

# Project Plan: Multi-Agent AI Presentation Generator & Beautifier

## Goal Alignment with Hackathon Criteria and Prizes

We are building a multi-agent presentation generator & beautifier using OpenAI's new open-weight models (gpt-oss-20b/120b). This aligns our project with the OpenAI Open Model Hackathon goals and prizes. In particular, we will target the "Most Useful Fine-Tune" category (sponsored by OpenAI, \$5,000) by fine-tuning gpt-oss on presentation-building tasks. We'll also design for a shot at "Best Overall" (\$10,000) by delivering a polished, impactful application.

The hackathon's judging criteria emphasize: effective use of gpt-oss, good design/UX (with user safety), high potential impact, and novelty . We have woven these into our plan:

- Application of gpt-oss: We will showcase the model's strengths uniquely – fine-tuning it for slide generation and leveraging its built-in agentic abilities (function calling, large context) . This ensures our solution isn't something any generic model could do .
- Design and UX: We plan a user-friendly interface with thoughtful workflow and safety checks, balancing a solid backend with a clean frontend .
- Impact: The tool addresses a broad need (fast, quality presentation creation), benefitting professionals, educators, and more. We'll highlight how it saves time and works seamlessly online – enabling access to the latest models, real-time data sources, and richer media generation for maximum accuracy and presentation quality.
- Novelty: Our multi-agent approach and fine-tuned open model concept is creative and unique – no existing product quite combines these. We're improving on known ideas (AI slide makers) with a new twist .

Sponsor Goals: OpenAI's hackathon prompt encourages "creative, unexpected" applications of these models – e.g. "fine-tune a model into something indispensable" . By fine-tuning gpt-oss for a specialized but widely useful task, we hit that mark. Hugging Face, as Best Overall sponsor, will appreciate an open-source community utility, and we plan to release our fine-tuned model on HF Hub.

## Competitive Benchmark

To gauge our fine-tuned model's success, we will compare it against several alternatives: the base gpt-oss-20b, state-of-the-art models like GPT-5, and other LLMs (e.g. GPT-4 or Claude). The goal is to demonstrate that fine-tuning yields a notable improvement in this task beyond what larger generic models or the base model can do.

- **Versus Base GPT-OSS-20B:** We expect a significant leap in performance after fine-tuning. The base model, when prompted to create slides, might output reasonable content but not in the structured slide format or with polished conciseness. Fine-tuning should drastically improve structured output fidelity and relevance. Research on slide generation supports this: a fine-tuned 8B model (AutoPresent) approached the performance of GPT-4 on slide tasks, whereas the base 8B model “could barely generate any slides” . By analogy, our fine-tuned 20B should far outperform the 20B base model on any slide-quality metric (structure, coherence, formatting). In practical terms, we anticipate the base model would require a lot of manual editing (or might even output a plain essay instead of slides), whereas the fine-tune will produce well-separated slides with titles and bullet points that need minimal editing. We will validate this with side-by-side examples – for instance, feeding the same document to base gpt-oss and to our fine-tuned model, and comparing results. Success means the fine-tuned output is clearly more “presentation-ready,” both to our eyes and via quantitative metrics (see Metrics section).
- **Versus GPT-5 / GPT-4:** GPT-5 is OpenAI's latest flagship model, extremely powerful generalist AI . It likely can generate presentations from a prompt (indeed, Microsoft 365 Copilot uses GPT-4 for transforming documents into slides). However, even top-tier models are not specifically fine-tuned for slide structuring. They might produce verbose slides or stray from a strict JSON schema unless carefully instructed. Our fine-tuned gpt-oss-20b, though smaller, is specialized for this one task. We hypothesize that on document-to-slide conversion, our model will be competitive with GPT-4/5 in quality, especially in adhering to structure. This isn't far-fetched: the AutoPresent project showed an 8B fine-tuned model could achieve “quality comparable to the closed-source GPT-4o model” on slide generation. In user studies, that fine-tuned model's slides were as preferred as GPT-4's in certain settings . With 20B parameters and a solid fine-tune, we could potentially approach GPT-5's performance in this niche domain. Of course, GPT-5 might still have an edge in raw language fluency or if asked extremely open-ended questions. But the value of our approach is specialization: judges will see that a well-targeted fine-tune can make an open model punch above its weight on a focused task – perhaps even rivaling much larger models' outputs for that task.
- **Versus Other LLMs:** We'll also consider comparisons with models like GPT-3.5/4 (via API) or Anthropic's Claude, and possibly other open models like LLaMA-2 (70B). These models can all be prompted to create slide outlines. However, none of them (to our knowledge) have been fine-tuned explicitly for slide generation. We expect our fine-tuned model to produce more consistent and structured results than prompting a generic model. For example, GPT-4 might sometimes generate slides in paragraph form

or include too much text per bullet if not prompted carefully, whereas our model will have learned the ideal format from training examples. Another point of comparison is formatting fidelity: we aim for JSON output (for easy ingestion into our app). Generic models might need complex prompt engineering to reliably output JSON with the correct schema. Our fine-tune will treat that as second nature, because JSON slide format will be part of the training. We can measure this by JSON validity rates – e.g. how often does the model output a correctly structured JSON for a slide deck without needing fixes. We expect near 100% from our fine-tune (because it's trained that way), versus a much lower success rate from an off-the-shelf model.

In summary, our fine-tuned gpt-oss-20b should clearly outperform the base model and offer comparable usefulness to top proprietary models in this domain. The competitive benchmark will be evidenced by: (a) side-by-side demo (base vs fine-tune on the same input), (b) possibly an example vs GPT-4 or GPT-5 output on the same topic (to show we're in the same ballpark of quality), and (c) any metrics we gather (e.g. slide structure score, user preference tests) showing the fine-tune's gains. We anticipate that judges and users will readily see that the fine-tune is not just marginally better, but transformative – turning a tedious multi-hour task into a quick AI-assisted draft, without needing the might of a 180B model. This is exactly the scenario where fine-tuning shines: a medium-sized open model, tuned well, beating or matching larger general models in a specialized, useful task .

## Metrics & Targets

To concretely measure our progress and validate usefulness, we will track a set of quantitative metrics. These will help demonstrate the fine-tune's impact in a measurable way:

- **Slide Format Accuracy:** This measures how well the model adheres to the required structured output (JSON with specific fields for each slide). We can define this as the percentage of generated outputs that are valid JSON and include all required components (e.g. a title and a bullet list for each slide, plus a visual/image hint). Our target is >95% format accuracy for the fine-tuned model. (In contrast, the base model might often deviate from format or require regex fixes). Essentially, we don't want the judge to ever see a malformed output. This metric is straightforward to compute automatically (parse the JSON; count success vs failures). Achieving near-perfect structured output reliability will underscore the fine-tune's robustness.
- **Editing Effort Reduction:** This is a "usefulness" metric: how much less post-generation editing is needed with the fine-tuned model versus the base model (or versus a naive approach). We might quantify this as "edit reduction %". For example, we can take a sample of test prompts/documents, generate slides with both base and fine-tuned models, and then simulate the editing process to make them presentation-ready. The difference could be measured in number of edits or time taken. If the base model output needs 20 edits and the fine-tuned needs 5, that's a 75% reduction. Our goal is to cut

editing by at least 50%. Another proxy: measure the length of the output or number of bullets – if the base output is overly verbose, a human would trim it; the fine-tuned should already be concise. This could tie into a “brevity score” (shorter is generally better as long as content is covered). Ultimately, we want to demonstrate that a user would spend far less time fixing a fine-tuned model’s slides. This kind of metric aligns with how Microsoft evaluated Copilot’s efficacy – e.g. by seeing if users keep the AI-generated slides or heavily modify them. We might even track a “presentation kept rate” internally: what fraction of generated slides are good enough to use with only minor tweaks. Target: a user would accept >80% of slides from our model as-is or with minor touch-ups, whereas with base model maybe only 30-40% of content would be kept without rewrite.

- Information Retention / Coverage: We should ensure the model isn’t achieving brevity by omitting key information. A metric here could be coverage score – does the slide deck cover the important points from the source document or prompt? To measure this, if we have reference summaries or outlines (especially since we’ll synthesize data with GPT-4, we can use those as a reference), we can use an automatic metric like ROUGE or embedding-based similarity for content. Alternatively, a manual evaluation: for each test document, mark X key points that should be in the slides, and see how many the model included. We aim for the fine-tune to capture a high percentage of the key ideas (say, >90% coverage of must-have points), whereas the base model might miss more or include fluff. SlidesBench (the research benchmark) uses reference-based metrics that check similarity to a ground-truth slide – we can draw inspiration from that. While we won’t have a perfect “ground truth” for arbitrary inputs, we will validate that the fine-tune isn’t sacrificing completeness.
- Slide Quality Score: Quality is somewhat subjective, but we can devise a simple rubric (possibly for human evaluation by our team or beta-users). Criteria might include structure (clear separation of ideas per slide), coherence (slides flow logically), conciseness (no overly long bullets), and correctness (no hallucinated facts). Each criterion can be rated 1-5, and we take an average. We expect the fine-tuned model to score higher on structure and coherence especially. If time permits, we’ll do a small user study – e.g. show 5 people two versions of a slide deck (one from base model, one from fine-tune, or one from fine-tune vs one from GPT-4) and ask which is better or to rate them. Our target is that fine-tuned slides are preferred in >80% of comparisons with the base model. This aligns with findings from the AutoPresent user study, where participants significantly preferred slides from the fine-tuned model over the base LLaMA slides. Even if we can’t do a large user study, a few anecdotal side-by-sides in the demo with qualitative feedback (“the fine-tuned model’s output is clearly more organized and on-point”) will strengthen our case.
- Efficiency Metrics: Given it’s a hackathon, we also consider the practicality: how fast is generation, can it handle typical input lengths? We plan to note the inference speed (e.g. X seconds for a Y-word document). While not a primary judged metric, if our fine-tuned model can generate a full slide deck in, say, 20 seconds, that’s powerful to mention



(especially compared to hours of human work). Also, the ability to run on reasonable hardware (since gpt-oss-20b can run on a single high-end GPU) could be a point: it's an open, efficient solution versus needing a cloud API.

- **Benchmark Scores (Stretch Goal):** If we have time, we might run our model on some established tasks. For example, SlidesBench provides automated metrics (element matching, content similarity, etc.) . We could take a subset of their test cases (if available) to see where we stand relative to the numbers reported for AutoPresent or GPT-4. This would be more for our knowledge and to cite in the presentation (to show we have academic grounding), rather than a requirement. A more direct benchmark could be comparing outputs with Microsoft Copilot on the same input (if we have access) to qualitatively show differences.

Our targets in summary: (1) Fine-tuned model consistently outputs correct slide JSON ( $\geq 95\%$  structured outputs). (2) Fine-tune reduces the editing/time required by at least half compared to base (or conversely, doubles the “usable content” generated). (3) Fine-tune slides capture essentially all key info from input (no significant omissions). (4) Fine-tune slides are subjectively preferred by users/judges over base model slides the vast majority of the time. We will report metrics like “X% reduction in edits” and show examples to validate them. By setting these concrete benchmarks, we can clearly demonstrate the value added by fine-tuning – it's not just a nifty idea, it's quantifiably better than the alternatives on this task.

## Key Features and Innovations




Our project's feature set is designed to maximize our odds of winning by hitting all the marks of an outstanding hackathon entry. Below are the key features, with notes on their implementation and how they address judging criteria:

-  **Fine-Tuned OpenAI Model (gpt-oss) for Slide Generation:** This is the core of our project. By fine-tuning gpt-oss on presentation data, we showcase the model's capabilities in a specialized, highly useful way – exactly what the “Most Useful Fine-Tune” prize is about . Fine-tuning will improve the quality of content (clear, on-point bullets) and ensure the model follows structured output formats. It also differentiates us from competitors who might just prompt GPT-4; we are investing in our own model specialization, which is novel and impressive. Technical detail: We'll likely use Hugging Face Transformers with the provided fine-tuning guide (which the hackathon resources link to) to train our model, possibly using parameter-efficient tuning (LoRA) to save time. This feature directly targets the Application of gpt-oss criterion – we're not just using the model, we're extending it in a unique way.
-  **Multi-Agent Architecture (Reasoning & Tool Use):** Rather than a single prompt-to-slides model, we implement multiple agents (as described) for a more robust and intelligent system. This design is innovative and leverages the model's agentic

features. Judges will appreciate that we're using "chain-of-thought and tool calling support" of gpt-oss to build a complex application, not just a trivial Q&A bot. The multi-agent approach also improves results (e.g., one agent's output becomes another's input, which is a form of scaffolded reasoning). It adds a "wow factor" – it's essentially an AutoGPT-like slide creator. For the demo, we might even show a quick snippet of the agents' conversation or plan (transparency can impress on novelty). This addresses Novelty of Idea and also Design, since our system design is quite thoughtful.

- 🎨 Automated Slide Design and Theming: We'll have built-in templates for slide design (color schemes, layouts) and let the AI apply them. The user can choose a theme or have the AI pick the best fit. This feature ensures the output is not just textually correct but also visually appealing and professional, which boosts our Design/UX score. It's one thing to spit out text; it's another to produce a nicely formatted deck ready for use. We saw competitors highlight this: Presentations.AI stresses "AI-powered customizable presentation templates" as a key feature, and we will deliver something similar. Implementing themes might involve HTML/CSS if we use a web-based viewer or using PowerPoint template files for export. Either way, it's a matter of mapping the AI's chosen style to a set of visual properties. This feature also allows potential branding (important for impact in business use-cases): e.g., user could input their company's brand colors and we adapt the theme, a capability that Presentations.AI markets heavily. Even if we just show a concept of it, mentioning it will signal we understand user needs in real-world scenarios (beyond hackathon).
- 🖼️ Integration of Visuals (Images/Icons/Media): Each slide will have relevant imagery or media generated or retrieved by the AI. This makes presentations more engaging. Technically, we'll integrate with image generation APIs (depending on what's available—since DALL·E 3 might require OpenAI API access, we might use Stable Diffusion locally or stock images which are free). We already have a plan to allow multiple providers (similar to Presenton's "Multi-Provider Support" for text and images). Visually rich output will make our demo stand out to judges (it's literally more eye-catching than plain bullet lists). It also shows we considered completeness – a presentation isn't just text. This contributes to Impact (slides with visuals are more effective for audiences) and shows a high level of effort in implementation (not trivial). We will need to handle this carefully within 31 days, but we can start with static relevant images (like use keyword search to fetch images) and upgrade to generative if time permits. Even a few illustrative images in the final output will significantly elevate the perceived quality.
- 🌐 Multi-Language Support: Our model will be instructed (and fine-tuned data permitting) to handle multiple languages for slide content. The UI will allow selecting the language of output. This is feasible because gpt-oss is presumably trained on a variety of languages (and if not, we can still attempt translation via the model). By including this, we broaden the project's impact – for example, a user can generate the same presentation in English and Spanish effortlessly. It aligns with "benefiting all of humanity"

goals by making knowledge more accessible. It's also a slight competitive edge: many existing tools focus on English. We can demo this feature by generating a slide deck in another language (if the judges are international, that's a plus). Technically, this means having training or prompting in other languages – we likely won't fine-tune separately per language, but we'll prompt the model (which is instruction-tuned) with "Respond in X language." And since we allow document input, a user could input a non-English document and get slides in that language (or even translated slides in another language). This global mindset can score points in Impact and possibly Novelty.

-  **Export and Compatibility:** As mentioned, we will support exporting to PPTX and PDF. This feature might seem logistical, but it's crucial for real-world usability – and the judges will value that we thought of it. It addresses the Design criterion in terms of user experience (the deliverable is immediately usable). We saw that being able to export to PowerPoint is a differentiator that even competitor comparisons talk about ("Tome does not support export to PowerPoint... Presentations.AI allows high-quality PowerPoint export" ). We'll implement this using an open-source library or by generating an OpenDocument/OOXML format. We will test the exported file on PowerPoint to ensure formatting is preserved (fonts, bullet indent, images placement, etc.). This feature will be highlighted in our demo – e.g., "Here's the PPT file generated – now open it in PowerPoint to show it works." It conveys a level of completeness that can impress judges (many hackathon projects stop at a web demo; we'll hand over a tangible output).
-  **Iterative Refinement & User Feedback Loop:** Unlike a one-shot generation, our tool allows iteration. The user can adjust the outline and request changes (e.g., "make slide 3 simpler," or "add a slide about X"). We might implement a simple chat-like interface for post-generation refinement, where the user types an instruction and the appropriate agent acts (similar to how one might prompt ChatGPT to modify its answer). This feature shows we prioritize user control and fine-tuning of results, which improves the Design score (a well-thought-out UX) and also addresses potential shortcomings of generative AI. It's also a subtle way to demonstrate the model's interactive abilities. Hackathon judges will appreciate that we didn't build a black-box but rather an assistive tool that cooperates with the user. In the demo, for instance, we can show editing one slide's text, or using a "regenerate" button. Having this fall-back for quality issues means we can still produce a great final presentation even if the initial generation had some quirks.
-  **Use of Model's Strengths (Context & Tools):** We specifically exploit features of gpt-oss that are unique or enhanced compared to other models:
  - **Large Context:** The ability to feed a large document (e.g. 50 pages of text) is a game-changer. We will highlight this by possibly showing an example of generating slides from a lengthy source (like a sample whitepaper). The model's "128k tokens context window" means it can "ingest large corpora" – we will push this to show something competitors with smaller contexts (like Llama-2 32k or GPT-3.5 16k) might miss. This directly shows off the Application of gpt-oss

criterion (“showcase the strengths of the model uniquely” ).

- Tool Usage / Function Calling: As noted, we’ll integrate at least one function (image search) and maybe a simple web search for facts if time allows. This shows that our project “applies the open models in an effective way” by using their agentic capabilities . For example, if the user’s document has outdated data, an agent could (optionally) do a quick web search for the latest stats and update a slide – demonstrating a live data integration. This is an ambitious stretch goal, so it might be something we mention (since the hackathon sponsors include vLLM who are interested in efficient tool use). Even if we only fully implement the image function, that itself is a proof of concept of model tool use.

All these features are chosen to maximize our hackathon scoring and create a compelling project. They also complement each other – fine-tuning improves base output, multi-agent structure improves reasoning and modularity, design features improve UX and polish, etc. The end result will be a project that feels complete and impactful in just one month, demonstrating a high level of skill and creativity from our team.

## Timeline and Milestones (31 Days)

To execute this plan in ~31 days, we’ll break the work into weekly sprints with clear milestones. Our team will likely work in parallel on model and app components, syncing often to integrate. Here’s our proposed timeline:

- Week 1: Setup and Research (Days 1–7)

Milestone: Properly equipped to train and run the model; initial project skeleton.

- Acquire & Prep Models: Download gpt-oss-20b weights from Hugging Face and set up the environment (we’ll use provided guides like “How to use gpt-oss with Transformers” for smooth setup). Test run a few prompts on the base model to gauge output style and performance on our hardware. If needed, install vLLM or Ollama for faster inference tests .
- Data Collection: Begin gathering training data for fine-tuning. Search for public slide decks with transcripts or speaker notes. Possible sources: SlideShare (if any text available), academic lecture slides vs papers (e.g., arXiv to slide conversions). Where direct data is lacking, use GPT-4 to generate synthetic pairs: e.g., feed it an article and ask for bullet-point slides. Aim to gather a few hundred examples to fine-tune on – focusing on quality over quantity given time. Also include different domains (tech, education, marketing) for generality.



- Design Brainstorm: Finalize the multi-agent design and how each agent will function. Decide on the interface architecture (likely a simple web app – maybe React or just HTML/JS + a Python backend via FastAPI or Flask to handle model calls). Sketch the UI screens (input form, outline view, slide viewer). Also define 3–5 theme styles we want to support (create basic CSS or PPT master slides for them).
- Task Assignment: If team has multiple members, split responsibilities: e.g. one focuses on model fine-tuning pipeline, one on front-end UI, one on building the PPT export logic, etc. Establish communication channels and version control (Git repo).
- Week 2: Prototype Development & Model Fine-Tuning (Days 8–14)

Milestone: Basic end-to-end generation with dummy data; fine-tuning job running.

- Model Fine-Tune Kickoff: Use the Hugging Face Transformers fine-tuning script (the hackathon resource “How to fine-tune gpt-oss using Transformers” will be handy). Start with gpt-oss-20b. We might do a LoRA fine-tuning to handle 20B on available GPUs (less memory intensive). Begin training on the dataset compiled in Week 1. This might take a few days to converge; we’ll monitor validation outputs (like see if bullet style is improving). If training is slow, consider smaller batches or use an online service with A100s for a few hours.
- UI & Front-End: Build the input form page and results page. Implement file upload handling and prompt input. For now, the “Generate” button can call a placeholder API that returns a hardcoded example (to develop the front-end independently of model work). Design the outline edit interface (e.g., list of slide titles with an edit button). Also implement the slide viewer (initially static content). We’ll ensure the front-end is modular so we can plug in real data easily when ready.
- Outline & Content Agents (Backend logic): Implement a basic version of the pipeline using the base model (since fine-tune may not be ready). For example, write a function that takes text input, uses gpt-oss (with a prompt template) to generate an outline (maybe using headings or JSON). Then another function to generate bullets for each outline point. This is a crude first cut, but by end of Week 2 we want to see a simple prompt result in some slide text. This will help us debug the logic and prompt formulations early. We’ll likely run this on a smaller model or short context to iterate quickly.
- Image Integration (Prototype): Choose one image source to test. For now, maybe use Pexels API (free stock photos). Write a small function `get_image(keyword)` that returns a representative image URL for a keyword. We can test this

separately (e.g. `get_image("climate change")` returns some relevant image). We won't integrate into the model's loop yet, but we lay the groundwork by obtaining API keys, etc.

- Team Checkpoint: By the end of week 2, have a meeting to demo the prototype: e.g., we input "Test topic" and see an outline or some dummy slides on the UI. Also review a sample output from the fine-tuning (if completed epoch1) to gauge if we need more data or adjustments. Adjust plan if something is lagging (e.g., if fine-tune is delayed, plan to rely more on prompt engineering; if UI is behind, simplify features, etc.).
- Week 3: Integration and Feature Completion (Days 15–21)

Milestone: Full system integrated – model, agents, UI – producing actual presentations.

- Integrate Fine-Tuned Model: By early Week 3, we expect to have a fine-tuned model ready (or at least a first version). Integrate it into our pipeline: load the fine-tuned weights for inference. Compare outputs: we'll run the same test prompt through base vs fine-tuned model to ensure fine-tuning helped (e.g., fine-tuned should output nicer bullets). We can cite this improvement in our submission (showing effectiveness).
- Refine Multi-Agent Pipeline: Now wire up all agent steps in the backend. After outline generation, feed outline into content generation agent automatically. Incorporate the image agent: for each slide content, extract a few keywords (maybe using an algorithm or another mini prompt) and call the image API to get an image link. Insert that into the slide data structure. Develop simple heuristics: e.g., if a slide title contains " vs. " or numbers, maybe it's a comparison slide – choose a different layout or icon. These rules can elevate the design aspect.
- Theme and Layout Implementation: Apply the chosen theme to the generated content. In a web app, this could mean adding CSS classes for the background and text styles. For PowerPoint, use a template PPTX where we can fill in text placeholders with our content. We can generate an XML or use Python-pptx to create slides, set the background color, etc., according to theme. By mid-week, test exporting a sample PPTX and opening it to verify formatting (this might require some tweaking).
- UI Polish – Outline Editor and Slide Viewer: Hook up the real data to the front-end. After generation, show the outline step if we haven't already. Let the user edit slide titles (implement this by simply taking their edits and regenerating content for those slides – or simpler, incorporate their titles into the next step). Ensure the slide viewer displays the text and images nicely (maybe a carousel or

a slide-by-slide view). Implement navigation controls if needed (next/prev slide).

- Multilingual & Other Options: Add the ability to choose language in the UI, and make sure our generation prompts respect it (e.g., include “in Spanish” in the system prompt if Spanish is selected). Quick test with a known text in another language to see if model outputs that language correctly. Also, implement the slide count option: perhaps adjust the outline agent prompt to generate exactly N slides if the user requested a number.
- Testing: Start testing the integrated system with various inputs:
  1. Short prompt (few words) vs long document.
  2. Technical content vs general topic vs narrative content (to see how it handles different styles).
  3. Different themes (ensure theme switching doesn't break layout).
  4. Edge cases: no images found for a weird topic, extremely large input (how is performance/memory).
  5. If any failure or low-quality output is observed, refine prompts or agent logic accordingly. We iterate now while there's still time to fix issues.
- Week 4: Optimization, Polish, and Demo Prep (Days 22–31)

Milestone: A polished, competition-ready project with documentation and demo.

- Performance Optimizations: Running a 20B model with multiple calls per generation might be slow. In week 4, we'll optimize:
  1. Possibly use vLLM for serving the model in memory efficiently (the hackathon sponsor vLLM provides a library for fast concurrent inference). This could speed up multi-agent calls by not reloading the model each time. If setup is complex, an alternative is to cache results for repeated runs or use a smaller model for non-critical agents (e.g., maybe use a 7B open model for the critique agent if fine).
  2. We may quantize the model further (if not using the provided MXFP4 quantization) to make it faster on our GPU. The HF blog notes the model is already quantized and can use FlashAttention etc. – we'll ensure those optimizations (like running on a machine with a recent GPU that supports it).

3. If generation is still too slow for a live demo, we'll prepare a cached example to show in the video (ensuring the video is smooth). But we'll also mention that with appropriate hardware (e.g., an NVIDIA 5090 GPU), it would be much faster – a nod to NVIDIA's sponsor interests.
- Full Run-through and UX Fine-tuning: Do complete test runs as if we are the end-user. Aim to fix any remaining rough edges:
    1. Ensure the UI is clear (add instructions or placeholder text like “Enter topic...”).
    2. Verify that after generation, the user can easily navigate the slides and trigger any re-generation for slide content if needed.
    3. Add a “Download PPT” and “Download PDF” button and test them.
    4. Implement any missing safety net (e.g., if the model somehow outputs a disallowed phrase, we filter or warn – though our domain is mostly safe).
    5. If possible, integrate a content fact-check step for extra impressiveness: e.g., use the model to double-check if any factual claim was made and perhaps append source info. (This is very advanced and likely skipped due to time, but even an attempt could be mentioned.)
  - Visual Appeal: As this is the final week, we'll spend time on the look and feel. Create a nice logo or landing screen for the app (small touches can make it memorable). Perhaps name the project (something catchy like “SlideWhiz” or “DeckGenie”) and include that in the UI. This helps with presentation and showing branding (plus looks good on our resume/project page).
  - Prepare Presentation & Video: This is crucial for hackathon submission. We'll create a short pitch deck (maybe ironically using our own tool to generate the first draft!) highlighting:
    1. The problem (making presentations is time-consuming or requires expensive tools).
    2. Our solution (multi-agent AI presenter) with its features.
    3. Quick demo of it in action.
    4. Why it's innovative (fine-tune, open-source, etc.).

## 5. Impact potential (time saved, wider access, etc.).

We'll then record a demo video (likely 2-3 minutes as required). The video will show the app interface: we'll walk through inputting a prompt/document, show the outline edit briefly, then show the final slides and maybe scroll through them. We'll point out how images were added and mention the model behind the scenes. We'll also explicitly call out how we used the open model and that it's running locally (e.g., show a terminal with it running to emphasize offline). The video should end with a strong note: e.g., "In 5 minutes, our AI co-pilot created a presentation that normally takes hours – imagine what this means for productivity and accessibility!" – leaving judges with a clear impression of impact.

- Devpost Submission Materials: Finalize the project write-up for Devpost. We'll include the required details, making sure to reference the judging criteria: e.g., Application of Model – describe fine-tuning and agent use; Design – describe our UI and user testing; Impact – include a hypothetical case study (like a teacher using it to prepare lessons, saving X hours weekly); Novelty – compare vs other tools (we can even reference that "no existing solution combines fine-tuned open models with multi-agent orchestration for presentations", which we have from our research). We will also acknowledge sponsors (mention using Hugging Face for model hosting, NVIDIA GPUs for running the model, etc., which subtly addresses those audiences).
- Buffer and Contingency: We'll reserve the last day or two for any unexpected issues or last-minute tweaks. Also, we plan to practice our pitch (if there's a live judging or Q&A) – be ready to answer technical questions (fine-tune process, how multi-agent works under the hood, etc.) and discuss future scope (like integrating this into MS Teams or adding voice-over generation as next steps, to show we have vision).

By following this schedule, we ensure we have a working prototype early and a refined product by the deadline. Regular testing and integration will mitigate the risk of last-minute failures. The timeline also shows our ability to plan and execute efficiently – something that reflects well both in the competition and on our resumes.

## Competitive Analysis & Differentiation

We expect top engineers in this hackathon, so it's likely other teams will also build impressive projects with gpt-oss. Here's why our project will stand out:

- Unique Combination of Techniques: Some teams might fine-tune the model for a specific domain (education tutor, medical Q&A, etc.), and others might build agent-based apps

(like a robot or a coder assistant). Our project does both – a fine-tuned model and an agent-based system – in a domain that is practical and demo-friendly. It's a new twist on the familiar "AI makes slides" idea, which hits the Wildcard ("unexpected use") note while still being broadly useful.

- **High Real-World Relevance:** Every knowledge worker has made slides; the utility is immediately clear. This could give us an edge on Potential Impact, as judges can instantly grasp how this could save time for many people. It's not a niche or whimsical demo – it's something that could be a product. In fact, sponsors like OpenAI and HF might see potential to showcase this as a success story of their open models enabling real productivity tools.
- **Polish and Presentation:** We are putting significant effort into the UI/UX and output quality. Many hackathon projects focus on the technical and have bare-bones interfaces; we'll deliver a professional-looking app and outputs. From our research, we know content quality and design matter (users/judges will notice if slides look sloppy or generic). By fine-tuning the model and refining the content, we avoid the pitfalls of generic AI output. And by including visuals and good formatting, our demo will simply look more impressive than a text-only console demo.
- **Open Source and Resume-worthy Engineering:** We will open source our code and (if allowed) our fine-tuned model. This demonstrates commitment to the community (which HF and OpenAI judges appreciate) and also means our project can live beyond the hackathon. From a resume perspective, this is great: we can show recruiters a GitHub repo of a working application, with documentation, and possibly a HuggingFace model link they can try. It's tangible proof of our skills in ML engineering, full-stack dev, and project management. Few hackathon projects achieve a level where they can be shared broadly; we aim to reach that level.
- **Alignment with Multiple Prize Categories:** Strategically, while our primary target is "Most Useful Fine-Tune" (OpenAI), our project ticks boxes in other categories too:
  - **Best Overall:** If we execute perfectly and have a bit of luck, we could contend for overall winner by scoring high in all criteria.
  - **For Humanity:** If we emphasize educational use (e.g., helping under-resourced schools create materials quickly, or non-profits preparing info decks), we can claim social benefit. We might not specifically aim for this prize, but it doesn't hurt to mention how our tool could help educators or community workshops, etc.

By positioning our project at the intersection of technical innovation and practical utility, we maximize our odds that at least one set of judges will champion it.

## **Resume & Career Highlights**

Beyond winning the hackathon, we also see this project as a stellar addition to our professional portfolio. Here's how it shines on a resume or in interviews:

- **Cutting-Edge Technologies:** We worked with OpenAI's latest open-source LLM (gpt-oss released 2025) and implemented fine-tuning, large-context processing, and AI agent orchestration. These are hot topics in AI – demonstrating experience in them will catch recruiters' eyes. We can truthfully say we built a mini-ChatGPT-like system but for a specialized task, including multi-model integration (LLM + image generation). This signals that we're not just users of AI APIs, but we deeply understand and can extend AI models.
- **End-to-End Project Delivery:** On our resume, this project shows full-stack capability: front-end web development, back-end server logic, and ML model training. It's a great example of bridging the gap between research and user-facing product. We can highlight specific accomplishments, e.g. "Fine-tuned a 20B-parameter language model to improve slide summarization by X%, and developed a React/Flask application for interactive use." Having metrics (like generation speed, or how much editing it reduced compared to normal) can also quantify the impact.
- **Teamwork and Agile Execution:** We delivered this in 31 days, which implies we can work under tight deadlines, divide tasks, and iterate quickly – all valuable in industry. If we lead the team, we can mention leadership; if we each took distinct roles (model specialist, UI developer, etc.), we highlight those contributions.
- **Open Source Contribution:** By releasing our project code or model, we add to our public profile. Employers often appreciate candidates who contribute to open-source projects. It also allows others to use or cite our project, increasing our visibility. We might even write a Medium article about our project (which, given the timely subject, could gain traction). This kind of initiative demonstrates passion and communication skills.
- **Hackathon Achievement:** If we win or place in the hackathon, that is a big resume boost in itself ("Winner of OpenAI's Open Model Hackathon 2025 – selected top out of 2000+ participants" sounds fantastic). Even if not, being a finalist or simply participating with a solid project is worth noting, especially given the prestigious sponsors involved. It shows we engage with the developer community and can hold our own in global competitions.
- **Demo-ready Project:** Post-hackathon, we'll have a polished demo that we can show in interviews or attach to portfolios. Many candidates talk about projects, but having a live demo where you can say "Type a topic, and watch what my AI builds" can leave a strong impression. It's the kind of project that can spark interesting discussions with interviewers (about how we did X or overcame Y challenge).

In summary, this project isn't just a hackathon entry – it's a demonstration of our abilities at the forefront of AI and software engineering. We will leverage it to open doors in career

opportunities, whether that's in AI research, product development, or startup endeavors. The fact that it addresses a real problem and has a potentially wide user base also suggests entrepreneurial potential (it could be turned into a startup or a product feature). So on a resume, we can emphasize both the innovation and the real-world relevance.

## **Conclusion: Maximizing Our Odds of Victory**

To conclude, our multi-agent presentation generator & beautifier project is carefully crafted to excel in the hackathon and beyond. By thoroughly researching the market and sponsor goals, we've ensured our idea is both novel and aligned with what judges seek. We will demonstrate:

- Superior use of gpt-oss: Fine-tuning and deploying it in a way that showcases its strengths (huge context, agentic tasks) that other projects likely won't . This directly addresses the hackathon's core challenge of applying these models in groundbreaking ways .
- Excellent design and polish: A smooth user experience with an interface that balances AI power with human control . We're not just showing a tech demo; we're showing a prototype of a product with real users in mind.
- Impact and usefulness: A solution that could save countless hours and empower those without design skills or big budgets. Our narrative will emphasize how this AI assistant can level the playing field (for example, a small startup can now create pitch decks as slick as a big company's, thanks to our tool). Judges considering impact will see that our project, if continued, could be used by many professionals, educators, and students around the world .
- Creativity and innovation: Combining multi-agent coordination with a fine-tuned model is unique. Even if others build slide generators, ours will have an extra "brain" behind it (multiple brains, in fact). We improve upon existing ideas (we've shown we know what's out there and how we're better), fulfilling the novelty criterion well .

We have also built in contingency and adaptability – if something doesn't work perfectly, our iterative refinement approach and user-in-the-loop design means we can still produce a great end result for the demo. This resilience will help ensure we can deliver on the day of submission.

Finally, our plan to present and pitch the project is as important as the project itself. We will make sure to tell a compelling story in our submission: one that highlights the technical achievement (training a model, orchestrating agents) and the practical achievement (making slide creation effortless). By clearly citing how we met each judging criterion and by delivering a visually engaging demo, we'll make it easy for judges to vote for us.



In essence, we aim to redefine what's possible with open models in the productivity space, echoing the hackathon's call to "redefine what open models can do" . If we execute this plan, our project will not only have a high chance of winning, but it will also be something we're proud to showcase long after the hackathon.

With 31 days of hard work and smart planning, we are confident that our multi-agent AI presentation builder can impress the judges and rise to the top – and in the process, become a shining example of our capabilities for our careers moving forward. We're excited to bring this idea to life and potentially secure a victory in the OpenAI Open Model Hackathon 2025!

Sources:

- OpenAI Open Model Hackathon rules and categories
- Hackathon prompt and sponsor goals
- Presenton open-source AI presentation tool features
- Presentations.AI vs Tome feature comparison
- HubSpot review of AI presentation makers (noting Tome's limitations)
- OpenAI GPT-OSS model capabilities (context window, tool use)
- Hugging Face blog on GPT-OSS (model size and usage) and OpenAI's announcement .