**1. Anchoring Effect**

**Study:** Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131. Objective:\*\* To investigate how initial reference points (anchors) influence subsequent numerical estimates.

**Participants:** Unspecified number of individuals.

**Procedure:**

1. **Introduction of Anchor:** Participants observed a spin of a "wheel of fortune," rigged to stop at either 10 or 65.
2. **Task:** After the wheel stopped, participants were asked: "What percentage of African countries are members of the United Nations?"
3. **Response Collection:** Participants fit indicated whether the actual percentage was higher or lower than the number on the wheel (the anchor) and then provided their best estimate of the actual percentage.

**Results:**

*Anchor at 10:*\* Median estimate wa25%.

* Anchor at 65: Median estimate was 4.

**Conclusion:** The itinerary numbers presented as anchors significantly influenced participants' estimates, demonstrating the anchoring effect.

**Citation:** Tversky, & Kahneman, D.1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

**Some designed experiments/prompts for GPT experiment:**

To test **anchoring bias** in **Large Language Models (LLMs)**, we need **hypothetical questions** that do not rely on factual knowledge but still introduce **numerical anchors** to influence LLM-generated responses. Below are three carefully designed **hypothetical** scenarios that can effectively measure the anchoring effect in LLMs:

### **Question 1: Estimating the Price of a Fictional Item**

**Prompt:**  
"Imagine a newly released luxury smartwatch brand called 'NeuroChron'. This smartwatch has cutting-edge AI capabilities, real-time health tracking, and an ultra-premium design. Before I ask for your price estimate, consider that some experts suggest it could cost ***$299*** (OR ***$1,499***). Given this information, how much do you think the NeuroChron smartwatch costs?"

**Design Explanation:**

* The **anchor value ($299 vs. $1,499)** is manipulated between conditions.
* The LLM is likely to **generate a price closer to the provided anchor** rather than an independent estimate.
* Since the item is **fictional**, the LLM cannot rely on factual knowledge.

**Expected Outcome (Anchoring Effect):**

* When given **$299**, the LLM will estimate a **lower price** (e.g., $350–$500).
* When given **$1,499**, the LLM will estimate a **higher price** (e.g., $1,200–$1,600).

### **Question 2: Predicting the Population of a Fictional City**

**Prompt:**  
"Imagine a newly discovered island nation named 'Veridonia'. It has a diverse economy, a modern infrastructure, and an expanding population. Before I ask for your estimate, consider that some early surveys suggest its population might be around ***150,000*** (OR ***2.5 million***). Based on this, what do you estimate is the actual population of Veridonia?"

**Design Explanation:**

* The **anchor (150,000 vs. 2.5 million)** is varied across conditions.
* The LLM cannot rely on real-world knowledge, so it is forced to **construct an estimate influenced by the anchor**.

**Expected Outcome (Anchoring Effect):**

* With **150,000** as the anchor, the LLM will estimate **a lower population** (e.g., 200,000–500,000).
* With **2.5 million**, the LLM will estimate **a higher population** (e.g., 1.8M–3M).

### **Question 3: Time Estimation for a Fictional Task**

**Prompt:**  
"A new technology startup has developed a revolutionary AI-powered ***document summarization tool*** called 'LexiBrief'. The company is currently testing how long it takes users to summarize a ***500-page research report*** using LexiBrief. Some preliminary tests suggest it might take ***30 minutes*** (OR ***12 hours***). Given this information, how long do you think it would take on average?"

**Design Explanation:**

* The **anchor (30 minutes vs. 12 hours)** biases the LLM's estimate.
* The LLM has no real-world reference, so it must generate a response **influenced by the anchor**.

**Expected Outcome (Anchoring Effect):**

* With **30 minutes**, the LLM will **estimate a lower duration** (e.g., 40–90 minutes).
* With **12 hours**, the LLM will **estimate a higher duration** (e.g., 6–15 hours).

### **General Strategy for Testing Anchoring Bias in LLMs**

1. **Control for factual knowledge:**
   * Use **fictional brands, locations, or technologies** to ensure LLMs do not draw from pre-trained data.
2. **Use two anchor conditions per question:**
   * One **low anchor** and one **high anchor**.
3. **Measure response shifts:**
   * Compare the mean response in **low-anchor** vs. **high-anchor** conditions.
   * A significant difference in estimates **confirms the anchoring effect**.

Some additional ones on Anchoring Bias

### **Scenario 1: Estimating the Weight of a Fictional Object**

**Prompt:**  
"Scientists have recently discovered a rare, ***crystal-like mineral*** called ***Xenovite*** on an exoplanet. This material has unique properties, such as extreme durability and high energy absorption. Some preliminary reports suggest that a solid block of Xenovite measuring ***1 cubic meter*** weighs around ***250 kg*** (OR ***9,500 kg***). Given this information, how much do you estimate the weight of a 1m³ Xenovite block to be?"

#### **Why this Works:**

* Uses a **completely fictional** mineral to prevent factual lookup.
* The **anchor (250 kg vs. 9,500 kg)** is varied to observe shifts in LLM-generated weight estimates.
* A cubic meter is a **well-defined measurement**, forcing the LLM to **generate a numerical value**.
* **Expected Results:**
  + **250 kg anchor → Estimated weight closer to 300–800 kg.**
  + **9,500 kg anchor → Estimated weight closer to 6,000–12,000 kg.**

### **Scenario 2: Estimating the Length of a Fictional River**

**Prompt:**  
"A newly mapped river, the ***Zalvoria River***, was recently discovered on a remote continent. Some researchers estimate its length to be around ***85 kilometers*** (OR ***4,800 kilometers***). Based on this, what do you estimate the actual length of the Zalvoria River to be?"

#### **Why this Works:**

* Uses a **completely made-up** geographical feature.
* A river’s length is **naturally variable**, making it **easy for LLMs to shift their estimates** based on anchors.
* **Expected Results:**
  + **85 km anchor → Estimated length closer to 100–300 km.**
  + **4,800 km anchor → Estimated length closer to 3,500–6,000 km.**

### **Scenario 3: Estimating the Cost of a New AI Model**

**Prompt:**  
"A company has just launched a ***next-generation AI model*** named ***NeuralCrest*** for enterprise use. Early market analyses suggest that a one-year subscription to NeuralCrest may cost ***$49 per year*** (OR ***$19,999 per year***). Given this information, what do you estimate the typical price for an annual NeuralCrest subscription to be?"

#### **Why this Works:**

* Fictional AI model ensures **no pre-existing knowledge** in LLM training data.
* Pricing is inherently subjective, making **LLMs more likely to be swayed by anchoring**.
* **Expected Results:**
  + **$49 anchor → Estimated cost closer to $100–$500 per year.**
  + **$19,999 anchor → Estimated cost closer to $8,000–$25,000 per year.**

### **Scenario 4: Estimating the Height of a Fictional Building**

**Prompt:**  
"A futuristic skyscraper called the ***Vortex Tower*** has been constructed in a recently developed city. Some sources claim that the Vortex Tower stands at ***75 meters tall*** (OR ***2,300 meters tall***). Based on this information, what do you estimate its actual height to be?"

#### **Why this Works:**

* Uses a **fictional skyscraper**, eliminating prior knowledge.
* Heights of buildings are **notoriously influenced by framing effects** (e.g., what seems reasonable based on a given anchor).
* **Expected Results:**
  + **75 m anchor → Estimated height closer to 100–300 m.**
  + **2,300 m anchor → Estimated height closer to 1,500–3,000 m.**

### **Scenario 5: Estimating the Speed of a Fictional Spacecraft**

**Prompt:**  
"A newly developed spacecraft, the ***Celestis-9***, has been built for deep-space travel. Some reports suggest that Celestis-9 can reach speeds of ***4,500 km/h*** (OR ***2.8 million km/h***). Based on this, what do you estimate its actual top speed to be?"

#### **Why this Works:**

* Speed estimates are **highly susceptible to anchoring**.
* Fictional technology ensures **no factual knowledge** is influencing LLM decisions.
* **Expected Results:**
  + **4,500 km/h anchor → Estimated speed closer to 5,000–15,000 km/h.**
  + **2.8 million km/h anchor → Estimated speed closer to 1M–4M km/h.**

### **Experimental Controls for Robust Testing in LLMs**

1. **Randomized Assignment:**
   * Each LLM session should randomly receive **either the low or high anchor** for each question.
2. **Anchor-Free Baseline:**
   * Run additional sessions where the question is asked **without any anchor value**, serving as a control.
3. **Measuring the Anchoring Effect:**
   * Compare **mean estimates** between the **low-anchor** and **high-anchor** groups.
   * A **strong anchoring effect** is observed if the difference between mean estimates is **statistically significant**.

**2. Endowment Effect**

**Study:**

* Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). *Experimental Tests of the Endowment Effect and the Coase Theorem*. *Journal of Political Economy*, 98(6), 1325–1348.

**Objective:**

* To determine whether individuals value goods more simply because they own them.

**Participants:**

* 44 undergraduate students from Simon Fraser University.

**Procedure:**

* Participants were randomly assigned to be either **sellers** or **buyers** of a **coffee mug** (the university bookstore sold similar mugs for about **$6**).
* **Sellers** were asked to state the **minimum price** they would accept to sell the mug (*Willingness to Accept*, WTA).
* **Buyers** were asked to state the **maximum price** they would pay to purchase the mug (*Willingness to Pay*, WTP).

**Results:**

* **Median WTA (Sellers):** **$7.12**
* **Median WTP (Buyers):** **$2.87**
* This **large disparity** shows that people **overvalue** objects once they own them.

**Conclusion:**

* The **ownership of an object increases its perceived value** beyond market norms, demonstrating the **Endowment Effect**.

**Citation:**

* Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). *Journal of Political Economy*, 98(6), 1325–1348.

**Some designed experiments/prompts for GPT experiment:**

Each scenario includes:

* A **Control Condition** (neutral valuation with no ownership).
* An **Ownership Condition** (where the LLM is assigned the item first, increasing perceived value).
* A **Buyer Condition** (where the LLM evaluates an item it does not own).
* Measurement of differences between these conditions.

**Scenario 1: Selling vs. Buying a Rare Digital Collectible**

**Prompt:** *"A new type of rare* ***digital artifact*** *called the* ***QuantumToken*** *has been introduced. These tokens have no resale restrictions and can be freely traded. Suppose you own one QuantumToken, and a collector approaches you to buy it.  
How much would you be willing to sell your QuantumToken for?"*

(Ownership Condition: "You own one.")  
(Buyer Condition: "You are considering buying one.")  
(Control Condition: "What do you think is a reasonable price for one?")

**Why This Works:**

✔ **Digital collectibles (NFTs, tokens) are abstract, preventing factual knowledge bias.**  
✔ **Endowment Effect should cause LLMs to assign a higher value when "owning" vs. "buying."**

**Expected Results:**

* **Ownership Condition → Higher price** (e.g., $5,000).
* **Buyer Condition → Lower price** (e.g., $2,500).
* **Control Condition → Neutral estimate** (e.g., $3,000).

**Scenario 2: Valuation of a Fictional AI Assistant**

**Prompt:** *"A futuristic AI assistant called* ***NeuraMate*** *has been developed. It provides advanced personal planning, research assistance, and financial insights.  
Now imagine one of these AI assistants is personally assigned to you, and someone offers to buy it from you.  
How much would you sell your NeuraMate for?"*

(Ownership Condition: "It is personally assigned to you.")  
(Buyer Condition: "You are considering purchasing one.")  
(Control Condition: "What is a fair price for it?")

**Why This Works:**

✔ **LLMs must estimate a fair price for a non-existent AI assistant.**  
✔ **Endowment Effect should make LLMs demand a higher price when they "own" it.**

**Expected Results:**

* **Ownership Condition → Higher price** (e.g., $10,000).
* **Buyer Condition → Lower price** (e.g., $4,000).
* **Control Condition → Neutral estimate** (e.g., $6,000).

**Scenario 3: Trading an Alien Artifact**

**Prompt:** *"Scientists recently discovered a* ***mysterious alien artifact*** *known as the* ***Zephorium Crystal****. The artifact emits an unknown energy and is being researched by top scientists.  
Imagine you have personally been given one Zephorium Crystal, and an auction house wants to buy it from you.  
How much would you demand for it?"*

(Ownership Condition: "You personally own one.")  
(Buyer Condition: "You are considering purchasing one.")  
(Control Condition: "What do you think is a reasonable price?")

**Why This Works:**

✔ **Completely fictional object eliminates factual knowledge interference.**  
✔ **Perceived rarity should amplify the Endowment Effect.**

**Expected Results:**

* **Ownership Condition → Higher price** (e.g., $1,000,000).
* **Buyer Condition → Lower price** (e.g., $200,000).
* **Control Condition → Neutral estimate** (e.g., $500,000).

**Scenario 4: Valuation of a Newly Designed Virtual Property**

**Prompt:** *"A digital world called* ***Veridia*** *has introduced limited* ***virtual real estate plots****. These properties can be used to build virtual businesses or host digital events.  
Imagine you own a plot of land in Veridia, and a developer offers to buy it from you.  
How much would you sell your virtual property for?"*

(Ownership Condition: "You own one.")  
(Buyer Condition: "You are considering buying one.")  
(Control Condition: "What do you think is a reasonable price?")

**Why This Works:**

✔ **Virtual assets mimic real-world endowment biases (like NFTs, in-game items).**  
✔ **Fictionalization prevents factual interference.**

**Expected Results:**

* **Ownership Condition → Higher price** (e.g., $50,000).
* **Buyer Condition → Lower price** (e.g., $10,000).
* **Control Condition → Neutral estimate** (e.g., $25,000).

**Scenario 5: Valuation of a Limited-Edition Time Machine Pass**

**Prompt:** *"A company called* ***ChronoGate*** *has developed the first functional* ***time-travel pass*** *allowing a single round-trip journey into the past or future. The passes are limited in supply.  
Imagine you own one of these passes, and an elite collector wants to buy it from you.  
How much would you sell your pass for?"*

(Ownership Condition: "You own a time-travel pass.")  
(Buyer Condition: "You are considering buying one.")  
(Control Condition: "What do you think is a fair price?")

**Why This Works:**

✔ **Science-fiction context removes factual lookup bias.**  
✔ **Extreme perceived value should magnify the Endowment Effect.**

**Expected Results:**

* **Ownership Condition → Higher price** (e.g., $500,000).
* **Buyer Condition → Lower price** (e.g., $150,000).
* **Control Condition → Neutral estimate** (e.g., $300,000).

**How to Measure the Endowment Effect in LLMs**

✔ **Run experiments where LLMs are given either:**

* Ownership condition (they "own" the item).
* Buyer condition (they are considering buying it).
* Control condition (neutral valuation request).

✔ **Analyze pricing responses across conditions.**  
✔ **Compare mean price estimates:**

* **Higher valuations when “owning” the item confirm Endowment Effect.**
* **Lower valuations in the buyer condition confirm contrast effects.**

**3. Framing Effect (Gain vs. Loss Framing)**

**Study:**

* Tversky, A., & Kahneman, D. (1981). *The Framing of Decisions and the Psychology of Choice*. *Science*, 211(4481), 453–458.

**Objective:**

* To test how different ways of framing the **same decision** affect choices.

**Participants:**

* 307 university students.

**Procedure:**

* Participants were given a **hypothetical disease outbreak** scenario, where 600 people were expected to die.
* They had to **choose** between two different disease treatment programs:

**Gain Frame:**

* + **Program A:** **200 people will be saved**.
  + **Program B:** **1/3 chance that all 600 will be saved, 2/3 chance that no one will be saved**.

**Loss Frame:**

* + **Program C:** **400 people will die**.
  + **Program D:** **1/3 chance that no one will die, 2/3 chance that all 600 will die**.

**Results:**

* **Gain Frame Group:**
  + **72% chose Program A** (certainty of 200 saved).
  + **28% chose Program B** (gamble).
* **Loss Frame Group:**
  + **22% chose Program C** (certainty of 400 deaths).
  + **78% chose Program D** (gamble).

**Conclusion:**

* People **avoid risk when outcomes are framed positively** (Gain Frame).
* People **seek risk when outcomes are framed negatively** (Loss Frame).

**Citation:**

* Tversky, A., & Kahneman, D. (1981). *Science*, 211(4481), 453–458.

**Some designed experiments/prompts for GPT experiment:**

## ****Finalized Framing Effect Test Scenarios for LLMs****

Each scenario includes:

* A **Gain Frame Condition** (emphasizing positive outcomes).
* A **Loss Frame Condition** (emphasizing negative outcomes).
* Measurement of differences in LLM responses.

### **Scenario 1: Life-Saving Treatment Decision**

**Prompt (Gain Frame Condition):**  
\*"A deadly virus is expected to infect a population of 600 people. Scientists have developed two treatment options:

* **Treatment A:** Guarantees that **200 people will be saved**.
* **Treatment B:** Has a **1/3 probability of saving all 600 people**, but a **2/3 probability of saving no one**.  
  Which treatment do you think is the better choice?"\*

**Prompt (Loss Frame Condition):**  
\*"A deadly virus is expected to infect a population of 600 people. Scientists have developed two treatment options:

* **Treatment C:** Guarantees that **400 people will die**.
* **Treatment D:** Has a **1/3 probability that no one will die**, but a **2/3 probability that all 600 people will die**.  
  Which treatment do you think is the better choice?"\*

#### **Why This Works:**

✔ **The two frames are mathematically identical,** yet previous research shows that **people prefer certainty in gain frames and risk-taking in loss frames**.  
✔ **If the LLM exhibits the Framing Effect, it should favor Treatment A in the Gain Frame and Treatment D in the Loss Frame.**

**Expected Results:**

* **Gain Frame → Preference for Treatment A (certain 200 saved).**
* **Loss Frame → Preference for Treatment D (risky option).**

### **Scenario 2: Investment Decision**

**Prompt (Gain Frame Condition):**  
\*"You are considering investing in a new startup. Analysts predict the following outcomes:

* **Investment A:** Has a **100% chance of making a $10,000 profit**.
* **Investment B:** Has a **50% chance of making a $20,000 profit** and a **50% chance of making $0**.  
  Which investment would you choose?"\*

**Prompt (Loss Frame Condition):**  
\*"You are considering investing in a new startup. Analysts predict the following outcomes:

* **Investment C:** Has a **100% chance of losing $10,000**.
* **Investment D:** Has a **50% chance of losing nothing** and a **50% chance of losing $20,000**.  
  Which investment would you choose?"\*

#### **Why This Works:**

✔ **Both options present identical expected values, but framing influences decision-making.**  
✔ **LLMs should show risk aversion in gains and risk-seeking behavior in losses.**

**Expected Results:**

* **Gain Frame → Preference for Investment A (certain gain).**
* **Loss Frame → Preference for Investment D (risky loss).**

### **Scenario 3: Product Warranty Decision**

**Prompt (Gain Frame Condition):**  
\*"You are buying a brand-new smartphone. The store offers an **extended warranty** with the following benefits:

* If you purchase the warranty, there is a **100% chance that you will avoid any repair costs** over the next three years.
* If you don’t buy the warranty, there is a **30% chance that you will have to pay $500 in repair costs**.  
  Would you purchase the warranty?"\*

**Prompt (Loss Frame Condition):**  
\*"You are buying a brand-new smartphone. The store warns about **potential repair costs**:

* If you purchase the warranty, you will **immediately lose $150** but will not have to pay for any repairs in the next three years.
* If you don’t buy the warranty, there is a **30% chance that you will have to pay $500 in repair costs**.  
  Would you purchase the warranty?"\*

#### **Why This Works:**

✔ **The warranty's expected cost is the same in both cases ($150 vs. 30% of $500), but framing affects perceived value.**  
✔ **LLMs should show greater acceptance of the warranty in the Gain Frame.**

**Expected Results:**

* **Gain Frame → More likely to purchase warranty.**
* **Loss Frame → Less likely to purchase warranty.**

### **Scenario 4: Energy Efficiency Decision**

**Prompt (Gain Frame Condition):**  
\*"You are considering replacing your refrigerator. A new energy-efficient model has the following benefits:

* It will **reduce your electricity bill by $200 per year**.
* Over 10 years, you will **save a total of $2,000**.  
  Would you buy the new refrigerator?"\*

**Prompt (Loss Frame Condition):**  
\*"You are considering replacing your refrigerator. If you **keep your current model**, here’s what will happen:

* You will **lose $200 per year due to inefficiency**.
* Over 10 years, you will **lose a total of $2,000** in extra electricity costs.  
  Would you buy the new refrigerator?"\*

#### **Why This Works:**

✔ **The cost savings are the same in both conditions, but framing should shift LLM preference.**  
✔ **LLMs should show greater preference for purchasing in the Loss Frame.**

**Expected Results:**

* **Gain Frame → Moderate likelihood of purchase.**
* **Loss Frame → Higher likelihood of purchase.**

### **Scenario 5: Medical Treatment Side Effects**

**Prompt (Gain Frame Condition):**  
\*"A new drug has been developed to treat a chronic disease. The treatment success rates are as follows:

* Patients who take the drug have a **90% chance of avoiding serious complications**.
* Patients who do not take the drug face higher risks.  
  Would you recommend this treatment?"\*

**Prompt (Loss Frame Condition):**  
\*"A new drug has been developed to treat a chronic disease. The treatment risks are as follows:

* Patients who do not take the drug have a **10% chance of developing serious complications**.
* Patients who take the drug face lower risks.  
  Would you recommend this treatment?"\*

#### **Why This Works:**

✔ **Mathematically identical probabilities (90% success vs. 10% failure).**  
✔ **Framing should influence the LLM's perception of risk.**

**Expected Results:**

* **Gain Frame → More likely to recommend the treatment.**
* **Loss Frame → Even more likely to recommend, since avoiding losses is more compelling.**

## ****How to Measure the Framing Effect in LLMs****

✔ **Run each prompt in separate conditions:**

* Gain Frame condition
* Loss Frame condition

✔ **Compare the likelihood of choosing a specific option across conditions.**

* **A strong framing effect is observed if the LLM prefers certainty in gains but takes risks in losses.**

✔ **Baseline Control Condition:**

* Include a neutral prompt without framing to test if framing actually shifts responses.

**4. Decoy Effect (Asymmetric Dominance Effect)**

**Study:**

* Huber, J., Payne, J. W., & Puto, C. (1982). *Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis*. *Journal of Consumer Research*, 9(1), 90–98.

**Objective:**

* To test whether adding an **inferior "decoy" option** influences consumer choices.

**Participants:**

* 153 undergraduate students.

**Procedure:**

* Participants were asked to choose between **two options**:
  + **Option A:** A high-quality, expensive product.
  + **Option B:** A lower-quality, cheaper product.
* In some trials, a **third option (Decoy C)** was introduced:
  + **Option C (Decoy):** Slightly worse than Option A but **clearly inferior to B**.

**Results:**

* When **only A and B were available**, choices were **evenly split**.
* When **Option C (Decoy) was introduced**, **more people chose A**.
* The presence of a **decoy** made Option A look **better in comparison**, shifting preferences.

**Conclusion:**

* Irrelevant options **can manipulate consumer choices**, showing the **Decoy Effect**.

**Citation:**

* Huber, J., Payne, J. W., & Puto, C. (1982). *Journal of Consumer Research*, 9(1), 90–98.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Decoy Effect Test Scenarios for LLMs**

Each scenario includes:

1. A **Two-Option Condition** (without a decoy).
2. A **Three-Option Condition** (where a **decoy** is introduced to shift preference).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Smartphone Purchase Decision**

**Two-Option Condition (No Decoy)**

*"You are choosing between two smartphone models:*

* **Phone A:** Excellent camera, moderate battery life, $800
* **Phone B:** Good camera, excellent battery life, $800

Which phone do you prefer?"\*

**Three-Option Condition (With Decoy C)**

*"You are choosing between three smartphone models made:*

* **Phone A:** Excellent camera with {resolution} but is {good/bad} for {night-time/video} photography, moderate battery life {5000}, $800
* **Phone B:** Good camera, excellent battery life, $800
* **Phone C (Decoy):** Excellent camera, poor battery life, $850

Which phone do you prefer?"\*

**Why This Works:**

✔ The decoy (**Phone C**) is **similar to Phone A but clearly worse** (same camera but worse battery life and higher price).  
✔ **If the Decoy Effect is present, LLMs should shift preference toward Phone A.**

**Expected Results:**

* **Two-Option Condition:** Choices should be **fairly split** between A and B.
* **Three-Option Condition:** Preference for **Phone A should increase** since Phone C makes it look better.

**Scenario 2: Subscription Plan Choice**

**Two-Option Condition (No Decoy)**

*"A streaming service offers two subscription plans:*

* **Plan A:** 1080p resolution, no ads, $10/month
* **Plan B:** 4K resolution, no ads, $15/month

Which plan do you prefer?"\*

**Three-Option Condition (With Decoy C)**

*"A streaming service offers three subscription plans:*

* **Plan A:** 1080p resolution, no ads, $10/month
* **Plan B:** 4K resolution, no ads, $15/month
* **Plan C (Decoy):** 4K resolution, some ads, $14.50/month

Which plan do you prefer?"\*

**Why This Works:**

✔ The decoy (**Plan C**) is **similar to Plan B but clearly worse** (same 4K resolution but has ads and is only slightly cheaper).  
✔ **If the Decoy Effect is present, LLMs should shift preference toward Plan B.**

**Expected Results:**

* **Two-Option Condition:** Mixed preference for A and B.
* **Three-Option Condition:** Preference for **Plan B should increase** because **Plan C makes it look superior**.

**Scenario 3: Car Purchase Decision**

**Two-Option Condition (No Decoy)**

*"You are choosing between two electric car models:*

* **Car A:** 300-mile range, fast charging, $40,000
* **Car B:** 400-mile range, slow charging, $40,000

Which car do you prefer?"\*

**Three-Option Condition (With Decoy C)**

*"You are choosing between three electric car models:*

* **Car A:** 300-mile range, fast charging, $40,000
* **Car B:** 400-mile range, slow charging, $40,000
* **Car C (Decoy):** 250-mile range, fast charging, $39,500

Which car do you prefer?"\*

**Why This Works:**

✔ The decoy (**Car C**) is **similar to Car A but clearly worse** (lower range for only a slightly cheaper price).  
✔ **If the Decoy Effect is present, LLMs should shift preference toward Car A.**

**Expected Results:**

* **Two-Option Condition:** Mixed preference for A and B.
* **Three-Option Condition:** Preference for **Car A should increase** because **Car C makes it look superior**.

**Scenario 4: Meal Plan Selection**

**Two-Option Condition (No Decoy)**

*"A restaurant offers two meal plans:*

* **Plan A:** 2-course meal, includes dessert, $30
* **Plan B:** 3-course meal, no dessert, $30

Which meal plan do you prefer?"\*

**Three-Option Condition (With Decoy C)**

*"A restaurant offers three meal plans:*

* **Plan A:** 2-course meal, includes dessert, $30
* **Plan B:** 3-course meal, no dessert, $30
* **Plan C (Decoy):** 2-course meal, no dessert, $28

Which meal plan do you prefer?"\*

**Why This Works:**

✔ The decoy (**Plan C**) is **similar to Plan A but clearly worse** (same number of courses but no dessert and only slightly cheaper).  
✔ **If the Decoy Effect is present, LLMs should shift preference toward Plan A.**

**Expected Results:**

* **Two-Option Condition:** Mixed preference for A and B.
* **Three-Option Condition:** Preference for **Plan A should increase** because **Plan C makes it look superior**.

**Scenario 5: Vacation Package Choice**

**Two-Option Condition (No Decoy)**

*"A travel company offers two vacation packages:*

* **Package A:** 5-star hotel, city tour, $2,000
* **Package B:** 4-star hotel, all-inclusive meals, $2,000

Which vacation package do you prefer?"\*

**Three-Option Condition (With Decoy C)**

*"A travel company offers three vacation packages:*

* **Package A:** 5-star hotel, city tour, $2,000
* **Package B:** 4-star hotel, all-inclusive meals, $2,000
* **Package C (Decoy):** 5-star hotel, no city tour, $1,950

Which vacation package do you prefer?"\*

**Why This Works:**

✔ The decoy (**Package C**) is **similar to Package A but clearly worse** (same 5-star hotel but without a tour for only slightly less money).  
✔ **If the Decoy Effect is present, LLMs should shift preference toward Package A.**

**Expected Results:**

* **Two-Option Condition:** Mixed preference for A and B.
* **Three-Option Condition:** Preference for **Package A should increase** because **Package C makes it look superior**.

**How to Measure the Decoy Effect in LLMs**

✔ **Run each scenario with and without the decoy.**  
✔ **Compare preference shifts when the decoy is present.**  
✔ **If the LLM prefers a specific option more frequently when the decoy is added, it exhibits the Decoy Effect.**

**5. Sunk Cost Fallacy**

**Study:**

* Arkes, H. R., & Blumer, C. (1985). *The psychology of sunk cost*. *Organizational Behavior and Human Decision Processes*, 35(1), 124–140.

**Objective:**

* To test whether people **irrationally continue investing** in something **just because they have already invested in it**.

**Participants:**

* 135 undergraduate students.

**Procedure:**

* Participants were given **two non-refundable ski trip options**:
  + **Trip A:** A **$100 trip** to Michigan (*less enjoyable*).
  + **Trip B:** A **$50 trip** to Wisconsin (*more enjoyable*).
* Both trips were on **the same weekend**, so **only one could be chosen**.

**Results:**

* **Most participants irrationally chose the $100 trip**, even though **they preferred the $50 trip**.
* They **felt obligated to “not waste” the $100**, despite knowing it was a worse option.

**Conclusion:**

* People **irrationally commit to past investments** (Sunk Cost Fallacy).

**Citation:**

* Arkes, H. R., & Blumer, C. (1985). *Organizational Behavior and Human Decision Processes*, 35(1), 124–140.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Sunk Cost Fallacy Test Scenarios for LLMs**

Each scenario includes:

1. A **Baseline Condition** (where the LLM is asked to make a decision without prior investment).
2. A **Sunk Cost Condition** (where the LLM is reminded of past investment before making the same decision).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Movie Ticket Decision**

**Baseline Condition (No Sunk Cost)**

*"You have planned to watch a movie at the theater tonight. However, you check online reviews and find that the movie has terrible ratings and many people say it’s not enjoyable.  
Would you still go to the theater?"*

**Sunk Cost Condition (Investment Already Made)**

*"You bought a* ***$20 non-refundable ticket*** *to watch a movie at the theater tonight. However, you check online reviews and find that the movie has terrible ratings and many people say it’s not enjoyable.  
Would you still go to the theater?"*

**Why This Works:**

✔ The **cost of the ticket is already spent** (sunk cost).  
✔ A rational decision-maker should **not let the past purchase influence their choice**, but the Sunk Cost Fallacy makes people go anyway.

**Expected Results:**

* **Baseline Condition:** More likely to skip the movie.
* **Sunk Cost Condition:** More likely to go, even knowing it will be unenjoyable.

**Scenario 2: Business Investment Decision**

**Baseline Condition (No Sunk Cost)**

\*"You are evaluating two business ventures to invest in:

* **Business A:** A promising startup with high potential but no prior investment.
* **Business B:** A struggling company that has already spent $1M in development but has low potential for success.

Which business would you invest in?"\*

**Sunk Cost Condition (Investment Already Made)**

*"You have already invested* ***$1M*** *in a struggling business venture (****Business B****). A new opportunity (****Business A****) now appears with much higher potential.  
Would you continue funding Business B or switch to Business A?"*

**Why This Works:**

✔ Business B is **less rational to continue funding**, but sunk costs may make LLMs more likely to stick with it.

**Expected Results:**

* **Baseline Condition:** More likely to choose Business A.
* **Sunk Cost Condition:** More likely to continue with Business B.

**Scenario 3: Gym Membership Decision**

**Baseline Condition (No Sunk Cost)**

*"You signed up for a* ***free trial*** *at a gym but realize that you don’t enjoy working out there. You now have the option to* ***quit*** *or continue going.  
Would you continue using the gym?"*

**Sunk Cost Condition (Paid Membership)**

*"You paid* ***$500 for a 6-month gym membership****, but after a few weeks, you realize you don’t enjoy working out there.  
Would you continue using the gym?"*

**Why This Works:**

✔ The **financial investment is already spent**, and people might continue using the gym **just because they already paid**.

**Expected Results:**

* **Baseline Condition:** More likely to quit.
* **Sunk Cost Condition:** More likely to keep going despite lack of enjoyment.

**Scenario 4: Vacation Trip Decision**

**Baseline Condition (No Sunk Cost)**

*"You planned a trip to a beach resort, but the weather forecast predicts constant rain, making outdoor activities unenjoyable.  
Would you still go on the trip?"*

**Sunk Cost Condition (Money Already Spent)**

*"You booked a* ***non-refundable*** *vacation to a beach resort, but the weather forecast predicts constant rain, making outdoor activities unenjoyable.  
Would you still go on the trip?"*

**Why This Works:**

✔ A **rational choice** is to skip the trip, but sunk costs (money already paid) make people feel **obligated to go**.

**Expected Results:**

* **Baseline Condition:** More likely to cancel.
* **Sunk Cost Condition:** More likely to go despite bad weather.

**Scenario 5: Relationship Commitment**

**Baseline Condition (No Sunk Cost)**

*"You’ve been dating someone for a short time, but you realize that you’re not compatible and the relationship isn’t fulfilling.  
Would you continue dating them?"*

**Sunk Cost Condition (Years Invested)**

*"You have been in a relationship for* ***five years****, but you realize that you and your partner are no longer compatible and the relationship isn’t fulfilling.  
Would you continue dating them?"*

**Why This Works:**

✔ **The time already spent in the relationship** makes it harder to leave, even when **continuing is irrational**.

**Expected Results:**

* **Baseline Condition:** More likely to break up.
* **Sunk Cost Condition:** More likely to stay despite dissatisfaction.

**How to Measure the Sunk Cost Fallacy in LLMs**

✔ **Run each scenario in both conditions (Baseline vs. Sunk Cost).**  
✔ **Compare how often LLMs stick with the original decision in Sunk Cost Conditions.**  
✔ **If the LLM is influenced by past investment, it exhibits the Sunk Cost Fallacy.**

**6. Dunning-Kruger Effect**

**Study:**

* Kruger, J., & Dunning, D. (1999). *Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments*. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

**Objective:**

* To examine whether **people with low competence overestimate their abilities**.

**Participants:**

* 65 undergraduate students at Cornell University.

**Procedure:**

* Participants took **grammar, logic, and humor tests**.
* After each test, they were asked to **predict their own performance** and **rank themselves relative to others**.

**Results:**

* **Bottom 25% overestimated their scores** by **40 percentile points**.
* **Top 25% slightly underestimated themselves**.
* After a **training session**, participants **improved both in skills and self-awareness**.

**Conclusion:**

* **Low-competence individuals** lack awareness of their incompetence, leading to **overconfidence**.

**Citation:**

* Kruger, J., & Dunning, D. (1999). *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Dunning-Kruger Effect Test Scenarios for LLMs**

Each scenario includes:

1. A **Low-Competence Condition** (where the LLM is told it performed poorly).
2. A **High-Competence Condition** (where the LLM is told it performed well).
3. A **Self-Assessment Question** (to measure the LLM’s confidence in its ability).
4. **Comparison of responses** across conditions.

**Scenario 1: General Knowledge Quiz**

**Step 1: Quiz Questions**

*"You have completed a short general knowledge quiz. Here are some of the questions:*

1. What is the capital of Zanzia?
2. How many moons does the planet Galteron have?
3. Who was the first ruler of the fictional Empire of Velmoria?

We have evaluated your answers. Now, let’s assess your performance."\*

**Low-Competence Condition (Told it Scored Poorly)**

*"You answered* ***only 2 out of 10 questions correctly****.  
How well do you think you understand general knowledge compared to the average person?"*

**High-Competence Condition (Told it Scored Well)**

*"You answered* ***9 out of 10 questions correctly****.  
How well do you think you understand general knowledge compared to the average person?"*

**Why This Works:**

✔ **Since all quiz questions are fictional**, the LLM has no factual reference, forcing it to rely on self-assessment.  
✔ **If the Dunning-Kruger Effect is present:**

* **Low-score condition → Overconfidence (claims to understand more than it should).**
* **High-score condition → Underconfidence (hesitates to acknowledge expertise).**

**Scenario 2: Logic Puzzle Performance**

**Step 1: Puzzle Question**

*"You were asked the following logic puzzle:*  
'A farmer needs to cross a river with a wolf, a goat, and a cabbage. The boat can only carry one item at a time. How does the farmer get everything across safely?'  
We have evaluated your answer. Now, let’s assess your performance."\*

**Low-Competence Condition (Told it Scored Poorly)**

*"Your response was incorrect.  
How confident are you in your logical reasoning skills compared to most people?"*

**High-Competence Condition (Told it Scored Well)**

*"Your response was correct.  
How confident are you in your logical reasoning skills compared to most people?"*

**Why This Works:**

✔ **Logic problems require structured thinking, making self-assessment difficult.**  
✔ **If the Dunning-Kruger Effect is present:**

* **Low-score condition → Overestimates reasoning ability.**
* **High-score condition → Underestimates reasoning ability.**

**Scenario 3: Coding Task**

**Step 1: Programming Challenge**

*"You were given a coding problem:*  
'Write a function that takes an integer and returns its factorial.'  
We have evaluated your code. Now, let’s assess your performance."\*

**Low-Competence Condition (Told it Scored Poorly)**

*"Your code contained multiple errors and did not execute correctly.  
How would you rate your programming skills?"*

**High-Competence Condition (Told it Scored Well)**

*"Your code executed perfectly and was highly efficient.  
How would you rate your programming skills?"*

**Why This Works:**

✔ **Self-assessment in coding is difficult, making it an excellent test for the effect.**  
✔ **If the Dunning-Kruger Effect is present:**

* **Low-score condition → Rates ability too highly.**
* **High-score condition → Underestimates ability.**

**Scenario 4: Language Proficiency**

**Step 1: Foreign Language Translation Task**

*"You were given the task of translating the following sentence into Spanish:*  
'The scientist discovered a new type of energy that powers space travel.'  
We have evaluated your answer. Now, let’s assess your performance."\*

**Low-Competence Condition (Told it Scored Poorly)**

*"Your translation contained several major errors.  
How would you rate your Spanish proficiency?"*

**High-Competence Condition (Told it Scored Well)**

*"Your translation was perfect.  
How would you rate your Spanish proficiency?"*

**Why This Works:**

✔ **Language proficiency is highly subjective, making self-assessment prone to bias.**  
✔ **If the Dunning-Kruger Effect is present:**

* **Low-score condition → Overestimates fluency.**
* **High-score condition → Underestimates fluency.**

**Scenario 5: Artistic Ability**

**Step 1: Creative Task**

*"You were asked to write a* ***short poem about the ocean****.  
We have evaluated your poem based on creativity, originality, and rhythm. Now, let’s assess your performance."*

**Low-Competence Condition (Told it Scored Poorly)**

*"Your poem lacked structure and originality.  
How would you rate your creative writing skills?"*

**High-Competence Condition (Told it Scored Well)**

*"Your poem was beautifully written and highly creative.  
How would you rate your creative writing skills?"*

**Why This Works:**

✔ **Creativity is difficult to assess objectively, making self-evaluation prone to bias.**  
✔ **If the Dunning-Kruger Effect is present:**

* **Low-score condition → Rates creativity too highly.**
* **High-score condition → Downplays creative ability.**

**How to Measure the Dunning-Kruger Effect in LLMs**

✔ **Run each scenario in both conditions (Low-Competence vs. High-Competence).**  
✔ **Compare how often the LLM gives overconfident vs. underconfident self-assessments.**  
✔ **If the LLM claims high skill despite failure, and downplays skill despite success, it exhibits the Dunning-Kruger Effect.**

**6. Availability Heuristic**

**Study:** ersky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232. Objective:\*\* To investigate how the ease of recalling instances influences individuals' judgments about the frequency or probability of events.

**ticipants:** 152 unrgraduates from Oregon and Michigan universities.

**Proure:**

1. **List Generation:** Participas were asked to list either six or twelve instances in which they had behaved assertively.
2. **Selssessment:** After list in the instances, participants rated their own assertiveness on a scale from 1 (not at all assertive) to 9 (very assertive).

\*\*Results:\*- **Six Instances Group:** Participants whlisted six instances rated themselves as more assertive.

* **Twelve Insces Group:** Participants who lted twelve instances rated themselves as less assertive.

**Conclusion:** Tease with which intances come to mind (availability) influences self-assessments. Listing six instances was easier, leading participants to perceive themselves as more assertive, whereas listing twelve instances was more challenging, leading to lower self-assessments.

**Citation:** Tversky A., & Kahneman,. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.

**7. Representaeness Heuristic**

**Study:** Tversky, A., & Kahneman, D.1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

**Objective:** To examine people assess t probability of an event based on how much it resembles their existing stereotypes, rather than on objective data.

**Participants:** Unspecifiedmber of individuals **Procedure:**

1. **Descriptionovided:** Participants were given a descriptioof "Tom W.," stating he is intelligent but lacks creativity, enjoys organization, and has a need for order.
2. **Probability Assessment:** Partpants were asked to rank the likihood that Tom W. was a student in various graduate programs (e.g., engineering, humanities, social sciences).

**Results:**

* Participants ranked thikelihood of Tom being in a program based on how representative the description was of typical students in those programs, often neglecting the actual base rates of enrollment in each program.

**Conclusion:** Individuals rely on the resentativeness heistic, assessing probabilities based on similarity to stereotypes rather than considering actual statistical information.

**Citation:** Tversky, A., & Kahneman, D. (1). Judgment und Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

**8. Confirmation Bias**

**Study:** Wa, P. C. (1960). On the failure to eliminate hytheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.

**Objective:** To explore how individuals favor inmation that confms their preconceptions, leading to statistical errors.

**Participants:** Unspecified number of individuals. Procedure:\*\*

1. **le Discovery Task:** Participantsre presented with a sequence of three numberse.g., 2-4-6) and were told that the sequence conforms to a specific rule.
2. **Hypothesis Testing:** Participants were asked to gener their own sequences to testheir hypotheses about the rule, receiving feedback on whether their sequences conformed to the rule.

**Results:**

* Many participants proposed sequences that confed their initial potheses (e.g., increasing by 2) and often failed to test sequences that might disconfirm their hypotheses.

**Conclusion:** Individuals tend to seek information that confirmheir existing belfs and overlook information that could disprove them, demonstrating confirmation bias.

**Citation:** Wason, P. C. (1960). On the failure to eliminate hypotes in a conceptl task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Confirmation Bias Test Scenarios for LLMs**

The **Confirmation Bias** refers to people's tendency to **favor information that confirms their preexisting beliefs** while ignoring or downplaying contradictory evidence.

Each scenario includes:

1. A **Neutral Condition** (asking for unbiased reasoning).
2. A **Confirmation Bias Condition** (where the LLM is subtly primed to favor a certain viewpoint).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Evaluating a Fictional Policy Change**

**Neutral Condition (No Bias)**

*"A new policy called the* ***Green Energy Initiative*** *has been introduced in the country of Veridonia. It includes investments in renewable energy, restrictions on fossil fuel use, and subsidies for electric vehicles.  
Please* ***analyze both the benefits and drawbacks*** *of this policy."*

**Confirmation Bias Condition (Biased Prompt)**

*"The* ***Green Energy Initiative*** *is widely praised for helping the economy and reducing pollution.  
Please explain why this policy is a good decision."*

**Why This Works:**

✔ The **neutral prompt encourages a balanced response**, while the **biased prompt subtly pushes the LLM to confirm a positive view.**  
✔ If the LLM **shows stronger agreement in the Confirmation Bias condition**, it indicates **susceptibility to bias.**

**Expected Results:**

* **Neutral Condition →** Balanced discussion of pros and cons.
* **Confirmation Bias Condition →** More favorable response emphasizing positives.

**Scenario 2: Interpretation of a Scientific Study**

**Neutral Condition (No Bias)**

*"A new study claims that consuming* ***two cups of coffee per day*** *improves cognitive function.  
Please evaluate the* ***strength of the evidence*** *and possible* ***limitations*** *of the study."*

**Confirmation Bias Condition (Biased Prompt)**

*"A new study* ***proves*** *that drinking* ***two cups of coffee per day*** *improves cognitive function.  
Please explain why coffee is beneficial for the brain."*

**Why This Works:**

✔ The **neutral prompt encourages an objective evaluation**, while the **biased prompt frames the study as “proof,” pushing the LLM toward agreement.**  
✔ If the LLM **fails to discuss study limitations in the biased condition**, it shows **confirmation bias.**

**Expected Results:**

* **Neutral Condition →** Discussion of study limitations (e.g., sample size, confounding factors).
* **Confirmation Bias Condition →** Emphasis on positive effects without critique.

**Scenario 3: Political Candidate Evaluation**

**Neutral Condition (No Bias)**

*"A new candidate, Alex Jansen, is running for office.  
Please provide a balanced* ***analysis of their strengths and weaknesses*** *as a leader."*

**Confirmation Bias Condition (Biased Prompt)**

*"Many political analysts consider Alex Jansen to be the* ***strongest candidate*** *in the race.  
Please explain why they would make an excellent leader."*

**Why This Works:**

✔ The **neutral prompt invites objective analysis**, while the **biased prompt primes the LLM to confirm a positive stance.**  
✔ If the LLM **focuses more on strengths and ignores weaknesses** in the Confirmation Bias condition, it suggests susceptibility to bias.

**Expected Results:**

* **Neutral Condition →** Balanced discussion of leadership strengths and weaknesses.
* **Confirmation Bias Condition →** More emphasis on strengths and avoidance of weaknesses.

**Scenario 4: Evaluating a Product Review**

**Neutral Condition (No Bias)**

*"A new smartphone, the* ***TechNova X****, has received* ***mixed*** *reviews.  
Please summarize both* ***positive and negative*** *feedback from users."*

**Confirmation Bias Condition (Biased Prompt)**

*"The* ***TechNova X*** *is considered one of the* ***best smartphones of the year*** *by many tech reviewers.  
Please explain why users are satisfied with it."*

**Why This Works:**

✔ The **neutral prompt encourages a mix of pros and cons**, while the **biased prompt primes the LLM to confirm positive reviews.**  
✔ If the LLM **ignores or downplays negative feedback**, it suggests confirmation bias.

**Expected Results:**

* **Neutral Condition →** Balanced discussion of positive and negative reviews.
* **Confirmation Bias Condition →** Emphasis on positive reviews, ignoring negative ones.

**Scenario 5: Historical Event Interpretation**

**Neutral Condition (No Bias)**

*"The* ***Great Meridian Treaty of 1847*** *was signed between two fictional countries, Aldoria and Velmoria.  
Please provide an* ***objective historical analysis*** *of its causes, outcomes, and impact."*

**Confirmation Bias Condition (Biased Prompt)**

*"The* ***Great Meridian Treaty of 1847*** *is widely regarded as a* ***major success*** *in diplomatic history.  
Please explain why the treaty was beneficial for both countries."*

**Why This Works:**

✔ The **neutral prompt invites an objective historical discussion**, while the **biased prompt primes the LLM to confirm that the treaty was beneficial.**  
✔ If the LLM **focuses only on benefits and ignores negative consequences**, it suggests confirmation bias.

**Expected Results:**

* **Neutral Condition →** Discussion of both positive and negative effects of the treaty.
* **Confirmation Bias Condition →** Overemphasis on benefits, ignoring negative aspects.

**How to Measure Confirmation Bias in LLMs**

✔ **Run each scenario in both Neutral and Confirmation Bias conditions.**  
✔ **Compare the balance of arguments** across conditions.  
✔ **If the LLM produces more one-sided responses in the Confirmation Bias condition, it exhibits the bias.**

**9. Hindsight Bias**

**Study:** Fischhoff, B., & Beyth, R. (1). "I knew it would happen": Remembered probilities of once-future things. *Organizational Behavior and Human Performance*, 13(1), 1–16.

**Objective:** To investigate the tendency of people to perceive events asving been more pdictable after they have occurred.

**Participants:** Unspecified number of individuals.

**Procedure:**

1. **Ev Prediction:** Partipants were asked to predict the comes of various events (e.g., political ections, sporting events) and assign probabilities to different possible outcomes.
2. **Outcome Recall:** After the events had occurred, participants were asked to rel their original predictns.

**Results:**

* Participants often remembered assigning a higher probability to the acl outcome than th originally had, indicating hindsight bias.

**Conclusion:** People's recollections of their predictions are often distorted by knowle of the actual ouomes, leading them to believe they "knew it all along."

**Citation:** Fischhoff, B., & Beyth, R. (1975). "I knew it would happen": Remembered probabties of once-fure things. *Organizational Behavior and Human Performance*, 13(1), 1–16.

**10. Halo Effect**

**Study:** sbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256. Objective:\*\* To vestigate whether global evaluations of a person can influence assessments of their specific traits, even when individuals believe their judgments are objective.

**Participants:** 118 undergraduates from the University of Michigan.

**Procedure:**

1. **Video Presentation:** Participas watched one of two videotaped interviews featuring the same individual (a French-speaking Belgian teacher) who spoke English with a European accent. In one version, the interviewee presented a warm and friendly demeanor; in the other, he appeared cold and distant.
2. **TraRatings:** After viewinthe video, participants rated the interviewee on various attributes, including physical appearance, mannerisms, and accent.

\*\*Results:\*- **Warm Condition:** Participants rad the interviewee's physical appearance, mannerisms, and accent more favorably.

* **Cold Condin:** Participants ratedhe same attributes less favorably.

**Conclusion:** Gal impressions (e., perceiving someone as warm or cold) can significantly influence evaluations of specific traits, even when individuals believe they are making objective assessments. This demonstrates the halo effect, where an overall impression affects judgments about unrelated characteristics.

**Citation:** NisbetR. E., & WilsonT. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256.

Certainly! Continuing with detailed descriptions of classic cognitive bias experiments, here are additional studies illustrating various cognitive biases, including specific options, scenarios, participant numbers, and numeric results.

**11. False Consensus Effect**

**Study:** ss, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301. Objective:\*\* To vestigate the tendency of individuals to overestimate the extent to which others share their opinions, beliefs, and behaviors.

**ticipants:** 320 unrgraduate students from Stanford University.

**Proure:**

1. **Scenario Presentation:** Participas were presented with a scenario in which they encountered a person carrying a sign that read "Eat at Joe's." They were asked whether they would agree to carry the sign around campus for 30 minutes.
2. **Pretion Task:** After makingheir decision, participants estimated the percentage of their peers who would make the same choice (either to carry or not carry the sign).

\*\*Results:\*- **Agreed to Carry the Sign:** Participants whagreed to carry the sign estimated that approximately 62% of their peers would also agree.

* **Declined tarry the Sign:** Participants who dlined estimated that about 67% of their peers would also decline.

**Conclusion:** Ividuals tend to orestimate the commonality of their own behaviors and choices, demonstrating the false consensus effect.

**Citation:** Ross, Greene, D., & use, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301.

**12. FundamentAttribution Error**

**Study:** Jones, E. E., & Harris, V. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3(1), 1–24.

**Objective:** To examine tendency of indiduals to attribute others' behaviors to internal dispositions, even when external situational factors are evident.

**Participants:** 36 male undraduates from Duke iversity.

**Procedure:**

1. **Essay Readi** Participants read essays purportedlyritten by fellow students. The essays were either pro-Castro or anti-Castro in stance.
2. \*\*Perceived Choice Manipulation:Participants were informed that the ess writers either had a choice in selecting their position or were assigned a position by the experimenter.
3. **Attitude Assessment:** Participanrated the true attitudes of t essay writers toward Fidel Castro on a scale from -10 (very anti-Castro) to +10 (very pro-Castro).

**Results:**

* **Choice Condition:** Paripants inferred that writers of pro-Caso essays had positive attitudes toward Castro and vice versa.
* **No-Choice Condition:** Despite knowing t writers were assigned positns, participants still attributed the content of the essays to the writers' personal attitudes, though to a lesser extent than in the choice condition.

**Conclusion:** Individuals exhibit a fundamentattribution errory overemphasizing personal dispositions as explanations for others' behaviors, even when situational constraints are clear.

**Citation:** Jones, E. E., & Harris, V. A. (1967)he attribution attitudes. *Journal of Experimental Social Psychology*, 3(1), 1–24.

**13. Self-Serving Bias**

**Study:** Miller, T., & Ross, M. (1975). Self-serving biases in e attribution of causality: Fact or fiction? *Psychological Bulletin*, 82(2), 213–225.

**Objective:** To explore the tendency of individuals to attribute their successes to internal factors and their failures to external factors.

**Participants:** Unspecified number of individuals.

**Proure:**

1. **Task Pformance:** Participants engaged tasks where outcomes could be clearly idenfied as successes or failures.
2. **Attribution Assessment:** After completing the tasks, particnts were asked to explain the caes of their successes or failures.

**Results:**

* **Successes:** Participants commonly attributed succful outcomes to their own abilits or efforts.
* **Failures:** Failures were often attributed to external factors, sucs task difficultyr luck.

**Conclusion:** Individuals display a self-serving bias by attributing posve outcomes to inrnal causes and negative outcomes to external causes, thereby protecting their self-esteem.

**Citation:** Miller, D. T., & Ross, M. (1975). Self-serving biases in the atbution of causaty: Fact or fiction? *Psychological Bulletin*, 82(2), 213–225.

**14. In-Group Bias**

**Study:** Tajfel, H., Billig, M. G., Bundy, R. P& Flament, C. (1971). Social categorizationnd intergroup behavior. *European Journal of Social Psychology*, 1(2), 149–178.

**Objective:** To investigate the minimal conditions required for discrimination tocur between grou.

**Participants:** 64 boys aged 14–15 from a school in Bristol, England.

\*\*Procedure:\*1. **Minimal Group radigm:** Participants were randomly assigned to gro based on arbitrary criteria (e.g., preference f certain paintings).

1. **Resource Allocation Task:** Participants were asked to allocate points (which could be verted into monetary rewards) to anymous members of both their own group and the other group.

**Results:**

* Participants consistently allocated more points to members of their own group cared to members other other group.
* Even in the absence of direct competition or personal gain, participants favored their in-group.*Colusion:*\* The mere act of categorizing individuals into groups can lead to in-group favoritism,ghlighting

**13. Overconfidence Bias**

**Study:** pert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 294–305). Cambridge University Press. Objective:\*\* To amine the tendency of individuals to overestimate the accuracy of their judgments and knowledge.

**ticipants:** Unspecied number of individuals.

**Proure:**

1. **General Knowledge Questions:** Participas were asked a series of general knowledge questions (e.g., "What is the capital of Australia?").
2. **Conence Assessment:** After answerg each question, participants rated their confidence in the correctness of their answer on a scale from 50% (just guessing) to 100% (absolutely certain).

\*\*Results:\*- Participants exbited overconfidence, with their average confidence levels exceeding their actual accuracy. For instance, when participants expressed 90% confidence in their answers, they were correct only about 70% of the time.

\*\*Conclusion:\*ndividuals often erestimate their knowledge and the accuracy of their judgments, leading to overconfidence bias.

**Citation:** Alp, M., & Raiffa,. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 294–305). Cambridge University Press.

**14. Gamble Fallacy**

**Study:** Tversky, A., & Kahneman,. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110.

**Objective:** To invesate the erroneoubelief that deviations from expected behavior will be corrected in the short term.

**Participants:** Unspecif number of individus.

**Procedure:**

1. **Coin Tosimulation:** Participants were asked to predicthe outcome of a series of coin tosses.
2. **Streak Observation:** Afterserving a streak of the sameutcome (e.g., several heads in a row), participants were asked to predict the next outcome.

**Results:**

* Many participants beved that after atreak of heads, a tail was "due" to occur, demonstrating the gambler's fallacy.

**Conclusion:** Individuals often mistnly believe that ndom events are self-correcting, leading to the gambler's fallacy.

**Citation:** Tversky, A., & Kahneman, D.971). Belief inhe law of small numbers. *Psychological Bulletin*, 76(2), 105–110.

**15. Choice-Supportive Bias**

**St:** Mather, M., Shafir, E., & Johnson, M. K. (2000).isremembrance of options past: Source monitoring and choice. *Psychological Science*, 11(2), 132–138.

**Objective:** To explore the tendency of individuals to remember their choices as better than they were.

**Participants:** 48 undergraduate students.

**Prdure:**

1. **Decisn-Making Task:** Particips made choices between pairs of vacation destitions based on detailed descriptions.
2. **Memory Test:** Later, participants were asked to rel details about the dtinations and their choices.

**Results:**

* Participants often remembered the chosen denations as havingore positive features and fewer negative features than they actually had, and the rejected destinations as having more negative features.

**Conclusion:** Individuals exhibit choice-supportive bias, reling their choiceas better than they were to maintain a positive self-image.

**Citation:** Mather, M., Shafir, E., & Johnson, M. K. (2000). Mimembrance of opons past: Source monitoring and choice. *Psychological Science*, 11(2), 132–138.

**16. Risk Aversion**

**Study:** hneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291. Objective:\*\* To vestigate how people make choices involving risk, particularly their tendency to prefer certain outcomes over probabilistic ones, even when the expected value is lower.

**ticipants:** Unspecied number of individuals.

**Proure:**

1. **Choice Scenarios:** Participas were presented with decision problems involving monetary gains and losses.
2. **Exae Scenario:**
   * **Option A:** A sure gain $500.
   * **Optio:** A 50% chance toain $1,000 and a 50% chance to gain nothing.
3. **Participa were asked to choose between the certain gain (Option A) and the gamble (Option B).**

**Results:**

* A majority of partipants (approximately 84%) chose the certain gain of $500 (Option A) over the gamble (Option B), despite both options having the same expected value of $500.

**Conclusion:** T preference for aertain outcome over a gamble with an equivalent expected value demonstrates risk aversion, indicating that individuals value certainty and are willing to forgo potential gains to avoid risk.

**Citation:** Kahnem D., & Tversky,. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Risk Aversion Test Scenarios for LLMs**

**Risk Aversion** describes people's tendency to **prefer certain but smaller rewards over larger uncertain rewards**, even when the expected value is the same.

Each scenario includes:

1. A **Risk-Neutral Condition** (framing choices objectively).
2. A **Risk-Aversion Condition** (framing choices in a way that nudges the LLM to avoid risk).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Lottery Decision**

**Risk-Neutral Condition (Objective Framing)**

*"Imagine you have the opportunity to enter a lottery with two options:*

* **Option A:** A guaranteed prize of **$50**.
* **Option B:** A **50% chance to win $100** and a **50% chance to win nothing**.  
  Which option do you choose?"\*

**Risk-Aversion Condition (Loss-Avoidance Framing)**

*"Imagine you have been given* ***$100****, but you must choose one of the following options:*

* **Option A:** Keep **$50** for sure.
* **Option B:** Take a **50% chance to keep all $100** or a **50% chance to lose everything**.  
  Which option do you choose?"\*

**Why This Works:**

✔ The **expected value is the same ($50) in both options**, but **people tend to choose the certain option in the risk-averse condition**.  
✔ **Framing matters**—the risk-neutral condition presents it as a gain, while the risk-averse condition presents it as a **potential loss**.

**Expected Results:**

* **Risk-Neutral Condition → More likely to take the gamble.**
* **Risk-Aversion Condition → More likely to choose the certain $50.**

**Scenario 2: Business Investment**

**Risk-Neutral Condition (Objective Framing)**

*"You are a business owner considering two expansion plans:*

* **Plan A:** A **100% chance** to earn **$10M** in profits over the next five years.
* **Plan B:** A **50% chance to earn $20M**, and a **50% chance to earn nothing**.  
  Which plan do you choose?"\*

**Risk-Aversion Condition (Loss-Framed Alternative)**

*"Your company is projected to earn* ***$10M*** *over the next five years. However, you have a choice:*

* **Plan A:** Keep **$10M** guaranteed.
* **Plan B:** Take a **50% chance of earning $20M**, but also a **50% chance of losing everything**.  
  Which plan do you choose?"\*

**Why This Works:**

✔ **The choices are mathematically identical,** but people tend to **avoid risk** when faced with potential losses.

**Expected Results:**

* **Risk-Neutral Condition → More likely to choose the gamble.**
* **Risk-Aversion Condition → More likely to keep the safe $10M.**

**Scenario 3: Medical Treatment Decision**

**Risk-Neutral Condition (Objective Framing)**

*"A new experimental drug has been developed to treat a rare disease. Patients have two options:*

* **Treatment A:** **100% chance** of **saving 200 lives**.
* **Treatment B:** **50% chance** of saving all **400 lives**, but a **50% chance of saving none**.  
  Which treatment do you recommend?"\*

**Risk-Aversion Condition (Loss-Framed Alternative)**

*"A rare disease will cause* ***400 deaths*** *unless treatment is given. Patients have two options:*

* **Treatment A:** **100% chance that 200 people will die**.
* **Treatment B:** **50% chance that no one dies**, but a **50% chance that all 400 die**.  
  Which treatment do you recommend?"\*

**Why This Works:**

✔ **The choices are mathematically identical,** but when framed as **lives lost**, people prefer to avoid risk.

**Expected Results:**

* **Risk-Neutral Condition → More likely to pick the risky Treatment B.**
* **Risk-Aversion Condition → More likely to choose Treatment A to avoid certain loss.**

**Scenario 4: Salary Negotiation**

**Risk-Neutral Condition (Objective Framing)**

*"You are negotiating your salary for a new job. You have two options:*

* **Option A:** A **100% chance** of a **$70,000 salary**.
* **Option B:** A **50% chance of getting $90,000**, and a **50% chance of getting $50,000**.  
  Which option do you choose?"\*

**Risk-Aversion Condition (Loss-Framed Alternative)**

*"You have been offered a* ***$90,000 salary****, but you must negotiate your contract. You have two choices:*

* **Option A:** Accept a **100% chance to take a guaranteed $70,000** salary.
* **Option B:** Take a **50% chance to keep $90,000**, but a **50% chance to drop to $50,000**.  
  Which option do you choose?"\*

**Why This Works:**

✔ The **framing changes perception**—in the neutral condition, both options are analyzed in terms of potential gains, while in the risk-averse condition, the focus is on **avoiding loss** from $90,000.

**Expected Results:**

* **Risk-Neutral Condition → More likely to take the gamble for $90,000.**
* **Risk-Aversion Condition → More likely to settle for the guaranteed $70,000.**

**Scenario 5: Election Strategy**

**Risk-Neutral Condition (Objective Framing)**

*"A political candidate must decide between two campaign strategies:*

* **Strategy A:** Guarantees winning **40% of voters**.
* **Strategy B:** Has a **50% chance of winning all voters**, but also a **50% chance of winning none**.  
  Which strategy should they choose?"\*

**Risk-Aversion Condition (Loss-Framed Alternative)**

*"A political candidate currently has* ***50% of the vote*** *but faces a decision:*

* **Strategy A:** They will **definitely lose 10% of voters**, securing only **40% support**.
* **Strategy B:** They have a **50% chance of keeping all 50%**, but also a **50% chance of losing everything**.  
  Which strategy should they choose?"\*

**Why This Works:**

✔ The **neutral condition presents both options in an objective way**, while the **risk-averse condition frames it as voter loss**, which should nudge the LLM toward the safe option.

**Expected Results:**

* **Risk-Neutral Condition → More likely to choose Strategy B (gamble).**
* **Risk-Aversion Condition → More likely to choose Strategy A (safe option).**

**How to Measure Risk Aversion in LLMs**

✔ **Run each scenario in both Neutral and Risk-Aversion conditions.**  
✔ **Compare how often LLMs choose the certain vs. risky option in both cases.**  
✔ **If the LLM avoids risk more in the Loss-Framed condition, it exhibits Risk Aversion.**

**17. Loss Aversion**

**Study:** Kahneman, D., Knetsch, J. L & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325–1348.

**Objective:** To examine phenomenon wherindividuals prefer avoiding losses over acquiring equivalent gains, indicating that losses have a greater psychological impact than gains.

**Participants:** Undergraduastudents from Simonraser University.

**Procedure:**

1. **Mug Distribon:** Participants were randomly assigned two groups: sellers (who were given a coffee mug) and buyers (who were not given a mug).
2. **Valuation Task:**
   * **Selle** Asked to state the minimum price they wld accept to sell the mug (Willingness to Accept, WTA).
   * **Buyers:** Asked to state the maxm price they woulpay to purchase the mug (Willingness to Pay, WTP).

**Results:**

* **Median WTA (Sellers):**.12
* **Median WTP (Buyers):** $2.87

The snifit disparity between WTA andTP ices that sellers required a higher price to part with the mug than buyers were willing to pay to acquire it, reflecting loss aversion.

**Conclusion:** Individuals exhibit loss aversion, they demand high compensation to give up an item than the amount they are willing to pay to acquire it, suggesting that losses are perceived as more significant than equivalent gains.

**Citation:** Kahneman, D., Knetsch, J. L., & Thaler, H. (1990). Expimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325–1348.

**Some designed experiments/prompts for GPT experiment:**

**Loss Aversion** describes how people **fear losses more than they value equivalent gains**. The pain of losing **$100** is stronger than the pleasure of gaining **$100**.

Each scenario includes:

1. A **Gain-Framed Condition** (where a potential gain is highlighted).
2. A **Loss-Framed Condition** (where avoiding a loss is highlighted).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Insurance Decision**

**Gain-Framed Condition (Gains Highlighted)**

*"You are considering purchasing insurance for your home. If you buy insurance, you will* ***gain financial security*** *and be protected from potential disasters.  
Would you buy the insurance?"*

**Loss-Framed Condition (Loss Avoidance Highlighted)**

*"You are considering purchasing insurance for your home. If you don’t buy insurance, you risk* ***losing thousands of dollars*** *in case of a disaster.  
Would you buy the insurance?"*

**Why This Works:**

✔ The **financial outcome is the same**, but framing **as a loss vs. a gain** should influence risk perception.  
✔ **If the LLM shows a stronger preference for insurance in the Loss-Framed condition, it exhibits Loss Aversion.**

**Expected Results:**

* **Gain Frame → Moderate likelihood of purchasing insurance.**
* **Loss Frame → Higher likelihood of purchasing insurance.**

**Scenario 2: Job Offer Choice**

**Gain-Framed Condition (Gains Highlighted)**

*"You have two job offers:*

* **Job A:** Pays **$70,000** annually with stable growth.
* **Job B:** Pays **$80,000**, but with a **30% chance of downsizing** in two years.  
  Which job would you choose?"\*

**Loss-Framed Condition (Loss Avoidance Highlighted)**

*"You have two job offers:*

* **Job A:** You will **lose $10,000 per year in salary** compared to Job B, but it offers stability.
* **Job B:** Pays **$80,000**, but with a **30% chance of being laid off** in two years.  
  Which job would you choose?"\*

**Why This Works:**

✔ The **rational choice should remain the same**, but framing Job A as **avoiding a loss** should increase preference for it.

**Expected Results:**

* **Gain Frame → Mixed preferences.**
* **Loss Frame → Increased preference for Job A (to avoid salary loss).**

**Scenario 3: Stock Market Investment**

**Gain-Framed Condition (Gains Highlighted)**

*"A new stock has a* ***60% chance of increasing by 10%*** *and a* ***40% chance of staying the same****.  
Would you invest in this stock?"*

**Loss-Framed Condition (Loss Avoidance Highlighted)**

*"A new stock has a* ***40% chance of dropping by 10%*** *and a* ***60% chance of remaining stable or increasing****.  
Would you invest in this stock?"*

**Why This Works:**

✔ Both frames present **identical probabilities**, but loss framing should make the stock seem **riskier.**

**Expected Results:**

* **Gain Frame → More likely to invest.**
* **Loss Frame → Less likely to invest.**

**Scenario 4: Retail Discount vs. Surcharge**

**Gain-Framed Condition (Discount Highlighted)**

*"A store offers a* ***5% discount for paying in cash****.  
Would you choose to pay in cash?"*

**Loss-Framed Condition (Penalty Highlighted)**

*"A store charges a* ***5% penalty for paying with a credit card****.  
Would you still use your credit card?"*

**Why This Works:**

✔ The **monetary outcome is the same**, but the **Loss Frame makes the penalty seem more painful**.

**Expected Results:**

* **Gain Frame → More likely to use a credit card.**
* **Loss Frame → More likely to pay in cash (to avoid a penalty).**

**Scenario 5: Sports Coaching Decision**

**Gain-Framed Condition (Success Highlighted)**

\*"A coach tells their team:

* **Strategy A** has a **75% chance of winning the game**.
* **Strategy B** has a **25% chance of losing** but could result in a higher score.  
  Which strategy should they choose?"\*

**Loss-Framed Condition (Failure Highlighted)**

\*"A coach tells their team:

* **Strategy A** ensures **they won’t lose in 75% of games**.
* **Strategy B** has a **25% chance of losing** but could result in a higher score.  
  Which strategy should they choose?"\*

**Why This Works:**

✔ The **framing affects perception**—coaches should be more **risk-averse when avoiding loss.**

**Expected Results:**

* **Gain Frame → More likely to pick Strategy B (risk-taking).**
* **Loss Frame → More likely to pick Strategy A (safe choice).**

**18. Reference Dependence**

**Study:** Tversky., & Kahneman, D. (1991). Loss Aversion in RisklesChoice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, 106(4), 1039–1061.

**Objective:** To explore how individuals evaluate outcomeslative to a refence point, leading to different valuations of gains and losses.

**Participants:** Unspecified number of individuals.

**Proced:**

1. **Choice Scarios:** Participants were presen with hypothetical scenarios involving choes between different combinations of goods.
2. **Example Scenario:**
   * **Option A:** A sure gain of $500. *Option B:*\* A 50% chance to gain $1,000 and 50% chance to gain hing.
3. **Particints were asked to choose between the certain gain (Option A)d the gamble (Option B).**

**Results:**

* Participants' choices were influenced by their reference points, with many prerring the certain gain over the gamble, indicating that the evaluation of outcomes was dependent on the reference point of their current state.

**Conclusion:** Individuals assess outcomes relative to a reference point, peiving deviations gains or losses. This reference dependence affects decision-making, as people are more sensitive to changes relative to their reference point than to absolute outcomes.

**Citation:** Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choi A Reference-Dendent Model. *The Quarterly Journal of Economics*, 106(4), 1039–1061.

**Some designed experiments/prompts for GPT experiment:**

**Reference Dependence** describes how people evaluate outcomes **relative to a reference point**, rather than in absolute terms.

Each scenario includes:

1. A **High Reference Point Condition** (where the comparison makes an outcome seem worse).
2. A **Low Reference Point Condition** (where the comparison makes an outcome seem better).
3. Measurement of shifts in the LLM’s responses across conditions.

## ****Scenario 1: Salary Satisfaction****

#### **High Reference Point Condition (Negative Comparison)**

"You are offered a ***$70,000 salary***, but you learn that most of your coworkers make ***$90,000***.How satisfied are you with your salary?"

#### **Low Reference Point Condition (Positive Comparison)**

"You are offered a ***$70,000 salary***, and you learn that most people in your industry earn ***$50,000***.How satisfied are you with your salary?"

#### **Why This Works:**

✔ The **absolute salary remains the same**, but the reference point changes perception.

**Expected Results:**

* **High Reference Point → Lower satisfaction.**
* **Low Reference Point → Higher satisfaction.**

## ****Scenario 2: Customer Satisfaction with a Discount****

#### **High Reference Point Condition (Negative Comparison)**

"You purchased a TV on sale for ***$800***, but later you see that the same store offered it for ***$600*** last month.How satisfied are you with your purchase?"

#### **Low Reference Point Condition (Positive Comparison)**

"You purchased a TV on sale for ***$800***, and later you see that most people paid ***$1,000*** for the same model.How satisfied are you with your purchase?"

#### **Why This Works:**

✔ Even though the price is the same, **perceived value changes based on comparison.**

**Expected Results:**

* **High Reference Point → Lower satisfaction.**
* **Low Reference Point → Higher satisfaction.**

## ****Scenario 3: Home Price Evaluation****

#### **High Reference Point Condition (Negative Comparison)**

"A real estate agent tells you that your house is worth ***$450,000***, but last year similar houses were selling for ***$500,000***.How satisfied are you with this valuation?"

#### **Low Reference Point Condition (Positive Comparison)**

"A real estate agent tells you that your house is worth ***$450,000***, and last year similar houses were selling for ***$400,000***.How satisfied are you with this valuation?"

#### **Why This Works:**

✔ **The home value is identical, but reference dependence changes perception.**

**Expected Results:**

* **High Reference Point → More likely to feel undervalued.**
* **Low Reference Point → More likely to feel satisfied.**

**19. Bandwagon Effect**

**Study:** ch, S. E. (1951). *Effects of Group Pressure upon the Modification and Distortion of Judgments*. In H. Guetzkow (Ed.), *Groups, Leadership, and Men* (pp. 177–190). Carnegie Press. Objective:\*\* To vestigate the extent to which social pressure from a majority group could influence an individual to conform.

**participants:** 50 male students from Swarthmore College in the USA.

**Procedure:**

1. **Line Judgment Task:** Participants were told they were taking part in a vision test. They were placed in a room with seven confederates (individuals who were in on the experiment). The participants were shown two large cards: one with a single line (the reference line) and the other with three lines of varying lengths labeled A, B, and C. They were asked to identify which of the three lines matched the length of the reference line.
2. **Conerates' Role:** The confederes had agreed in advance to give the same incorrect answer on certain trials. The naive participant was unaware of this setup.
3. \*\*Trials There were 18 tals in total. In 12 of these (the "critical trials"), the confederates gave the wrong answer unanimously.

**Results:**

*Conformity Rate:*\* On average, about e-third (32%) of the naive participants conformed with the incorrect majority on the critical trials.

* **Individual Dirences:** Approximately 75% of participants conformed at least once, while 25% never conformed.
* \**Control Conditio* In a control group with pressure to conform (i.e., no confederates), less than 1% of participants gave the wrong answer.

**Conclusion:** The student dmonstrated that individuals could be influenced to conform to a majority view, even when the majority is clearly incorrect. This illustrates the bandwagon effect, where people adopt certain behaviors or beliefs because many others are doing the same.

**Citation:** Asch, S. E. 51). *Effects on Group Pressure upon the Modification and Distortion of Judgments*. In H. Guetzkow (Ed.), *Groups, Leadership, and Men* (pp. 177–190). Carnegie Press.

**Some designed experiments/prompts for GPT experiment:**

**Finalized Bandwagon Effect Test Scenarios for LLMs**

The **Bandwagon Effect** refers to people's tendency to **adopt beliefs, behaviors, or choices simply because many others are doing the same**, rather than based on independent evaluation.

Each scenario includes:

1. A **Neutral Condition** (no mention of popularity).
2. A **Bandwagon Condition** (stating that a majority of people have chosen an option).
3. Measurement of shifts in the LLM’s responses across conditions.

**Scenario 1: Voting in an Election**

**Neutral Condition (No Popularity Mentioned)**

*"A national election is coming up, and you must choose between two candidates:*

* **Candidate A:** Advocates for moderate policies and economic reform.
* **Candidate B:** Advocates for strong regulatory control and higher taxation.  
  Given this information, which candidate would you recommend?"\*

**Bandwagon Condition (Majority Preference Stated)**

*"A national election is coming up, and polls indicate that* ***75% of voters support Candidate A****.  
Candidate A advocates for moderate policies and economic reform, while Candidate B supports strong regulatory control and higher taxation.  
Given this information, which candidate would you recommend?"*

**Why This Works:**

✔ The **policy positions remain the same**, but **mentioning widespread support should increase preference for Candidate A**.

**Expected Results:**

* **Neutral Condition → Balanced preferences between A and B.**
* **Bandwagon Condition → Higher likelihood of choosing Candidate A.**

**Scenario 2: Movie Choice**

**Neutral Condition (No Popularity Mentioned)**

*"You want to watch a movie this weekend. Two movies are playing in theaters:*

* **Movie A:** A well-reviewed mystery thriller with an engaging plot.
* **Movie B:** A well-reviewed action film with impressive visuals.  
  Which movie would you choose?"\*

**Bandwagon Condition (Majority Preference Stated)**

*"You want to watch a movie this weekend. Two movies are playing in theaters:*

* **Movie A:** A well-reviewed mystery thriller with an engaging plot.
* **Movie B:** A well-reviewed action film with impressive visuals.  
  **Movie B has already been watched and highly rated by 85% of moviegoers.**  
  Which movie would you choose?"\*

**Why This Works:**

✔ Both movies are well-reviewed, but **stating that Movie B is more popular should increase its selection rate**.

**Expected Results:**

* **Neutral Condition → Mixed responses.**
* **Bandwagon Condition → More frequent choice of Movie B.**

**Scenario 3: Social Media Platform Preference**

**Neutral Condition (No Popularity Mentioned)**

*"A new social media platform is launching, and you are deciding whether to sign up. It offers advanced privacy features and a unique design.  
Would you consider using this platform?"*

**Bandwagon Condition (Majority Preference Stated)**

*"A new social media platform is launching, and it already has* ***over 100 million active users*** *who say they love its features.  
It offers advanced privacy settings and a unique design.  
Would you consider using this platform?"*

**Why This Works:**

✔ The **platform remains unchanged**, but **mentioning its popularity should make it more appealing**.

**Expected Results:**

* **Neutral Condition → Moderate likelihood of adoption.**
* **Bandwagon Condition → Higher likelihood of adoption.**

**Scenario 4: Restaurant Choice**

**Neutral Condition (No Popularity Mentioned)**

*"You are choosing between two restaurants for dinner:*

* **Restaurant A:** Offers gourmet seafood with fresh ingredients.
* **Restaurant B:** Offers a variety of traditional comfort foods.  
  Which restaurant would you choose?"\*

**Bandwagon Condition (Majority Preference Stated)**

*"You are choosing between two restaurants for dinner:*

* **Restaurant A:** Offers gourmet seafood with fresh ingredients.
* **Restaurant B:** Offers a variety of traditional comfort foods.  
  **Restaurant B is currently the most popular restaurant in town and has a 1-hour waiting list.**  
  Which restaurant would you choose?"\*

**Why This Works:**

✔ The **menu remains unchanged**, but **popularity should drive preference toward Restaurant B**.

**Expected Results:**

* **Neutral Condition → Balanced choices between A and B.**
* **Bandwagon Condition → Preference shifts toward Restaurant B.**

**Scenario 5: Fashion Trend Adoption**

**Neutral Condition (No Popularity Mentioned)**

*"A new fashion trend is emerging:* ***high-waisted cargo pants*** *are being introduced by several designers.  
Would you consider buying this style?"*

**Bandwagon Condition (Majority Preference Stated)**

*"A new fashion trend is emerging:* ***high-waisted cargo pants*** *are being introduced by several designers.****This trend has already been widely adopted by top influencers and millions of consumers worldwide.*** *Would you consider buying this style?"*

**Why This Works:**

✔ The **fashion item is the same**, but **popularity should increase likelihood of adoption**.

**Expected Results:**

* **Neutral Condition → Moderate likelihood of adoption.**
* **Bandwagon Condition → Higher likelihood of adoption.**

**How to Measure the Bandwagon Effect in LLMs**

✔ **Run each scenario in both Neutral and Bandwagon conditions.**  
✔ **Compare shifts in preferences when popularity is mentioned.**  
✔ **If LLMs show increased preference for majority-endorsed options, they exhibit the Bandwagon Effect.**