The concept involves an ABB GoFa cobot to perform a "sandwich assembly", serving as a viable showcase of our system's ability to translate natural language commands into structured physical tasks in a collaborative environment. The setup focuses on speech-to-intent, safety validation, trajectory generation, and execution while minimizing mechanical complexity through the use of simulated props.

The interaction begins when a user makes a natural language command, for example: "GoFa, make a classic sandwich on plate two, nice and neat." The system's Automatic Speech Recognition (ASR) and Large Language Model (LLM) processes this input to generate a plan of the task. Before execution, a simulation (via Unity) displays a preview of the motion. This allows the operator to verbally refine parameters, for example: saying "A bit slower on the second sandwich", before the robot executes assembly. Once confirmed, the robot retrieves ingredients and assembles them on the specified plate, leveraging the GoFa's integrated torque sensors to operate safely alongside human operators.

To ensure reliability and hygiene,, "ingredients" are modeled as uniform props, which are rigid 100mm x 100mm tiles fabricated from 3d-printed PLA or PETG. These tiles visually represent components like bread, protein, cheese, and lettuce but with consistent mechanical properties. Each tile features a 10mm vertical thickness and 45° chamfered top edges for easier grasping. The robot utilizes a soft-touch parallel gripper equipped with silicone-overmolded fingertips with a configured stroke width and low grip force to effectively handle fragile tiles securely without crushing them. By supporting distinct "recipes" (e.g., classic vs. veggie), the system provides semantic variety for the LLM to process without introducing complex physics of organic food. The workstation is divided into three pre-calibrated zones:

Ingredient Zone: Consists of a fixed tray with specific slots for each tile type.

Assembly Zone: A pad that provides a custom passive alignment fixture designed to mechanically contain the stack. The fixture uses guide walls to correct minor misalignment and includes relief slots that allow the gripper to grasp the completed stack from the bottom.

Serving Zone: Three numbered plates, allowing the user to specify destination. Since the position of every slot and the assembly fixture is calibrated offline, the system relies on a static location mapping. The stacking logic calculates vertical drop-off points by incrementing the height coordinate by 10mm for each layer relative to the fixture's surface.

The demonstration exercises the full Words2Motion pipeline. Intent extraction occurs as the LLM parses spoken audio into a JSON object defining the task type, recipe, target plate, and speed. Safety Validation is handled by a logic layer that verifies the request within physical constraints (e.g., maximum stack height) and workspace boundaries before passing the data to the motion planner.Trajectory generation produces cartesian paths for the cobot. If the user intervenes during the preview or execution phase with commands like "Stop" or "Change speed", the system dynamically updates the trajectory or halts immediately.