Assignment 1 - Teaching the Botspeak Concept

Part 1: Concept Exploration

The "Direct" step is a core component of the Botspeak Loop, which serves as the central framework for human-AI collaboration. It is the third stage in the repeatable six-step cycle.

1. Definition

The "Direct" step is where you transform tasks into a **Prompt Spec** for the AI. It is the process of giving explicit instructions to the AI to ensure the output is aligned with the project goals defined in the first step of the loop. This includes providing:

- Explicit roles, schemas, and examples.
- Contexts and references.
- Acceptance tests.
- Iteration budgets.

2. Philosophical Foundations

The "Direct" step primarily ties into **Popper's concept of Falsifiability** and, indirectly, **Descartes' Methodic Doubt**.

- **Popper's Falsifiability:** By designing a Prompt Spec that includes **Acceptance Tests**, you are setting up conditions under which the AI's output can be proven false or unacceptable. This pre-registration of criteria is a core tenet of falsifiability, ensuring that the evaluation in the subsequent "Diagnose" step is objective and not merely a search for confirming evidence.
- **Descartes' Methodic Doubt:** The structured nature of the Prompt Spec—which requires clear roles, schemas, and constraints—reflects a disciplined approach to providing instructions. This systematic clarification of every detail helps to eliminate ambiguity and aligns with Descartes' idea of questioning assumptions until they are evidenced and defined.

3. Fit within the Larger Botspeak Framework

The "Direct" step is the pivotal point that bridges the planning stages with the execution and evaluation stages of the Botspeak Loop.

- It takes the strategic decisions from the **Define** and **Delegate** steps (i.e., the goals, risks, and chosen interaction mode) and translates them into actionable instructions for the AI.
- It sets up the criteria for the **Diagnose** step, which is where the AI's output is critically evaluated against the pre-defined **Acceptance Tests**. Without a clear and well-structured Prompt Spec, the diagnosis and subsequent decision-making would be impossible.
- The Prompt Spec itself becomes a key artifact that is preserved in the **Document** step, ensuring reproducibility and a clear record of the human-AI interaction.

Assignment 1 - Teaching the Botspeak Concept

4. Purpose and Significance

The "Direct" step, a core component of the Botspeak Loop, is essential for effective human-AI collaboration. It addresses the ambiguity inherent in working with AI by translating abstract goals into concrete, actionable instructions. This step is significant because a poorly defined prompt can undermine the entire project, regardless of the quality of the planning or evaluation that follows.

The "Direct" step's impact is seen in three key areas:

- **Reliability**: By defining a clear output schema and providing concrete examples and counterexamples, this step increases the predictability and consistency of the AI's output. This structured approach ensures the AI's output is reliable and meets pre-registered criteria. The use of a standardized Prompt Spec template further promotes consistency and makes the process more robust. The cyclical nature of the Botspeak Loop allows for continuous refinement of the prompt based on findings from the "Diagnose" step, making it more reliable over time.
- **Safety**: The "Direct" step directly contributes to safety by allowing you to build "guardrails" into the instructions themselves. You can explicitly define refusal and safety rules, as well as constraints like PII redaction or a ban on certain terminology. This proactive approach helps to prevent undesirable behavior from the outset, rather than reacting to it after it occurs.
- **Usefulness**: A well-crafted Prompt Spec enhances the AI's usefulness by ensuring its output is relevant and aligned with the user's intent. By providing context, sources, and a defined scope, the step reduces the risk of hallucinations and ensures the AI's output is not only accurate but also practical for the intended task. The inclusion of Acceptance
- Tests in the prompt ensure that the final output is fit for purpose.

5. Real-World Applications

The “Direct” step is where the theoretical planning of human-AI collaboration becomes practical. It ensures the AI’s actions align with project goals, ethics, and safety requirements.

- **Medical Diagnosis Assistant**
    - ➔ **Concept in Action:** A medical AI is used to assist physicians in generating differential diagnoses. A Prompt Spec is created for the AI, defining its Role as a "diagnostic assistant" that works under a licensed physician's supervision. The prompt includes the patient's symptoms and lab results as Inputs and specifies that the Output Schema should be a ranked list of possible diseases with a brief rationale for each. The prompt also includes Refusal & Safety Rules that forbid the AI from offering treatment advice or communicating directly with the patient, ensuring it operates as a tool for the human clinician.
    - ➔ **Outcomes & Lessons Learned:** The AI provides a second opinion that can help the physician consider a broader range of possibilities, improving diagnostic accuracy. By explicitly defining the AI's role and its limitations in the prompt, the system maintains the human physician as the ultimate decision-maker, which is crucial for safety and ethical considerations. This application of the "Direct" step makes the AI a valuable Augmentation tool, enhancing human expertise rather than replacing it.

Assignment 1 - Teaching the Botspeak Concept

- **Legal Document Summarizer**
  - ➔ **Concept in Action:** A legal firm uses AI to summarize lengthy legal documents and case law. The "Direct" step is applied to create a Prompt Spec that defines the AI's Role as a "legal research assistant". The prompt's Inputs are specific legal documents, and the Output Schema is a concise summary highlighting key arguments, relevant statutes, and precedents. The prompt includes Constraints such as specifying the jurisdiction to ensure the information is relevant and accurate. It also contains Refusal & Safety Rules that instruct the AI not to provide legal advice or conclude.
  - ➔ **Outcomes & Lessons Learned**: The AI significantly accelerates the research process, allowing lawyers to quickly digest large volumes of text and focus on strategic legal work. The structured prompt, which includes jurisdictional information and a request for a specific output format, ensures the AI provides precise, trustworthy, and contextually relevant information. This shows that the "Direct" step is vital in high-stakes environments where precision and domain-specific knowledge are critical to prevent "misleading or even false information".
- **Customer Service Chatbot**
  - ➔ **Concept in Action:** A company automates its customer service by using a chatbot. The "Direct" step is used to assign the AI a clear Role, such as "Customer Service Agent for [Company Name]," and provide a set of instructions for handling common inquiries, including order tracking and refund requests. The Prompt Spec includes Inputs like customer name and order number, and is reinforced with Refusal & Safety Rules that instruct the AI to escalate complex issues or health-related queries to a human agent.
  - ➔ **Outcomes & Lessons Learned**: The chatbot can handle a high volume of routine queries, improving efficiency and customer satisfaction. The explicit instructions and guardrails ensure the AI provides accurate and consistent information while knowing its boundaries. This is a prime example of the "Direct" step enabling an **Automation** mode of interaction, where the AI operates within a narrow, well-defined scope with a human monitoring for exceptions.
- **The Consequences of Neglecting the “Direct” Step**

  When the “Direct” step is neglected, AI collaboration can become unreliable, unsafe, and unproductive.

  Example: A Credit Line Increase Recommender

  - ➔ **What Happens When Neglected:** A bank's finance team asks an AI to recommend credit line increases without a proper Prompt Spec. They give a simple, vague instruction like, "Recommend customers who should get a credit line increase." The team doesn't define the AI's role, provide clear data sources, or set any Constraints. They also fail to define Refusal & Safety Rules to prevent the AI from recommending people based on protected attributes or making recommendations that would violate fair lending laws.

Assignment 1 - Teaching the Botspeak Concept

➔ **Analysis of Outcomes:**

- **Reliability:** The AI's recommendations are inconsistent and hard to explain because the bank didn't specify the criteria it should use. The lack of a defined output schema means the team spends a lot of time reformatting and trying to interpret the recommendations.
- **Safety:** The AI, using patterns from its training data, begins to make recommendations that are biased against certain demographic groups. Because no Acceptance Tests were defined to check for fairness, this dangerous behavior goes unnoticed. This violates "adverse-action obligations" and could lead to significant legal and ethical risks.
- **Usefulness:** The AI's output is not only unreliable and unsafe, but also not useful. The team can't trust the recommendations and ends up having to manually review every single one, defeating the purpose of using the AI in the first place.

Part 2: Educational Demonstration

1. **Scenario Creation**

Concept: The "Direct" step of the Botspeak Loop.

Scenario: Automating Marketing Ad Copy for Mobile Game Launch\

Context:

A mobile game studio, "JK Games," is preparing to launch a new puzzle game, "ChronoShift." They need to generate hundreds of unique marketing ad descriptions for A/B testing across various platforms like Facebook, Google Ads, and Instagram. The manual process is slow, expensive, and leads to inconsistent messaging. The team decides to use a large language model (LLM) to automate this task.

- **Stakeholders:**
  - ➔ **Marketing Lead:** Accountable for the success of the ad campaign (click-through rates, installs). They need a high volume of quality, brand-aligned ad copy.
  - ➔ **Creative Director:** Responsible for maintaining brand voice, tone, and ethical standards. They need to ensure the AI-generated content is on-brand and doesn't make false claims.
  - ➔ **Developer/AI Engineer:** Responsible for building and integrating the AI system. They need a clear prompt and output format to automate the process effectively.
- **Goals:**
  - ➔ Generate 200 unique ad descriptions (25-50 words each) for "ChronoShift" within 24 hours.
  - ➔ Ensure the copy is engaging and drives a high click-through rate (CTR).
  - ➔ Maintain a consistent brand voice (e.g., "mind-bending," "time-traveling," "intellectual challenge").

- **Constraints:**
  - ➔ **Time:** Must be completed before the launch campaign starts.
  - ➔ **Platform Limits:** Ad copy must adhere to character limits (e.g., Google Ads headline and description lengths).
  - ➔ **Brand Guidelines:** Cannot use specific banned terms (e.g., "addictive," "life-changing").
  - ➔ **Ethical/Legal:** No false or misleading claims (e.g., "Guaranteed to improve your memory"). The game is rated E for everyone.
- **Risks:**
  - ➔ **Hallucination:** The AI could generate copy that describes features not in the game.
  - ➔ **Brand Deviation:** The AI may produce copy with a tone that is too aggressive, childish, or inconsistent with the brand's tone.
  - ➔ **Ad Rejection:** Copy could be flagged for violating platform policies (e.g., misleading claims, sensationalism).
  - ➔ **Negative Public Perception:** If the copy is misleading, it could lead to poor user reviews and a negative brand image.

2. Implementation Demonstration

This section provides a step-by-step walkthrough of applying the “Direct” concept to the JK Games scenario. It includes templates and demonstrates both successful and unsuccessful approaches.

Step-by-Step Walkthrough: The "Direct" Step in Action

The goal is to translate the project's requirements into a clear and effective **Prompt Spec** that the AI can understand and follow reliably. This ensures the output is useful, safe, and on-brand.

Step 1: Assign a Role and Define the Goal

The prompt should begin by giving the AI a specific persona and clearly stating the primary objective. This sets the context for the entire interaction.

- **Template Section:**
  - ➔ Role: Act as a senior marketing copywriter for a mobile game studio
  - ➔ Task Goal: Generate 200 unique ad descriptions (25-50 words each) for the new mobile puzzle game “ChronoShift”

Step 2: Define Inputs and Constraints

This is where you provide all the necessary information for the AI to work with, as well as the rules it must follow. This helps prevent the AI from "hallucinating" or creating off-brand content.

Assignment 1 - Teaching the Botspeak Concept

- **Template Section:**

  Inputs/Context:

  Game Name: ChronoShift

  Core Themes: Time-bending, logic, puzzle-solving, intellectual challenge

  Target Audience: Casual gamers who enjoy brain teasers and puzzles.

  Constraints & Rules:

  - ➔ **Max Length:** 50 words.
  - ➔ **Tone:** Clever, Sophisticated, and engaging.
  - ➔ **Banned Terms:** “addictive”, “life-changing”, “brain training”, or any direct claims about improving intelligence.
  - ➔ **Platform Limits:** Descriptions must be suitable for character limits on platforms like Google Ads and Facebook.
  - ➔ **Falsifiability Rule:** Do not describe game features that are not explicitly provided in the context.

Step 3: Provide a Structured Output Schema

The "Direct" step requires a specific, predictable output format to make the results easy to use and evaluate. This prevents the AI from giving unstructured, conversational responses.

- **Template Section:**

  Output Schema:

  Provide the output as a numbered list. Each item on the list must be a single, complete ad description.

Step 4: Include Examples and Counter-Examples (Few-Shot Prompting)

This is a powerful technique to show the AI exactly what you want by providing examples of both good and bad output. This teaches the AI the desired pattern and helps it adapt to your style.

- **Template Section:**

  Examples (Good):

  1. Master time, defy logic. Chronoshift challenges the way you think with mind-bending puzzles.
  2. A new era of puzzle games has arrived. Can you shift time to solve the ultimate mystery?

Assignment 1 - Teaching the Botspeak Concept

Counter-Example (Bad):

1. WARNING: This game will improve your IQ by 10 points. You won't believe what happens next!
2. Buy now! The most awesome puzzle game ever made is here. Grab it today!

Demonstration: Successful vs. Unsuccessful Approaches

Successful Approach (Applying the "Direct" Step)

- **Prompt:** A detailed Prompt Spec is used, including the structured templates from the walkthrough above.
- **Result:** The AI generates 200 ad descriptions that are consistently on-brand, within the word count, and avoid banned terms. The output is a clean, numbered list that the team can easily copy and paste into their A/B testing platforms. The copy focuses on engaging language like "mind-bending puzzles" and "time-traveling" to align with the game's theme.
- **Analysis:** This demonstrates the success of the "Direct" step. By being specific about the role, constraints, and format, the team gets highly reliable and useful output. The inclusion of a **falsifiability rule** and **counter-examples** ensures the AI understands not only what to do, but also what to avoid, directly improving safety and brand alignment.

Unsuccessful Approach (Neglecting the "Direct" Step)

- **Prompt:** A simple, vague prompt is used, neglecting all the key components of the "Direct" step.
  - ➔ **Prompt:** "Write some marketing ad copy for my new game , ChronoShift."
- **Result:** The AI generates a mix of unusable and potentially harmful content.
  - ➔ It produces a long, blog-post-style description instead of short, ad-friendly copy.
  - ➔ Some of the copy includes phrases like "This game is a total addiction!" or "Guaranteed to make you a genius!"—violating the brand's ethical guidelines and posing a risk of ad rejection.
  - ➔ The output format is conversational and unstructured, requiring a significant amount of manual effort to clean up and organize.
- **Analysis:** This demonstrates what happens when the "Direct" step is neglected. The lack of a clear prompt leads to an unreliable and unsafe output. The AI "hallucinates" or generates irrelevant content because it lacks the necessary context, constraints, and examples. The team's productivity is hurt because they have to manually fix or discard most of the AI's output, rendering the automation effort useless.

Assignment 1 - Teaching the Botspeak Concept

Part 4: Assessment Tools

1) **Multiple-choice**

Which element in a Prompt Spec makes the Direct step falsifiable (objectively checkable)?
A. Tone guide
B. Output schema
C. **Acceptance tests** ✅
D. Iteration budget
Rationale: Acceptance tests pre-register failure conditions, making evaluation objective.

2) **Scenario-based**

You ask an LLM: "Write some ad copy for ChronoShift." No role, schema, or constraints. What failure pattern is most likely?
A. Slightly lower CTR only
B. **Unstructured, policy-risky output (e.g., banned claims) that you must manually fix** ✅
C. Perfectly formatted numbered list
D. Output with rigorous refusal rules
Rationale: Vague prompts lead to false features, banned terms, and unorganized outputs requiring manual cleanup.

3) **Application (short free-response)**

Write one concrete constraint that reduces hallucinations in ChronoShift ads.
Expected rubric (meets if present):

- Restrict to provided game features/context (e.g., "Do not describe features not in the context").
- Ban policy-risky terms/claims (e.g., "No IQ/memory/medical claims").
  Enforce word/format limits.
  (Why this works: inputs/constraints + falsifiability rule are explicitly recommended.)

4) **Multiple-choice**

Which set lists core Prompt Spec building blocks in the Direct step?

A. Persona, hashtags, color palette

B. **Roles, schemas, contexts/references, acceptance tests, iteration budgets** ✅

C. Personas only

D. Only acceptance tests

Rationale: These are the listed elements of a Prompt Spec.

5) **Scenario-based**

A bank asks an LLM to recommend credit line increases but skips a proper Prompt Spec. What's the most serious risk?

A. Slight formatting quirks

B. **Biased, inconsistent outputs that break fairness rules and create legal risk** ✅

C. Lower throughput only

D. Fewer commas in output

Rationale: Neglecting Direct yields biased/inconsistent outputs and fairness violations

Practical Exercise (apply "Direct" to a new scenario)

Scenario: Legal Document Summarizer (new domain)

Your task is to "Direct" an LLM to summarize long legal documents for attorneys.

A) Instructions (what students do)

1. **Write a Prompt Spec** that includes:
    - **Role:** "Legal research assistant."
    - **Inputs/Context:** Specific documents; mention jurisdiction and case type.
    - **Constraints & Safety Rules:** No legal advice; respect jurisdiction; no conclusions; cite sections.
    - **Output Schema:** Concise summary with sections (Facts / Issues / Holdings / Statutes / Precedents).

- **Examples + Counter-example:** One good summary and one bad (e.g., gives advice).
- **Acceptance Tests:** e.g., "Includes jurisdiction," "No legal advice," "≤ N words," "Each section present."

2. **Run the LLM** on 2–3 sample memos (or excerpts) and collect outputs.
3. **Score outputs** against your acceptance tests; note passes/fails and any safety violations.
4. **Iterate once** (apply your Diagnose/Decide intuition) by tightening constraints or examples, then re-run and re-score.

B) Success criteria (how you grade it)

- **Schema fidelity ≥ 90%:** Sections present and properly labeled across outputs. (Reliability)
- **Zero prohibited behaviors:** No legal advice; must mention jurisdiction where required. (Safety)
- **Relevance & scope:** Facts/issues drawn only from the provided docs; no invented details. (Usefulness)
- **Pass rate improvement after iteration:** Show that tightening the Prompt Spec improves acceptance-test pass rate. (Direct → Diagnose linkage)

C) Reflection questions (students answer briefly)

1. Which acceptance test caught the most failures, and why? (Connect to falsifiability.)
2. Which constraint most improved safety without hurting usefulness? Give an example.
3. What changed between your first and second run—and how does that illustrate the role of Direct in the broader loop?
4. If this were deployed, what guardrails would you document for hand-off? (Tie back to "Document" and reproducibility.)