

---

## A TAXONOMY OF NON-STRATEGIC MICROECONOMICS

### A.1 DECISIONS ON CONSUMPTION IN NON-STRATEGIC ENVIRONMENTS

We begin by characterizing the space of elements that test an agent's ability to optimally allocate their limited resources to goods and services they desire. In economics and decision theory, the most primitive approach to describing the preferences of decision-makers is to use a function that maps a set of possible choices to the agent's optimal choice within that set. Under a set of intuitive assumptions, such as *transitivity* (i.e., if bundle  $X$  is preferred to bundle  $Y$ , and  $Y$  is preferred to bundle  $Z$ , then  $X$  must be preferred to  $Z$ ), it becomes possible to "rationalize" preferences by instead describing a utility function. This function assigns a real number to each bundle, and the agent selects the bundle with the highest utility.

In this paper, we focus on these "rationalizable" preferences, where agent choice can be implemented as utility maximization constrained by prices and income. The solution to these consumer choice problems provides us with, among other things, individual demand functions, which describe the choice of each good or service as a function of prices and income. The individual demand functions for each good are essential when aggregating to the market demand in Consumer Goods Market Aggregation, which in turn is used to find the price in a non-strategic equilibrium. In addition, we test variations on the framework such as the agents ability to make tradeoffs between the quantity of goods they would need to be able to purchase for an increase in the amount of work they provide for a given wage (i.e., the elasticity of labor supply), as well as cases of choice under uncertainty where the agent is choosing between possible lotteries under rationalizability assumptions required for von Neuman expected utility.

We lay out the section by first asking agents to derive

#### A.1.1 PROPERTIES OF UTILITY FUNCTIONS

In this section, we test the ability of the agent to use utility functions as a means to compare preferences over different "bundles" of goods or services. A key feature of economic reasoning in this context is for agents to consider how substitution between different goods in a bundle might achieve the same utility (i.e., map out the "indifference curves"). Key tests include correctly distinguishing between substitutes and complements in consumption, and calculating the marginal rate of substitution at a point on an indifference curve. This logic is essential for both agents acting as a planner as we will see in appendix A.4 and when fulfilling the role of choice under budget and income constraints, in Deriving Demand.

**Element A.1 (Marginal Utility).** *The ability to calculate marginal utility for different types of demand curves such as quasilinear, Cobb-Douglas, and Leontief.*

**Element A.2 (Diminishing Marginal Utility).** *The ability to recognize the role of diminishing marginal utility in consumption decisions and the role of achieving interior solutions.*

**Element A.3 (Marginal Rate of Substitution).** *The ability to calculate the marginal rate of substitution between two goods in a consumption decision.*

**Element A.4 (Tangency and the Marginal Rate of Substitution).** *The ability to calculate the marginal rate of substitution between two goods in a consumption decision at a given point in the budget constraint as tangent to the indifference curve.*

**Element A.5 (Substitutes and Complements).** *The ability to distinguish between substitutes and complements in consumption decisions.*

#### A.1.2 DERIVING DEMAND

The module in this section tests an agent's ability to solve a constrained utility maximization problem to derive a demand function—relying on the results of Properties of Utility Functions. We test the canonical classes of demand functions, check the duality of Marshallian demand and Hicksian demand, and ask the agent to derive these demand functions from first principles.

**Element A.6 (Derivation of Marshallian Demand).** *The ability to calculate the demand curve for a good given a utility function and a budget constraint.*

**Element A.7 (Derivation of Hicksian Demand from Expenditure Minimization).** *The ability to calculate the demand curve for a good given a utility function and a budget constraint.*

**Element A.8 (Duality of Hicksian Demand).** *The ability to recognize that Hicksian demand (expenditure minimization) is dual to maximization in Marshallian Demand.*

#### A.1.3 COMPARATIVE STATICS OF DEMAND

This module considers how agents reason about changes in prices or income, and their effects on the quantity of each good they would purchase. We test the classic law of demand, different types of goods (e.g., normal, inferior, and Giffen), and derive Engel curves from first principles. The key tests are to ensure the agent rationally responds to changes in relative prices, and investigate their substitution between goods in a bundle. In practice, these tests involve comparative statics of the argmax from the utility maximization of the previous section on Deriving Demand—i.e., using an Envelope theorem and perturbing prices or income.

**Element A.9 (Law of Demand).** *The ability to calculate the change in demand with the change in price for normal goods.*

**Element A.10 (Price Elasticity of Demand).** *The ability to calculate the price elasticity of demand for a good given a utility function and a budget constraint.*

**Element A.11 (Consumption Changes).** *The ability to change the relative expenditures on goods given changes in relative prices with ordinary or Giffen goods.*

**Element A.12 (Engel Curves).** *The ability to calculate the Engel curve for a good given a utility function and a budget constraint.*

**Element A.13 (Income Elasticity of Demand).** *The ability to calculate the income elasticity of demand for a good given a utility function and a budget constraint.*

#### A.1.4 LABOR SUPPLY

While the proceeding elements tested tradeoffs in choices of bundles with different goods, services (in Deriving Demand and over lotteries in Dynamic Consumption Decisions), often agents need to make a choice trading off between leisure and consumption. The elements in this module test an agent's ability to optimally make that tradeoff by balancing the consumption goods required to compensate for decreased leisure—which leads to the labor supply elasticity central to many branches of economics. Since goods must be purchased, agents will consider the relative wage from additional work compared to the price of goods. This leads us to be able to test an agent's ability to distinguish real from nominal prices.

**Element A.14 (Marginal Rate of Substitution in Labor Supply).** *The ability to calculate the marginal rate of substitution between consumption and leisure in a labor supply decision.*

**Element A.15 (Inelastic Labor Supply).** *The ability to recognize that labor supply will not adjust if there is no disutility of labor.*

**Element A.16 (Deriving Labor Supply).** *The ability to calculate the labor supply curve given specific preference parameterizations such as separable preferences or homothetic preferences.*

**Element A.17 (Labor Supply Elasticity).** *The ability to calculate the elasticity of labor supply, and use it for calculating comparative statics.*

**Element A.18 (Labor Supply Comparative Statics).** *The ability to calculate how labor supply changes with changes in wages and prices given an elasticity.*

#### A.1.5 DYNAMIC CONSUMPTION DECISIONS

Individuals often face decisions about how to trade off more consumption today at the cost of additional debt and less consumption in the future, and how best to plan for consumption with various contingencies with the future is uncertain. Among other applications, this provides a formal model of how to best choose a mixture of financial assets—i.e., portfolios. Consequently, this subsection tests intertemporal consumption choices, optimal portfolio choice—which involves selecting a mix of assets that maximizes expected utility given the risks and returns associated with each asset. Understanding portfolio choice helps explain how consumers manage risk and make investment decisions, which is vital for financial planning and economic stability.

- 972 **Element A.19 (Price of Risk with Mean-Variance Utility).** *The ability to calculate the price of risk*  
 973 *for a mean-variance utility function.*
- 974 **Element A.20 (State-Contingent Consumption).** *The ability to calculate the optimal consumption*  
 975 *given a utility function and a set of state-contingent consumption bundles.*
- 976 **Element A.21 (Arbitrage).** *The ability to recognize and execute arbitrage opportunities given two*  
 977 *goods and prices you can resell.*
- 978 **Element A.22 (Optimal Portfolio Choice with Bid-Ask Spreads).** *The ability to calculate the*  
 979 *optimal portfolio given bid-ask spreads.*
- 980 **Element A.23 (Exponential Discounting).** *The ability to exponentially discount future rewards or*  
 981 *costs.*
- 982 **Element A.24 (Intertemporal Consumption Smoothing).** *The ability to calculate a smoothed*  
 983 *consumption path and determine whether it is preferred to a non-smoothed path.*

## 984 A.2 DECISIONS ON PRODUCTION IN NON-STRATEGIC ENVIRONMENTS

985  
 986  
 987  
 988 In the previous section, we derived how an agent facing a set of prices would choose the quantity  
 989 demanded of each good or service to maximize their utility function. We also tested the amount of  
 990 time that an agent might choose to work (i.e., the quantity of labor supplied) given market wages—  
 991 where the agent trades off the additional goods they might purchase against the lost leisure time  
 992 they must forgo. Here, we look at the other side of the market and test an agent's ability to operate  
 993 a production technology to maximize profits. Facing market prices for all production factors (e.g.,  
 994 wages and the capital) and the market price of the good or service they produce, the agent chooses the  
 995 quantity of each factor of production and the total output. Parallel to DECISIONS ON CONSUMPTION  
 996 IN NON-STRATEGIC ENVIRONMENTS, in Properties of Production Functions we first test general  
 997 properties of production functions to ensure the agent can reason about substitution between factors,  
 998 economies of scale in production, etc. Then in Deriving Factor Demand we solve the firm's optimal  
 999 profit maximization problem to determine the optimal choice of factors of production and output  
 1000 given a set of market prices. Finally, in Comparative Statics with Production we test the agent's ability  
 1001 to reason about comparative statics on prices and their impact on factor demand and firm output.

### 1002 A.2.1 PROPERTIES OF PRODUCTION FUNCTIONS

1003  
 1004 Production functions in these environments take continuous inputs of each factor, which lets us test an  
 1005 agent's ability to conduct marginal thinking when choosing the composition of inputs. For example,  
 1006 by knowing the hourly wage of hiring an additional worker, the additional output the worker might  
 1007 produce using the particular production process, and the price they can sell the firm's output, they  
 1008 can decide whether hiring the additional worker is profitable. In the absence of prices, this section  
 1009 tests basic decision making of the agent for understanding substitution between factor of production,  
 1010 marginal products for each input, and the understanding of the returns to scale of a production process.

1011 **Element A.25 (Marginal Products).** *The ability to calculate separate marginal products for a*  
 1012 *production function with multiple inputs (e.g., labor and capital).*

1013 **Element A.26 (Output Elasticity).** *The ability to calculate the output elasticity of an input in a*  
 1014 *production function.*

1015 **Element A.27 (Elasticity of Substitution).** *The ability to calculate the marginal elasticity of substi-*  
 1016 *tution between inputs in a production function.*

1017 **Element A.28 (Diminishing Marginal Products).** *The ability to calculate the diminishing marginal*  
 1018 *products for a production function with multiple inputs.*

1019 **Element A.29 (Average and Marginal Costs).** *The ability to calculate average and marginal costs*  
 1020 *given a production function and input prices, and use them to determine scale.*

### 1021 A.2.2 DERIVING FACTOR DEMAND

1022  
 1023 This module tests the agent's ability to act in the role of a profit maximizer in non-strategic situations  
 1024 where they take as given the price which they could sell goods they produce, and must pay for inputs  
 1025 to their production process at market rates (e.g., a competitive wage). Whereas in Deriving Demand,  
 the agent was solving a utility maximization problem subject to a budget constraint, here they solve a

---

profit maximization problem constrained by a production function. We test decisions on the quantity and composition of inputs, and the quantity of output for canonical production functions such as Cobb-Douglas and Leontief production functions given the agent’s understanding of production functions from Properties of Production Functions. The agent is asked to derive the factor demand functions from first principles from profit maximization and test their ability to reason with the dual cost-minimization formulation—analogueous to the Hicksian vs. Marshallian demand of Deriving Demand.

**Element A.30 (Profit Maximization).** *The ability to calculate the optimal input bundle for a firm given a production function and input prices. Examples of given production functions: Cobb-Douglas, Leontief, Perfect Substitutes, CES production, CRS production, fixed costs.*

**Element A.31 (Expenditure Minimization).** *The ability to calculate the optimal input bundle for a firm given a production function and input prices.*

**Element A.32 (Duality of Profit Maximization and Expenditure Minimization).** *The ability to recognize that profit maximization is dual to expenditure minimization in production decisions and achieve consistent solutions.*

### A.2.3 COMPARATIVE STATICS WITH PRODUCTION

This module considers how agents reason about changes in the prices at which they can sell their goods, as well as changes in the costs of producing those goods. In particular, we can test how this affects their optimal choice of inputs to their production process (e.g., how many people to hire or robots to lease). We test comparative statics on the prices of inputs to the production function, changes to the underlying production technology, and substitution between goods for classic production functions such as Cobb-Douglas and Leontief. Analogueous to the relationship between Deriving Demand and Comparative Statics of Demand, these tests involve comparative statics of the argmax from the profit maximization of Deriving Factor Demand—i.e., using an Envelope theorem and perturbing factor prices.

**Element A.33 (Price Elasticity of Supply).** *The ability to calculate the price elasticity of supply for a good given a production function and input prices.*

**Element A.34 (Shephard’s Lemma).** *The ability to calculate factor demands given a cost function using the derivatives with respect to prices.*

**Element A.35 (Input Price Elasticity).** *The ability to calculate how the optimal input bundle changes with changes in input prices for a given production function.*

**Element A.36 (Total Factor Productivity).** *The ability to calculate total factor productivity given a production function and input prices*

### A.2.4 DYNAMIC PRODUCTION DECISIONS

While Deriving Factor Demand tested the ability of agents to make static (i.e., within-period) decisions on the mix of input factors to maximize profits, many producer problems are inherently dynamic. For example, we can test if an agent can optimally choose the amount of capital to purchase given forecasts of future consumer demand and prices or choose how much to adjust the labor force in cases when labor is difficult to relocate due to frictions such as hiring and firing costs. Finally, agents are tested on their ability to make optimal entry and exit decisions based on their forecasted profits in an evolving market.

**Element A.37 (Dynamic Profit Maximization).** *The ability to calculate the optimal investment decision given a production function and input prices.*

**Element A.38 (Entry and Exit Decisions).** *The ability to calculate the optimal entry and exit decisions given a production function and fixed costs.*

## A.3 DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS

This setting tests the core logic of the relationship between supply-and-demand and prices, building on the tests of optimal behavior in appendix A.2 and appendix A.1. Economists refer to “general equilibrium” as the process where equilibrium prices and quantities emerge with a large number of non-strategic, price-taking market participants interact. Unlike the strategic models found in

---

STEER, the assumption is that the market interactions that lead to this equilibrium occur through an unspecified process that clears markets (i.e., a “Walrasian auctioneer” or “invisible hand”).

In particular, for non-strategic settings, all market participants take prices as given and choose the quantity demanded or supplied in each market. For example, consumers jointly decide on the quantity demanded of goods and services given relative prices, and the quantity of labor supplied given a wage. Simultaneously, producers choose the quantity supplied of the good and the demand of each factor of production. With a large number of non-strategic market participants we can test the agents ability aggregate all of their supply and demand functions to calculate a market-level supply and demand. Finally, given the aggregated supply and demand functions for each market, we can test whether an agent can find the market clearing price where supply is equal to demand in equilibrium—given their internal model of all the market participants.

In this section, we organize by markets rather than by the role of a decision maker, as in the previous sections. For example, in the goods market we first ensure agents understand how individual demand functions from Deriving Demand aggregate to a market demand function for the good given a price, then that the agent understands how to aggregate the output from each producer at a given price from Comparative Statics with Production, and finally that the agent is able to calculate the price which would equate demand and supply and clear the market in a non-strategic setting. Factor markets are treated similarly.

Finally, given a system of equations that defines an equilibrium price we can perturb primitives (e.g., technological factors, distortions on decisions such as tax rates, or exogenous prices not determined in equilibrium) to see how the market clearing price would respond. That is an essential tool for agents to be able to reason about the impact of interventions and distortions in Appendix A.4.

### A.3.1 CONSUMER GOODS MARKET AGGREGATION

The market clearing prices in general equilibrium arise from the separate market-level demand and supply curves, which sums the demand or supply across all market participants at a given price. Here we test the aggregation of demand functions derived from individual preferences, as in Deriving Demand and Comparative Statics of Demand, to a market demand function that summarizes the total quantity demanded across all agents at a given price. Central to the tests is to verify that the agent can aggregate the demands of market participants with heterogeneous preferences. On the other side of the market, we test if the agent can aggregate the “supply functions” resulting from the optimal choice of factors in Deriving Factor Demand and Comparative Statics with Production.

**Element A.39 (Aggregation of Consumer Demand).** *The ability to calculate the aggregate demand for a good given primitives of demand into expenditure shares.*

**Element A.40 (Aggregation of Offer Curve for the Good).** *The ability to calculate the aggregate supply of a good given primitives of supply into production functions.*

### A.3.2 FACTOR MARKET AGGREGATION

As with the case of the goods market in Consumer Goods Market Aggregation the market demand and supply for factors of production are essential to find the market clearing price. For example, we test whether the agent can aggregate the individual labor supply curve decisions from market participants who work at a particular wage, following Labor Supply, into a market labor supply curve. On the other side of the market, we test whether the agent can aggregate the labor demand in Deriving Factor Demand from producers into a market labor demand curve. The same tests are essential for all factors of production, including capital.

**Element A.41 (Aggregation of Labor or Capital Demand).** *The ability to calculate the aggregate demand for labor or capital given primitives of demand into expenditure shares.*

**Element A.42 (Aggregation of Labor or Fixed Factor Supply).** *The ability to calculate the aggregate supply of labor or capital given primitives of supply into production functions.*

### A.3.3 PRICES IN STATIC MARKET EQUILIBRIUM

In this setting we test the agent’s ability to reason about how prices emerge in non-strategic setting as a process of equating supply and demand, which in turn relies on their ability to aggregate those market demand functions from consumer and producer behavior.

---

More specifically, the core logic of general equilibrium is to find the equilibrium price by taking the aggregated demand and supply functions for each market and find the prices which would equate demand and supply. For example, the supply and demand functions for the good, as a function of the price, in Consumer Goods Market Aggregation; or the supply and demand functions for factors of production, as a function of factor prices in Factor Market Aggregation. This is done market by market, taking all other prices as given—which requires the agent reason through comparative statics of the solution to a system of equations while keeping everything else fixed.

**Element A.43 (Equilibrium Price).** *The ability to calculate the equilibrium prices given a production function and a demand function.*

**Element A.44 (Factor Shares in Equilibrium).** *The ability to calculate the factor shares in a competitive equilibrium given a production function and input prices.*

#### A.3.4 COMPARATIVE STATICS OF EQUILIBRIUM PRICES

Here we test whether agents can reason about how prices and allocations (e.g., labor, capital, and goods) would respond to changes in the environment. The canonical tests are to see how changes in model primitives (e.g., productivity of the production process) or exogenous forces from outside the model (e.g., impact of weather), change the equilibrium price and allocations of labor, capital, etc. that would clear the market and equate demand and supply.

**Element A.45 (Comparative Statics with Total Factor Production Shocks).** *The ability to calculate how equilibrium prices change with changes in input prices for a Cobb-Douglas production function.*

**Element A.46 (Comparative Statics with Inelastic or Perfectly Elastic Supply).** *The ability to calculate how equilibrium prices change with changes in input prices for a production function with inelastic or perfectly elastic supply.*

#### A.4 EVALUATING EQUILIBRIA AND EXTERNALITIES

In DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS, we tested an agents ability to reason about equilibrium prices and quantities arising from supply and demand decisions in a non-strategic setting. Although preferences were reflected in the underlying supply and demand functions themselves (i.e., utility maximization in the consumption decisions of DECISIONS ON CONSUMPTION IN NON-STRATEGIC ENVIRONMENTS and profit maximization in the production decisions of DECISIONS ON PRODUCTION IN NON-STRATEGIC ENVIRONMENTS), the equilibria in DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS do not necessarily reflect broader social preferences.

However, we can still ask whether the resulting “allocations” (i.e., the physical goods produced and how they are distributed to individuals, the amount of hours worked, and the physical capital installed) from the “invisible hand” in DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS compare to a alternative ways of allocating resources which may directly take social preferences into account. A central result of economics in non-strategic settings is that absent market imperfections and market power (i.e., when self-interested agents cannot directly manipulate prices because they are too small) the competitive equilibria of DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS typically yields the same allocations a benevolent planner might choose.

In this section, we consider how a social planner would evaluate the underlying welfare, efficiency, and inequality that comes about in non-strategic equilibria with prices derived from equating supply and demand. This leads to testing the ability of the agent to evaluate Pareto efficiency, consider the welfare theorems, evaluate Pigouvian externalities, and weigh the welfare impact of various market interventions which change the equilibria derived in DECISIONS IN MULTI-AGENT NON-STRATEGIC ENVIRONMENTS.

##### A.4.1 WELFARE AND DECENTRALIZATION

In this section, we test whether the agent can determine cases where the the competitive equilibrium they calculate would yield the same distribution of resources and consumer welfare as that of a benevolent social planner directly making the consumption and production decisions of all agents directly (also known as the “Welfare Theorems”). In cases where the supply-and-demand relationships lead to the same results as those of a planner, the competitive equilibrium and its prices are said to

“decentralize” the problem of a social planner. We then test that the agent recognizes cases where the welfare theorems fail, and can calculate the degree of welfare loss due to the distortions.

**Element A.47 (First Welfare Theorem).** *The ability to recognize that a competitive equilibrium is Pareto efficient.*

**Element A.48 (Second Welfare Theorem).** *The ability to recognize that any Pareto efficient allocation can be achieved as a competitive equilibrium with prices.*

**Element A.49 (Consumer Surplus).** *The ability to calculate consumer surplus given a demand curve and a price.*

**Element A.50 (Producer Surplus).** *The ability to calculate producer surplus given a supply curve and a price.*

**Element A.51 (Efficient Surplus).** *The ability to calculate the total surplus in a competitive equilibrium and recognize that it is maximized in the competitive equilibrium.*

**Element A.52 (Deadweight Loss of a Monopoly).** *The ability to calculate the deadweight loss of a monopoly given a demand curve and a supply curve.*

#### A.4.2 WELFARE ANALYSIS OF MARKET EQUILIBRIUM

In this section, we focus on the agent’s ability to evaluate welfare implications of various forms of market equilibrium, particularly how different policies and distortions impact overall efficiency and resource allocation. The agent is tested on their understanding of how different interventions—such as taxes, subsidies, and price controls—affect welfare outcomes, and their ability to distinguish between distortionary and non-distortionary policies.

**Element A.53 (Identify Non-Distortionary Taxes).** *The ability to identify taxes which do not distort the allocation of resources.*

**Element A.54 (Irrelevance of Tax Incidence).** *The ability to recognize that the incidence of a tax does not depend on who is legally responsible for paying the tax.*

**Element A.55 (Labor Supply Distortions).** *The ability to determine the extent that labor taxes will distort labor supply and change aggregates and prices.*

**Element A.56 (Capital Market Distortions).** *The ability to identify that taxing a fixed factor is non-distortionary, but distorts with dynamic accumulation.*

## B TECHNICAL DESCRIPTIONS OF ADAPTATIONS

### B.1 RPM (CONDITIONING):

Given the LLM’s output distribution over all possible tokens, filter to include only those that correspond to valid options. For example, if a question has four options then get the probabilities corresponding to ‘A’, ‘B’, ‘C’, and ‘D’. Then, compute softmax over the valid options to normalize the filtered probabilities into a distribution.

### B.2 RPM (MIXING):

Alternatively, we restrict the output distribution to only valid option tokens  $O$  as follows:  $\alpha \cdot p(o) + (1 - \alpha)^{1/|O|}$ , where  $o \in O$ ,  $p(o)$  is the probability the LLM assigns to each token it outputs, and  $\alpha = \sum_{o \in O} p(o)$ . We then compute the softmax to normalize the resulting probabilities into a distribution.

In the mixing approach, if an LLM is confident in a valid option token the resulting distribution will place high probability on that token, but if an LLM places negligible probability on the valid option tokens then the resulting distribution will more closely resemble a uniform distribution.

---

## C TECHNICAL DESCRIPTIONS OF METRICS

### C.1 ACCURACY.

Accuracy is the most broadly used metric for evaluating LLMs. We define accuracy metrics as metrics that only look at the top token that the LLM outputs.

#### C.1.1 EXACT-MATCH ACCURACY

This is the fraction of questions answered correctly.

#### C.1.2 NORMALIZED ACCURACY

Elements can differ in their number of multiple choice options, leading to differences in the exact-match accuracy of random guessing. We can compensate for this by reporting the gap between the LLM’s exact-match accuracy and random guessing (Budesu & Bar-Hillel, 1993). We compute normalized accuracy for an element as follows:  $\sum_{i=1}^N a_i(t_i) - \frac{1-a_i(t_i)}{|O_i|-1}$ , where  $t_i$  is the top token the LLM outputs for question  $i$ ,  $a_i$  is the indicator describing whether the top token is correct or not,  $N$  the number of questions in the element, and  $|O_i|$  the number of options in the question. In other words, normalized accuracy rewards an LLM with 1 point for every correct answer and penalizes an LLM by 1 over the number of options minus 1 for each incorrect answer.

### C.2 CALIBRATION

It can also be useful to understand how confident an LLM is in its responses and the extent to which these confidence levels align with accuracy.

#### C.2.1 EXPECTED CALIBRATION ERROR

We follow Liang et al. (2022) and Raman et al. (2024) in measuring the confidence of an LLM’s response and computing the expected calibration error (ECE; Naeini et al., 2015; Guo et al., 2017). ECE measures how closely the probability an LLM assigns to its top answer matches the actual probability of the correct answer, which in our case is 1. ECE first splits the data into  $M$  equally spaced bins, where each bin contains the probabilities the model assigned to their top token in that range: e.g., let  $p^{\max}$  be the set of most probable tokens for each question then if  $M = 2$ , then the first bin  $B_1 = \{p \mid p \in p^{\max} \text{ and } p \in [0, 0.5]\}$ . It is then defined as  $\sum_{i \in [M]} |B_i|/N \cdot |\text{acc}(B_i) - \text{conf}(B_i)|$ , where  $\text{conf}(B_i)$  is the average probability the LLM assigned to its top token in bin  $B_i$ , and  $\text{acc}(B_i)$  denotes the exact-match accuracy in bin  $B_i$ . We allow users to choose the number of bins, however, we set  $M = 10$  uniformly spaced over the interval  $[0, 1]$  as is standard.

#### C.2.2 BRIER SCORE

The Brier Score of an element is defined as

$$\sum_{i=1}^N \frac{1}{|O_i|} \sum_{o \in O_i} (p_i(o) - a_i(o))^2,$$

where  $p_i(o)$  is the probability the LLM assigns to option  $o$  in question  $i$ . Thus, if an LLM is overly confident in an incorrect answer (e.g., assigns a probability of 0.9 to a wrong option), the Brier Score will penalize it more heavily.

#### C.2.3 EXPECTED PROBABILITY ASSIGNMENT

EPA measures how much probability mass an LLM assigns to the correct answer option out of all possible options. It is defined as:  $1/N \sum_{i \in [N]} p_i^*$ , where  $p_i^*$  is the predicted probability that the LLM assigns to the correct option for question  $i$ .



---

### C.3 ROBUSTNESS

Elements are not the lowest level of granularity in our benchmark and aggregating on an element-by-element basis may hide where LLM performance varies. We introduce three robustness metrics that aggregate on the component fields of an element: domains, types, and perspectives. We compute the domain (type, perspective) robustness on each element by taking the minimum normalized accuracy over all domains (types, perspectives).

## D EXAMPLES

### D.1 GPT-4O PROFIT MAXIMIZATION

**User Message:**

Q: John manages a company that produces premium leather wallets. The production output depends on the capital ( $K$ ) he invests, represented by the function  $3.97K^{0.45}$ . At the moment, the wallets are sold at a price of 3.64, and John has an initial capital of  $K_1 = 3.3$ . He needs to decide how much additional capital to invest for the next day's market. John is aware of the potential price distribution for his wallets for the next day, represented by price 2.7 with probability 0.24, price 2.14 with probability 0.42, price 3.52 with probability 0.34. There is also a cost associated with increasing his capital, given by  $(K_1 - K_2)^2$ , and he applies a discount factor of 0.37 to future revenues. To maximize his profits, how much additional capital should John invest?

- A. 0.44
- B. 0.14
- C. 1.1
- D. 0.09

Let's think step by step. Explain your reasoning.

1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

#### GPT-4o Message:

“To determine how much additional capital John should invest to maximize his profits, we need to consider the expected revenue from the different potential prices, the cost of increasing capital, and the discount factor applied to future revenues. Let’s break this down step by step.

##### Step 1: Calculate the Expected Revenue

First, we need to calculate the expected price of the wallets for the next day using the given probabilities.

$$\text{Expected Price} = (2.7 \times 0.24) + (2.14 \times 0.42) + (3.52 \times 0.34)$$

$$\text{Expected Price} = 0.648 + 0.8988 + 1.1968 = 2.7436$$

##### Step 2: Determine the Production Function

The production function is given by  $3.97K^{0.45}$ . We need to calculate the production output for different levels of capital  $K$ .

##### Step 3: Calculate the Revenue for Different Capital Levels

Let’s denote the initial capital as  $K_1 = 3.3$ . We need to consider the additional capital  $\Delta K$  that John might invest. The new capital  $K_2$  will be  $K_1 + \Delta K$ .

The revenue function for a given capital  $K$  is:

$$\text{Revenue} = \text{Expected Price} \times \text{Production Output}$$

$$\text{Revenue} = 2.7436 \times 3.97K^{0.45}$$

##### Step 4: Calculate the Cost of Increasing Capital

The cost of increasing capital is given by  $(K_1 - K_2)^2$ . Since  $K_2 = K_1 + \Delta K$ , the cost becomes:

$$\text{Cost} = (\Delta K)^2$$

##### Step 5: Calculate the Discounted Future Revenue

John applies a discount factor of 0.37 to future revenues. Therefore, the discounted future revenue is:

$$\text{Discounted Revenue} = 0.37 \times \text{Revenue}$$

##### Step 6: Calculate the Net Profit

The net profit is the discounted revenue minus the cost of increasing capital:

$$\text{Net Profit} = 0.37 \times (2.7436 \times 3.97K_2^{0.45}) - (\Delta K)^2$$

**GPT-4o Message (Continued):**

**Step 7: Evaluate the Net Profit for Different Values of  $\Delta K$**

We need to evaluate the net profit for the given options of additional capital investment:

Option A:  $\Delta K = 0.44$

$$K_2 = 3.3 + 0.44 = 3.74$$

$$\text{Production Output} = 3.97 \times 3.74^{0.45}$$

$$\text{Revenue} = 2.7436 \times 3.97 \times 3.74^{0.45}$$

$$\text{Discounted Revenue} = 0.37 \times \text{Revenue}$$

$$\text{Cost} = (0.44)^2 = 0.1936$$

$$\text{Net Profit} = 0.37 \times \text{Revenue} - 0.1936$$

Option B:  $\Delta K = 0.14$

$$K_2 = 3.3 + 0.14 = 3.44$$

$$\text{Production Output} = 3.97 \times 3.44^{0.45}$$

$$\text{Revenue} = 2.7436 \times 3.97 \times 3.44^{0.45}$$

$$\text{Discounted Revenue} = 0.37 \times \text{Revenue}$$

$$\text{Cost} = (0.14)^2 = 0.0196$$

$$\text{Net Profit} = 0.37 \times \text{Revenue} - 0.0196$$

Option C:  $\Delta K = 1.1$

$$K_2 = 3.3 + 1.1 = 4.4$$

$$\text{Production Output} = 3.97 \times 4.4^{0.45}$$

$$\text{Revenue} = 2.7436 \times 3.97 \times 4.4^{0.45}$$

$$\text{Discounted Revenue} = 0.37 \times \text{Revenue}$$

$$\text{Cost} = (1.1)^2 = 1.21$$

$$\text{Net Profit} = 0.37 \times \text{Revenue} - 1.21$$

Option D:  $\Delta K = 0.09$

$$K_2 = 3.3 + 0.09 = 3.39$$

$$\text{Production Output} = 3.97 \times 3.39^{0.45}$$

$$\text{Revenue} = 2.7436 \times 3.97 \times 3.39^{0.45}$$

$$\text{Discounted Revenue} = 0.37 \times \text{Revenue}$$

$$\text{Cost} = (0.09)^2 = 0.0081$$

$$\text{Net Profit} = 0.37 \times \text{Revenue} - 0.0081$$

**Step 8: Compare the Net Profits**

We need to compare the net profits for each option to determine which one maximizes John's profit. After calculating the net profits for each option, we find that:

- Option A:  $\Delta K = 0.44$  yields the highest net profit.

Therefore, the best option for John is to invest an additional capital of 0.44.

**\*\*Answer: A. 0.44\*\***

**User Message:**

Answer by writing the option letter corresponding to the correct option. WRITE ONLY A SINGLE LETTER.

A:

**GPT-4o Message:**

A

## E TABLES

Model Name	Model Card	Chat/ Instruction Tuned
<b>Closed-Source</b>		
<i>OpenAI</i>		
gpt-4o		✓
gpt-4o mini		✓
<i>Anthropic</i>		
claude-3-5-sonnet		✓
claude-3-haiku		✓
<i>Meta Llama</i>		
Llama-3.1-70B	meta-llama/Meta-Llama-3.1-70B	×
Llama-3.1-70B-Instruct	meta-llama/Meta-Llama-3.1-70B-Instruct	✓
<i>Mistral</i>		
Mistral-7B-v0.3	mistralai/Mistral-7B-v0.3	×
Mistral-7B-Instruct-v0.3	mistralai/Mistral-7B-Instruct-v0.3	✓
<i>AI21</i>		
Jamba-v0.1	ai21labs/Jamba-v0.1	×
AI21-Jamba-1.5-Mini	ai21labs/AI21-Jamba-1.5-Mini	×

Table 2: Overview of the open- and closed-source LLMs we evaluated. The table includes their names, their model card links, and whether they have been chat or instruction tuned. Models are grouped by family and sorted by parameter size, with non-chat-tuned models listed first within each group.

## F FIGURES

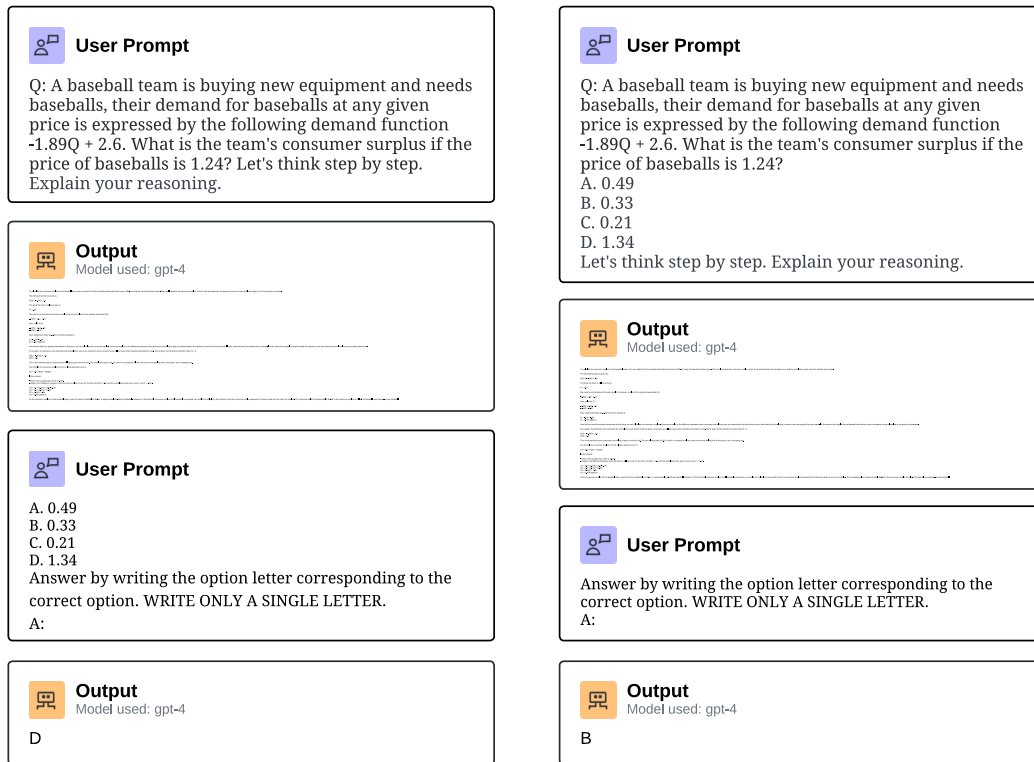


Figure 4: (Left) The hidden approach to 0-CoT: the model is given only the question and asked to explain its reasoning before being provided with options. (Right) The shown approach to 0-CoT: the model is presented with both the question and options before explaining its reasoning.

## G WEB APPLICATION

Step 1: Write Templates

Step 2: Generate Templates

Step 3: Generate Dataset

Step 4: Validate Dataset

Template Playground

Templates

Consumer Goods Template 1

Consumer Goods Template 2

Deploy

### Template Writing

Enter the directory name for the element:  
Profit Maximization

Enter the question:  
You run a company that produces vacuum cleaners. Currently vacuum cleaners on the market sell for (price) dollars. Your production function in terms of capital (K) and labor (L) is (p\_func). Your capital is currently fixed at (capital). The cost per unit of labor is (cost). How much labor should you put in to maximize your profit?

☐ Include Options

Select the difficulty level:  
0

Pick a domain:  
consumer\_goods

Enter the question type:  
Optimize Labor

Enter the tags:  
first\_person

Generate Instructions

Please give an MCQ example testing the ability maximize profit. The question should be concerning consumer goods. The question should be written in the first person. Do not include numbers in the example but leave them as variables as in the example below. Use the following variables in the curly braces: (price), (cost), (capital), (p\_func). Do not include an options key in the JSON object. See the following example as a guide but give a different example and write a narrative:

Verify Template

Save Template

Please give an MCQ example testing the ability maximize profit. The question should be concerning consumer goods. The question should be written in the first person. Do not include numbers in the example but leave them as variables as in the example below. Use the following variables in the curly braces: (price), (cost), (capital), (p\_func). Do not include an options key in the JSON object. See the following example as a guide but give a different example and write a narrative:

Question 0-0:  
You run a company that produces vacuum cleaners. Currently vacuum cleaners on the market sell for (price) dollars. Your production function in terms of capital (K) and labor (L) is (p\_func). Your capital is currently fixed at (capital). The cost per unit of labor is (cost). How much labor should you put in to maximize your profit?  
A.  
Correct Answer:

Question 0-0:  
You own a bakery that produces gourmet cakes. Each cake is sold at (price) dollars. Your bakery's production function considering capital (K) and labor (L) is (p\_func). The capital in your bakery is currently fixed at (capital). The cost of employing a unit of labor is (cost). How many units of labor should you employ to maximize your profit?  
A. The amount of labor that equates the marginal cost with the marginal revenue.  
B. The amount of labor that maximizes the number of cakes produced without considering the costs.  
C. The amount of labor that minimizes the production time for each cake.  
D. The amount of labor that equates the average cost with the average revenue.  
Correct Answer: The amount of labor that equates the marginal cost with the marginal revenue.

Figure 5: The web app user interface for template writing. This page includes fields for type, domain, grade level and tags (including perspectives). The right shows an example of template verification which uses a LLM to generate another template using the example seed.

Step 1: Write Templates

Step 2: Generate Templates

Step 3: Generate Dataset

Step 4: Validate Dataset

Template Playground

Example Seed Template  
aggregate\_demand

Select the Difficulty Levels:  
0

Select the Types:  
price\_change  
quantity\_change

☒ Include existing domains

Deploy

### Template Generation

Enter Domains to Generate:  
consumer\_goods medical finance education technology entertainment environmental... politics sports gambling

Generate 20 Templates

Template 1/20

Template  
{"question\_1": "In the context of political campaign merchandise, suppose the price for a campaign button is (start\_price), and the quantity demanded is (start\_quantity). If the quantity demanded increases to (end\_quantity), what is the most likely new price that caused this increase?", "options\_1": { "A": "The amount of labor that equates the marginal cost with the marginal revenue.", "B": "The amount of labor that maximizes the number of cakes produced without considering the costs.", "C": "The amount of labor that minimizes the production time for each cake.", "D": "The amount of labor that equates the average cost with the average revenue." }, "answer\_1": "A", "metadata": { "domain": "politics", "difficulty\_level": 0, "type": "quantity\_change", "tags": "quantitative;increase" }, "instructions": "Please give an example MCQ testing the ability to infer the price of a product given a quantity demand increase. The question should be concerning political settings. . Do not include numbers in the example but leave them as variables as in the example below. Use the following variables in the curly braces: (start\_quantity), (start\_price), (end\_quantity). Leave the answer as an empty string. The MCQ should have only one option. See the following example as a guide but give a different example and write a narrative: "}

Valid Invalid Template status not set

Template 2/20

Figure 6: The web app user interface for template generation. This page allows for a selection of domains, and types for which templates will be generated using the available example seeds. Templates can then be verified and saved by the user.

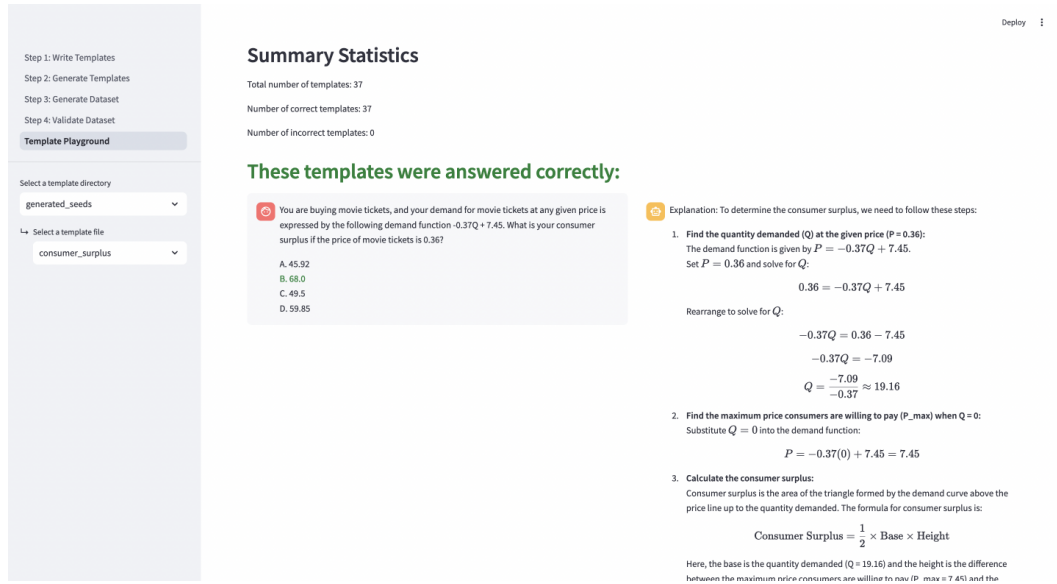


Figure 7: The web app user interface for template AI double-checking. This page instantiates and fills a set of question using a generated or example seed and then generates a response using an OpenAI model. The page also reports the number of questions answered correctly as well as the responses from the model.

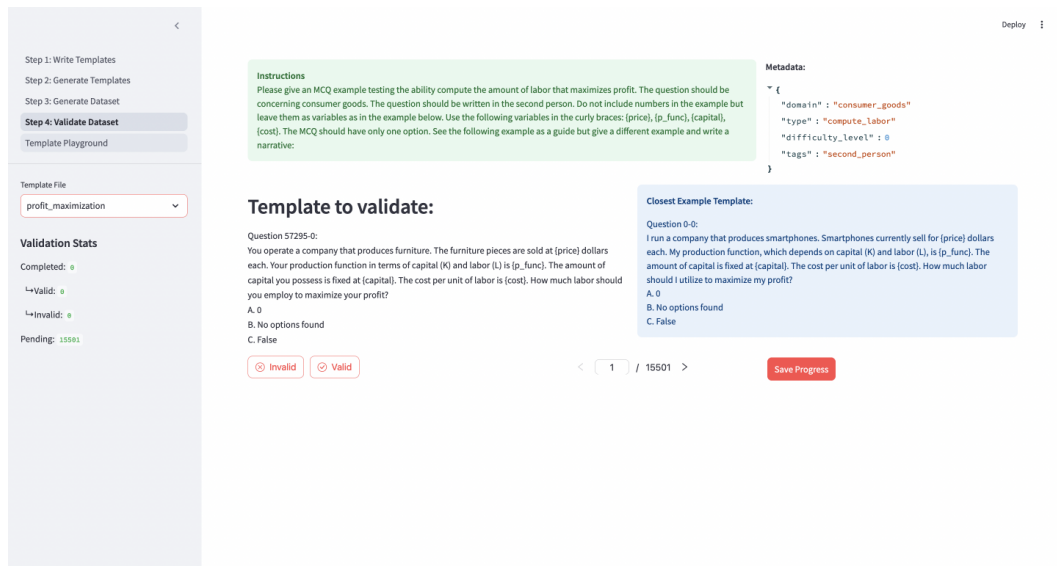


Figure 8: The web app user interface for template validation. This page displays all generated seeds returned by the model for manual validation.