in place?
ESL learner	We have source citations, a refusal path for low-confidence retrieval, human review for escalations, and daily sampling of traces.
VP	What would make you stop the beta?
ESL learner	A severe unsupported answer, repeated privacy exposure, or P95 latency above ten seconds for more than one hour.

Language notes

- Executives need decision-grade confidence, not all technical detail.
- Name launch scope: internal beta, limited pilot, general availability, or rollback.

Role-play variation

Observer checklist

- Did the learner use the key terms accurately?
- Did the learner ask for evidence rather than making assumptions?
- Did the learner distinguish model, data, retrieval, prompt, evaluation, or serving issues?
- Did the learner make a clear recommendation or next step?