BizFermi：フェルミ推定による ビジネスアイデアの市場規模推定データセット

高橋洸丞 [1]　広田航 [1]　進藤尚希 [1,2]　有馬幸介 [1]　石垣達也 [3]
[1] ストックマーク株式会社 [2] 電気通信大学
[3] 産業技術総合研究所
{kosuke.takahashi, wataru.hirota, naoki.shindo, kosuke.arima}@stockmark.co.jp
ishigaki.tatsuya@aist.go.jp

概要

LLM による事業アイデア生成において市場規模推定は重要だが、その評価基盤は未整備である。本研究では、ビジネスアイデアに対するフェルミ推定の「推定値」と「導出ステップ」を人間と LLM 双方で収録したデータセット「BizFermi」を構築した。分析の結果、LLM は数値の妥当性が高い一方でステップの論理性が低い傾向があることや、人間同士でも推定過程の不一致が大きいことなどを明らかにした。本データは公開予定である。

1　はじめに

大規模言語モデル（LLM）の発展により、創造的なアイデア生成の研究が注目されている。この潮流は、AI for Science における研究アイデア生成 [1] からビジネスアイデア生成 [2] に至るまで、広範な分野へと波及している。

LLM により高速かつ大量のアイデア生成が可能になったことで、現在は大量の候補を自動生成し、その中から有望なアイデアを選抜（ランキング）するアプローチが主流となりつつある。 本研究では、特に需要の強い新規事業策定のためのビジネスアイデア生成に着目する。 新規事業の意思決定においては、アイデアの実現可能性や新規性に加え、最終的に「どれだけの収益が見込めるか」という市場規模の推定が決定的な要因となる [3, 4]。 しかし、ビジネスアイデアと市場規模推定に関しては公開データセットが存在せず、そもそも人間や LLM がどのような論理ステップを経て市場規模を算出し、その推定値がどの程度妥当であるかは明らかになっていない。

ビジネス実務における市場規模推定では、TAM

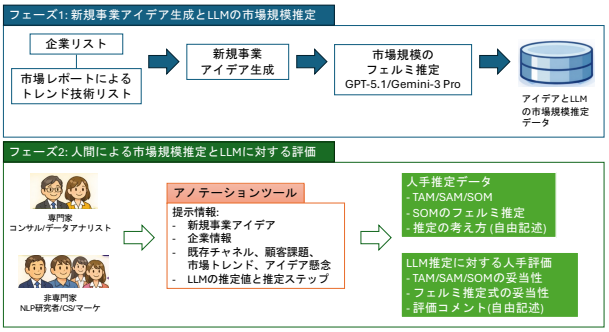

図 1　BizFermi 作成のイメージ

（理論上狙える最大市場）、SAM（現実に狙える提供範囲）、SOM（短中期に獲得可能な市場）という 3 層の指標が標準的に用いられる。 しかし、新規事業の立ち上げ初期段階では、これらの算出に必要な統計データが十分に整備されていないことが多い。 そのため、ターゲット層の人口や予定単価などの推定可能な因子を設定し、論理的な仮定を積み上げて最終的な数値を導出する「フェルミ推定」の手法が、市場規模推定の初期段階ではよく用いられる。このときフェルミ推定は、単に最終的な数値が当たっているかだけでなく、そこに至るまでの因数分解や計算ステップが尤もらしいかどうかが、推定結果の説明性と納得感を支える重要な要素になる [5]。

市場規模予測をタスクとして扱う場合、最終的な「推定値」だけでなく、それを導くための「推定ステップ（因数分解と仮定）」の両方が評価の対象とならなければならない。実際、市場規模推定の品質は、フェルミ推定においてどの因子を採用し、どのような値を置き、どの順序で積み上げたかという推定ステップの妥当性に強く依存する。近年、LLM に数値を算出させること自体は容易になった一方で、推定が誤ったときに「数値設定が誤っている」のか、それとも「論理構造（推定ステップ）自

体が破綻している」のかを区別できなければ、モデルやプロンプトの改善方針を定めにくい。さらに、比較対象となる人間の推定も、個人の実務経験や前提知識に強く依存する。市場規模推定は正解が一つに定まらないタスクであり、専門家と非専門家では因数分解の粒度や納得の基準が大きく異なりうる。[6, 7]。 ゆえに、LLM の自動評価基準を確立するためには、人間同士の推定がどこまで一致し、どこで分岐するのかという構造的理解が不可欠である。

本研究ではこの問題意識に基づき、ビジネスアイデアを対象に、人手によるフェルミ推定（推定ステップおよび推定値）を収集し、LLM による推定と比較・評価可能なデータセット **BizFermi** を構築した（図 1）。本データセットを用い、推定値と推定ステップを分離して評価することで、LLM の適用可能性と限界、および人間による推定のばらつきの構造と専門性による差異を明らかにすることを目指す。

本研究の貢献　本稿の貢献は以下の通りである。

- BizFermi データセットの構築：ビジネスアイデア市場規模推定に特化し、人手フェルミ推定（推定ステップ＋推定メモ＋数値）と、LLM 推定への人手評価を同一条件で収録（公開予定）。
- 推定値と推定ステップを分離した評価枠組み：市場規模の推定値と、因数分解（推定ステップ）の妥当性を切り分けて評価可能にし、誤りの所在を観測できる形にした。
- 専門家／非専門家差と人間不一致の構造化：推定のどの部分で人間が分岐し、専門性が評価にどう影響するかを分析し、将来のモデル化・評価設計に必要な論点を提示する。

本研究では、LLM を市場規模推定に用いる際の信頼性と、評価の基準となる人間推定の構造を明らかにするため、次の RQ (Research Question) を定める。**RQ1:** LLM はビジネスアイデアに対して妥当な TAM/SAM/SOM を提示できるか。**RQ2:** LLM は SOM の因数分解（フェルミ推定ステップ）を妥当な形で構成できるか。**RQ3:** 市場規模推定において、専門家と非専門家の推定はどの段階でどのように分岐するか。

2　問題設定

本研究は、企業×技術を前提として記述された新規事業アイデアを入力とし、SOM（市場投入後 3 年以内に狙える市場規模）の推定を主な対象とする。ここで重視するのは最終的な数値そのものではなく、因子へ分解して仮定を明示しながら積み上げるフェルミ推定の過程である。たとえば「製造業向け予知保全 AI」というアイデアでは、SOM を「3 年以内に獲得できる導入社数 × 1 社あたり年間単価」と分解できる。導入社数を「ターゲット企業数 2,000 社 × 到達率 25% × 成約率 12%」と置けば 60 社となり、年間単価を 800 万円と仮定すれば SOM は約 4.8 億円/年になる。ただし、顧客単位を企業ではなく工場で数える、単価をライン数ベースにする、継続率やアップセルを組み込むといった因子設計の違いだけで、推定値は桁レベルで変わりうる。

このため、LLM の推定が外れたときも、数値の過大・過小なのか、因子分解の構造が崩れているのか（欠落・過剰・関係の取り違え）、あるいは両方なのかを切り分けて検証できなければ改善方針を定めにくい。そこで本研究では、推定値と推定ステップを分離して評価できる設計を採用し、LLM と人間の差異を自然に観測できるようにする。

3　アノテーションデータの収集

BizFermi は、同一サンプルに対して (A) 人手によるフェルミ推定 と (B) LLM によるフェルミ推定の人手評価 を併せて収録するデータセットである。各サンプルは「企業 × 技術 × 新規事業アイデア」から構成され、技術は IMARC 等の市場調査レポートでトレンドとして取り上げられる技術群からランダムに選定し、gemini-3-flash-preview を用いてアイデアを生成する。

アノテーションは、専門家としてフェルミ推定の実務経験を持つコンサルタントとデータアナリストの 2 名、非専門家として NLP 研究者やカスタマーサポート、マーケット担当など 6 名が担当した。アノテーションツールでは、各新規事業アイデアについて、(i) 新規事業アイデア本体、(ii) 企業の既存チャネル・顧客課題・アイデアの懸念点・関連市場情報、(iii) GPT-5.1/gemini-3-pro-preview による市場規模の推定値・説明および推定ステップを提示した。

その上でアノテータには、まず (A) 人手推定 として、SOM のフェルミ推定（分解式／推定ステップ）と推定の考え方（自由記述）を記入し、加えて TAM・SAM の推定値も入力してもらった。SOM はフェルミ推定結果から自動計算される。この際、SOM は「市場投入後 3 年以内に狙える市場規模」として推定するよう統一的に指示した。

続いて (B) LLM 推定の評価（GPT-5.1 / gemini-3-pro-preview）として、TAM・SAM・SOM の妥当性を 6 段階（大きく過大／過大／妥当／過小／大きく過小／その他）で評価し、さらにフェルミ推定の式についても 4 段階（過大になりそう／妥当／過小になりそう／不明）で評価したうえで、LLM のフェルミ推定に対する所見を自由記述で記録した。

4　収集したデータの分析

本章では、BizFermi に収録した人手推定と LLM 推定を用い、研究質問 RQ1–RQ3 に答える。分析対象は、アノテーション総数 124 件であり、同一アイデアに複数アノテータが付与した結果を含む。対応する独立したアイデア数は 43 件である。

4.1　RQ1: LLM はビジネスアイデアに対して妥当な TAM/SAM/SOM を提示できるか

図 2 に、LLM が提示した市場規模（TAM/SAM/SOM）が妥当と評価された割合を示す。専門家の評価では、推定値の妥当率は概ね 3〜6 割に分布し、特に SOM では Gemini-3 Pro が約 6 割、GPT-5.1 も約 5 割である。。一方、非専門家の評価では同じ推定値に対する妥当率が 2〜3 割程度まで低下する。LLM の推定値は「一定程度受け入れられる」局面がある一方で、その評価像は一様ではなく、評価者の専門性に依存することが分かる。

この差が単なる採点の甘辛に留まらないことは、表 1 からも示唆される。専門家では、LLM-SOM と人間 SOM の相関はほぼゼロであり、LLM が示す SOM の大小関係が専門家の推定と整合しにくい。一方、非専門家では Gemini-3 Pro が強い相関を示す（Spearman 0.68, Kendall 0.55）。よって RQ1 に対して分析結果からわかることは、推定値の妥当性は約 6 割程度で、その妥当性判断や人間推定との整合は専門性によって大きく変わるということである。

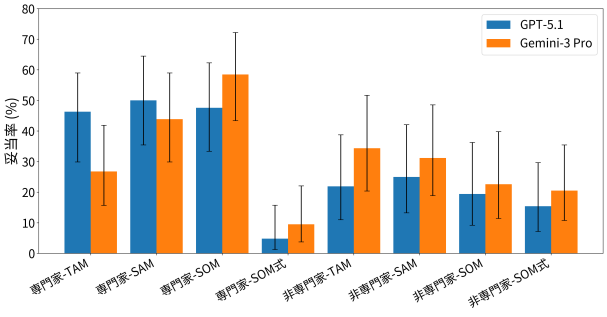

図 2　LLM 推定の妥当率. SOM 式はフェルミ推定の因数分解の妥当性評価. 縦ひげは 95% 信頼区間.

表 1　人間 SOM と LLM-SOM の相関
r:Pearson(log10), ρ:Spearman, τ:Kendall, ξ:Chatterjee

グループ	モデル	r	ρ	τ	ξ
専門家	GPT-5.1	-0.01	0.00	0.02	0.06
専門家	Gemini-3 Pro	-0.06	-0.06	-0.03	-0.11
非専門家	GPT-5.1	0.03	0.20	0.19	0.21
非専門家	Gemini-3 Pro	0.51	0.68	0.55	0.39

4.2　RQ2: LLM はフェルミ推定ステップを妥当な形で構成できるか

図 2 で、SOM 式について妥当と評価された割合も示す。ここで最も顕著なのは、推定値よりも推定ステップの妥当率が大幅に低い点である。専門家評価では SOM 式の妥当率は約 5〜10%に留まり、非専門家でも約 15〜20%程度である。LLM が提示する数値が一定割合で妥当と見なされる場合があっても、その背後の因数分解は妥当と判断されにくいというねじれが存在している。

さらに、専門家／非専門家の関係が推定値の場合と逆転する点も重要である。推定値については専門家の方が相対的に高く妥当と判断する一方で、推定ステップについては専門家の方がより厳しく判定している。これは、推定ステップの評価が、重要因子の欠落・過剰や因子間関係の取り違え、SOM（市場投入後 3 年以内）と整合しない仮定の混入といった「構造」の検査を要するためだと解釈できる。

以上より、RQ2 に対する結論として、現状の LLM は推定ステップの構成にボトルネックを抱えており、特に専門家ほどその因数分解を尤もらしいものとして受け取りにくい傾向が示された。

4.3　RQ3: 専門家と非専門家の推定・評価はどの段階でどのように分岐するか

RQ3 を検討するため、推定が分岐する中身を、(a) アノテータ間の一致、(b) 推定作法（観点・単価・時間軸・TAM 定義）、(c) 因数分解に含める因子の種類、の順に整理する。

(a) 人間同士の一致の分岐（SOM の相関）図 3 の Spearman 相関ヒートマップでは、人間同士でも相関は一様ではない。専門家同士でも高い一致が保証されず、非専門家同士には強い負の相関も見られる（例：-0.80）。

(b) 推定作法の分岐（前提・見立てのタイプ）推定作法の差を定量化するため、本研究では二段階の手順を採った。まず、各アノテーションに付随する記述（推定式・因数・メモ）を対象に、

Qwen/Qwen3-30B-A3B-Thinking-2507 を用いて、専門家・非専門家・LLM を分ける特徴を探索的に抽出させた。その結果、差が現れやすい軸として 視点・単価定義・時間軸・TAM 定義の 4 軸が得られた。次に、この 4 軸に基づき、Qwen3 モデルで再分類し、各グループにおける割合を算出した（図 5）。本研究で用いる 4 軸は、二項（A/B）に加え、判断が難しい場合を「混合/不明」とする。視点：推定の中心が 供給側（自社の営業・提供キャパ、獲得可能数など）か、需要側（市場母数、普及率、TAM/SAM など）か。単価定義：単価が 価値/実費（ROI、削減額、原価等）に基づくか、相場/予算（価格帯、Budget、TCV/ACV 等）に基づくか。時間軸：立ち上がりを遅行（導入リードタイム、稟議、摩擦等）として捉えるか、先行（需要発生・イベント/トレンド起点）として捉えるか。TAM 定義：TAM を獲得可能な上限（目標最大）として置くか、理論上の総枠（市場最大）として置くか。

結果として専門家と非専門家、ならびに LLM の間に体系的な偏りが見られる。たとえば、専門家は需要側の視点に強く寄る一方、非専門家や LLM は供給側の比率が相対的に高い。また、時間軸について専門家は遅行（実装・浸透の制約を織り込む）に寄るのに対し、非専門家は混合／不明が多数となりやすい。さらに TAM の定義は、専門家が獲得目標寄りであるのに対し、LLM（特に GPT-5.1）が理論総枠寄りに偏る。これらは、同じ市場規模推定でも「何を起点にし、どの時間解釈で、定義をどう置くか」が揃っていないことを示す。

(c) 因数分解の中身の分岐（因子カテゴリ）図 4（因子カテゴリ構成比）では、数量（人数・社数など）、割合・率、単価・金額、その他の比率がアノテータごとに大きく異なる。つまり、人間同士で推定値が近くなる場合があっても、同じ種類の因子を同じ粒度で積み上げているとは限らない。推定ステップの妥当性評価が伸びにくい（RQ2）背景として、こうした推定作法・因子設計の分岐が関与している可能性が高い。

以上より、専門家／非専門家の分岐は、最終数値の違いだけでなく、推定の前提（視点・時間軸・TAM 定義など）と、因数分解に含める因子の構成に現れる。そして LLM の推定値が「妥当」と見なされる局面があっても、推定作法まで含めた整合は取りにくく、評価像（何を良しとするか）も専門性によって変化することがわかる。

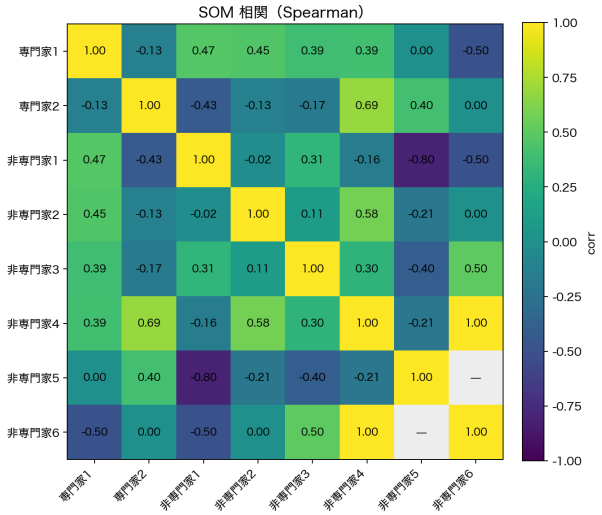

図 3　アノテータ毎の SOM の Spearman 相関係数

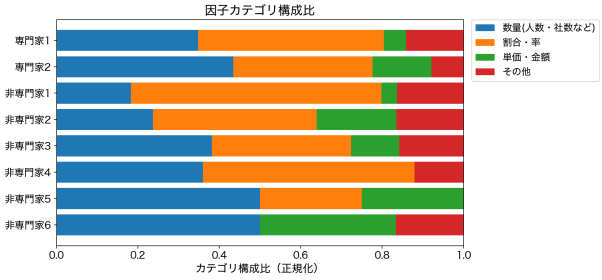

図 4　アノテータ毎の SOM のフェルミ推定時の因数分解における因数のカテゴリ分類

5　おわりに

本研究では、新規ビジネスアイデアの人手フェルミ推定と LLM のフェルミ推定評価を収集し、アノテータを専門家と非専門家の 2 属性に分けつつ分析を行った。そして、新規ビジネスアイデアのフェルミ推定では、推定値と推定ステップが別物として振る舞い、さらに人間側の推定作法も一様ではないということが明らかにした。今後の自動化・モデル化では、定義（TAM/SAM/SOM の置き方、時間軸（3 年以内の SOM）、観点（需要側／供給側）、因子カテゴリ（何を掛け合わせるか）といった推定ステップの構造を、学習・評価の中心に据える必要がある。

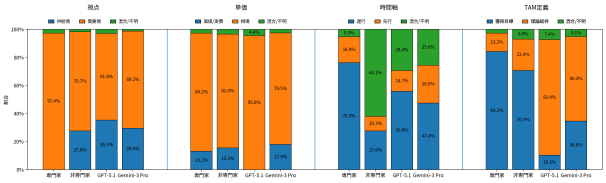

図 5　推定の視点、単価の定義、売上につながるまでの時間軸の捉え方、TAM の捉え方における専門家、非専門家、GPT-5.1, Gemini-3 Pro の違い

参考文献

[1] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025.

[2] Wataru Hirota, Chung-Chi Chen, Tomoko Ohkuma, Tomoki Taniguchi, and Tatsuya Ishigaki. Overview of PBIG shared task at AgentScen 2025: Product business idea generation from patents. In Chung-Chi Chen, Tatsuya Ishigaki, Sophia Ananiadou, and Hiroya Takamura, editors, **Proceedings of the 2nd Workshop on Agent AI for Scenario Planning**, pp. 35–42, Montreal, Canada, 16 August 2025. -.

[3] Steven N. Kaplan and Per Strömberg. Characteristics, contracts, and actions: Evidence from venture capitalist analyses. **The Journal of Finance**, Vol. 59, No. 5, pp. 2177–2210, 2004.

[4] Robert G. Cooper, Scott J. Edgett, and Elko J. Kleinschmidt. Portfolio management in new product development: Lessons from leading firms. Working Paper 60, McMaster University, Michael G. DeGroote School of Business, Innovation Research Centre, Hamilton, Ontario, Canada, February 1997. Innovation Research Working Group. Part I.

[5] Philip M. Anderson and Cherie Ann Sherman. Applying the fermi estimation technique to business problems. **Journal of Applied Business and Economics**, Vol. 10, No. 5, pp. 33–42, 2010.

[6] 菊野慎太郎, 新一郎. フェルミ推定を取り入れた「標本調査」単元の開発と実践. 静岡大学教育実践総合センター紀要, Vol. 31, pp. 147–158, March 2021.

[7] Vildan Salikutluk and Frank Jäkel. Deliberation in guesstimation. **Cognitive Science**, Vol. 49, No. 8, p. e70090, August 2025. Published online 2025-08-13. PMID: 40802877. PMCID: PMC12349749.