LLM 偽針混入テスト：誤抽出を考慮した情報抽出時の評価フレームワーク

叶内 晨 [1,2]　深澤 祐援 [1,2]　角野 為耶 [1,3]　林 翔太 [1]　小原 正大 [1]
[1] 株式会社ブリングアウト　[2]NLPeanuts Inc.　[3] 株式会社 Axcreator
{shin.kanouchi, yusuke.fukasawa}@nlpeanuts.com
{nasuka.sumino, shota.hayashi, shota.kohara}@bringout.cloud

概要

本研究では，大規模言語モデル（LLM）の情報抽出時の評価手法である Needle in a Haystack LLM テストをより実践的に拡張した LLM 偽針混入テストを提案する．従来手法は LLM の抽出精度を Recall のみで評価するため，現実の情報抽出で頻出する誤抽出の問題を考慮していない．提案手法では，Precision を評価指標に加えて誤抽出を定量化するとともに，抽出対象と文脈のドメインを一致させ，さらに偽正解情報を混入することで実践により近い誤抽出を評価可能とする．実験の結果，提案手法による現実的な設定下で LLM による誤抽出が増加し，誤抽出を網羅的に評価する必要性が示された．

1　はじめに

大規模言語モデル（LLM）[1] や検索拡張生成システム（RAG）[2] は，膨大な情報から必要な知見を効率的に抽出する用途で，ビジネスや研究など多様な領域で活躍している．近年では，長大な文脈（Long Context）を扱う LLM が提案されており [3, 4]，それを用いて商談記録や製品資料などの長文を解析するサービス[1)] の開発が進んでいる．しかし，既存の LLM は文脈が長大になるほどハルシネーションが発生したり出力品質が低下する問題がある [5, 6, 7]．そのため実用化にあたっては，LLM の出力を厳密に評価する必要がある．

LLM による情報抽出性能を測定する手法の 1 つに，Needle in a Haystack LLM テスト（以下，針テスト）がある [8, 9]．針テストでは，膨大な文脈（干し草の山）の中に少量の特定情報（針）を人為的に埋め込み，モデルがその針を正しく抽出できるかを評価する．針の数や文脈長などのパラメータを変更することで網羅的な評価が可能であり，モデルの情報抽出性能を測る指標として有用である [10, 11]．

しかし，既存のテストには以下 3 つの課題がある．

1. 埋め込んだ針をモデルが漏れなく抽出可能かを Recall で評価するため，誤抽出は検知しない．
2. 埋め込む針と抽出クエリのドメインが文脈と異なるため，抽出が容易で現実的な問題設定となっていない．
3. 針に類似した誤抽出が起きやすい情報が文脈中に存在しないため，実タスクよりも誤抽出が起きにくい．

そこで本研究では，針テストを拡張した新たな評価フレームワークである LLM 偽針混入テストを提案する．具体的には以下の改良により，実運用環境により近い設定で LLM の情報抽出性能を評価する．

1. Precision を評価指標に加えて誤抽出を評価．
2. 埋め込む針と抽出クエリを文脈と同じドメイン情報に変更．
3. 偽正解情報 (偽針) を文脈に混入．

実験では営業商談の文字起こしデータを対象とし，文脈と関連した針と偽針が情報抽出精度にどう影響するかを分析する．実験の結果，複数の商用 LLM において偽針が抽出精度に大きく影響することから，提案する評価手法の実用性が示された．

2　タスク概要

図 1 に LLM 偽針混入テストの全体図を示す．2.1 節で従来の針テストについて説明し，2.2 節で本研究で提案する偽針混入テストについて説明する．

2.1　Needle In a Haystack LLM テスト

Needle in a Haystack LLM テスト [8, 9]（以下，針テスト）は，長大な文脈に特定の情報（針）を挿入し，

1） https://www.bringout.biz

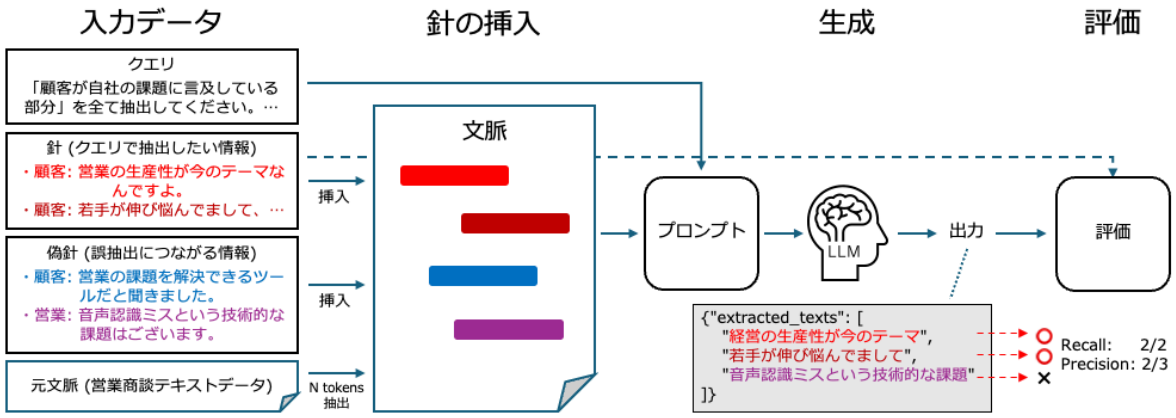

図 1　LLM 偽針混入テストの全体図

LLM がその針を適切に抽出できるかを測定するフレームワークである．文脈長や針の数・位置，LLM の種類などのパラメータを設定できるため，様々な状況を考慮した網羅的な検証が可能であり，モデルがどこまで針を見落とさず正しく情報抽出できるかを評価できる．針テストは以下の手順で行う．

1. **文脈 (Long Context) の用意**: 利用する元データからパラメータで設定した先頭 N tokens を抽出し，針を埋め込むための文脈とする．
2. **針の挿入**: 文脈に対して，事前に用意した針となる情報を挿入する．針の挿入数や挿入位置はパラメータで設定する．従来手法は文脈と無関係な針が使われ，例えば「イチジクは完璧なピザを作るための材料のひとつ」などがある．
3. **抽出プロンプトの作成**: 針に対応するクエリを用いて，「{ クエリ } に関する情報を全て抽出してください」などのプロンプトを作成する．上記の針に対応するクエリとして「完璧なピザを作るために必要な秘密の食材」が挙げられる．
4. **LLM による生成 (情報抽出)**: 埋め込んだ針を抽出するため，GPT-4[3] や Gemini[4] などの任意の LLM に対して針入りの文脈と抽出プロンプトを入力し，抽出結果を得る．
5. **評価**: モデルの出力に正解情報 (挿入した針) が含まれるかを，別の LLM を用いて Recall で評価する．入力の針をそのまま正解に用いる場合と，正解文を別途用意する場合の両方がある．

2.2　提案手法: LLM 偽針混入テスト

本研究では，針テストを拡張した LLM 偽針混入テストを提案する．以下に挙げる改善により，現実的な誤抽出を定量的に捉える仕組みを構築する．

評価指標の拡張　従来手法は正解の抽出対象が複数でも出力を単一の文字列としており，情報の抽出個数が自明ではない．そのため，評価する際は別の LLM を用いて正解の針を基準として何個抽出できたかという Recall ベースになっており，誤抽出を考慮していない．そこで提案手法では，LLM を用いた抽出結果を JSON の配列として得ることでモデルが抽出した情報の数を明確にし（付録 A），Precision，F1 スコアを含めた総合的な評価を行う．また，抽出時には情報をそのまま抽出することを指示し，正解の針と抽出結果の文字列一致で評価する．ただし，抽出結果が完全な文ではない場合も考慮し，抽出した文字列が正解の針の一部であれば正解とする．

文脈と関連した針とクエリの使用　従来手法は文脈と無関係な針を埋め込むため，実際のタスクに比べて情報抽出が容易であった．そこで本研究では，より現実的な設定として，文脈と関連した針とクエリを使用する．例えば，営業商談を文脈に用いるケースでは「顧客の課題」のような営業現場で頻出な情報を針とクエリとする．

一方で，同じドメインに属する針を利用すると，既存の文脈に同種の情報が存在する可能性があり，それらも針と同様に抽出され評価が煩雑になる恐れがある．この問題を回避するため，本研究では針挿入前に最大 3 回まで抽出プロンプトを用いて文脈から情報を抽出し，抽出された表現を含む文を事前に文脈から除去する．これにより，元から文脈に存在する生の正解情報が評価を妨げるリスクを可能な限り抑え，モデルの情報抽出性能をより正確に評価できるようにしている．

偽針の混入　本研究では，正解の針と見間違えやすい情報（偽針）を文脈に挿入することで，LLM が誤って偽針を抽出しないかを検証する．例えばクエ

表 1　クエリと針・偽針の例

カテゴリ	クエリ	針 (クエリで抽出したい情報)	偽針 (誤抽出につながる情報)
顧客の課題	顧客が自社の課題に言及している部分	[針 1] 顧客: 若手の営業が伸び悩んでまして，早急に対応する必要がある... [針 2] 顧客: ... 商談準備と議事録にすごい時間かかっちゃっているんですよね．[針 3] 顧客: 営業の生産性も今のテーマなんですよ，ここも改善しろと言われていて．	[偽針 1] 営業: 一部音声認識ミスが起きてしまうという技術的な課題はございますが，... [偽針 2] 営業: 営業が伸び悩んでいるようなお客さんに使っていただいて... というお声をいただいています．[偽針 3] 顧客: 営業の課題を解決できるツールだと聞いたのですが，本当ですか？．営業: 例えば別のお客様からですと....
顧客の予算	顧客が自社の予算に言及している部分	[針 1] 顧客: 今期は予算カツカツで結構厳しいんですよね．[針 2] 顧客: 余裕はあるんですが，500 万円からは経営会議にかけなきゃだめなんです．[針 3] 顧客: 一括は厳しいですが，毎月 100 万円であれば問題ないと思います．	[偽針 1] 営業: 1000 万円以上の年間一括契約の場合，1 年間の無料コンサルティングも付きますがいかがでしょうか．[偽針 2] 顧客: 他のお客さんで，予算の都合で半分を来年度に払うというパターンもあります．[偽針 3] 顧客: 顧客: メンバーによってピンキリですが，トップ営業だと月に 2000 万くらい売り上げるんですよ．
ピザ (無関係)	完璧なピザを作るために必要な秘密の食材	[針 1] イチジクは完璧なピザを作るための材料のひとつ．[針 2] 生ハムは完璧なピザを作るための材料のひとつ．[針 3] 山羊のチーズは完璧なピザを....	[偽針 1] アボカドは完璧なハンバーガーを作るための食材のひとつ．[偽針 2] オリーブオイルは完璧なパスタを作るための食材のひとつ．[偽針 3] サーモンは完璧な海鮮丼を....
シスコ (無関係)	サンフランシスコでの一番の楽しみ	[針 1] サンフランシスコでの一番の楽しみは，晴れた日にドローレス公園に行くことです．[針 2]... 楽しみは，ゴールデンゲートブリッジを自転車で渡ることです．[針 3]... 楽しみは，アルカトラズ島に行くことです．	[偽針 1] ボストンでの 1 番の楽しみは，美術館に行くことです．[偽針 2] サンフランシスコでの一番の辛い経験は，ツイン・ピークスのハイキングです．[偽針 3] サンフランシスコの友達は，いつもサンドイッチを食べています．

表 2　実験時のパラメータ一覧

パラメータ名	値
対象商談	商談 1 〜 5
文脈長 (tokens)	4k，8k，12k
針カテゴリ	課題，予算，ピザ，シスコ
針の数	1，2，3
偽針の数	0，1，2，3
針埋込開始位置	10%，40%，70%
針埋込終了位置	100%
モデル	GPT-4o, GPT-4o-mini, Gemini-1.5-Pro, Gemini-1.5-Flash

表 3　GPT-4o による針の種類毎の抽出精度

針タイプ	偽針	P	R	F1
文脈に無関係	偽針なし	1.000	0.991	0.995
	偽針あり	0.930	0.995	0.961
文脈に関係	偽針なし	0.915	0.993	0.952
	偽針あり	**0.735**	**0.960**	**0.832**

リが「顧客の課題」の場合，課題という文字列を含むが顧客の課題ではない文や，顧客ではなく営業側の課題が記述された文などの紛らわしい情報を意図的に挿入することで，正解の針との見分けがつきにくい状況を作り出す．これにより，LLM が正しい情報だけを的確に抽出できるか，あるいは偽針に惑わされて誤抽出してしまうかが評価可能となる．実運用では多くの情報が混在した中から的確に目的の情報を抽出する必要がある．偽針を用いた評価で LLM の誤抽出リスクを把握することで，実タスクでの安全性や信頼性を確認することが可能になる．

3　実験

3.1　実験設定

2.2 節で提案した LLM 偽針混入テストを，営業ドメインのデモ商談記録の文字起こしデータ（付録 B）に対して行った．データの各行の先頭には話者分離結果の「営業:」「顧客:」が埋め込まれている．元動画は 5 商談用意し，それぞれ動画長は 30 分前後で，文字起こし結果は平均 14k 文字となった．

表 1 に実験に利用したクエリと針を示す．文脈に関連した針として，営業ドメインにおける代表的な概念である「顧客の課題」カテゴリなどを用意し，文脈に無関係な針として従来手法で使われる「ピザ」カテゴリなどを用意した．また，モデルの誤検出リスクを評価するための偽針として，例えば「顧客の予算」では価格情報だが顧客の予算ではないものなどを用意した．

実験に用いたパラメータを表 2 に示す．針を埋め込む際は，パラメータの針埋込開始位置から針埋込終了位置まで等間隔に挿入した．また，針と偽針の順番は乱数で決定した．評価には，全てのパラメータに基づく網羅的な抽出結果を用い，分析時には同一母集団を分析項目に応じてセグメント分けし，それぞれマイクロ平均値を算出した．

3.2　実験結果

表 3 に，文脈に関係・無関係な針を埋め込んだ場合と，偽針を埋め込んだ場合の GPT-4o による抽出精度を示す．文脈に無関係な針のみを埋め込んだ場合，GPT-4o の F1 スコアは 0.995 となり，ほぼ正確に針を抽出した．追加で偽針を埋め込んだ場合は F1 スコアが 0.961 まで低下した．これは，偽針を誤抽出したことが原因で，実際に Recall はほぼ変わらず Precision が 0.930 に低下している．

一方，文脈に関係する針を埋め込んだ場合，Precision が 0.915，F1 スコアが 0.952 まで低下した．これはクエリと針が文脈と関連する内容になり，モデルが正解の針以外の余計な情報を抽出するようになったことが原因であった．文脈に関係する針と偽針の両方が埋め込まれた最も困難な条件では，F1 スコアは 0.832 まで低下した．Precision は 0.735 で

表 4　LLM ごとの抽出精度

モデル名	P	R	F1
GPT-4o (2024-08-06)	**0.852**	**0.981**	**0.912**
GPT-4o-mini (2024-07-18)	0.646	0.927	0.761
Gemini-1.5-Pro	0.729	0.961	0.829
Gemini-1.5-Flash	0.516	0.961	0.671

表 5　針カテゴリごとの抽出精度 (GPT-4o)

針のカテゴリ	P	R	F1
ピザ (文脈無関係)	0.998	0.993	0.995
シスコ (文脈無関係)	0.900	0.995	0.945
顧客の課題	0.742	0.954	0.834
顧客の予算	0.807	0.982	0.886

あり，文脈に関係する針と偽針が共存すると，誤抽出がさらに増加する結果となった．さらに Recall も 0.960 まで低下しており，タスクが困難になることで，モデルがこれまで抽出に成功していた正解の針を抽出し損ねることを確認した．実際のタスクではこういった複雑な状況が多く，この条件下での精度検証や改善が求められる．また，今回の検証では正解の針として明確な事例を用いたが，実世界では文脈外知識や推論が必要な場合も多く，抽出精度はさらに低下することが予測される．

LLM ごとの抽出精度　表 4 に LLM ごとの抽出精度を示す．GPT-4o は全体的に抽出精度が高く，F1 スコアは全体で最高の 0.912 であった．Gemini-1.5-pro の F1 スコアは 0.829 で，文脈に関係したクエリにおいて文脈情報を過剰に抽出する傾向があった．また，GPT-4o-mini と Gemini-1.5-Flash は偽針の誤抽出が大幅に増え，Precision が大きく低下した．

本研究では従来手法同様に全モデルで可能な限り共通のプロンプトを利用し，モデルごとの最適化はしていない．そのため，モデルに合わせて抽出基準を調整することで，抽出精度のさらなる向上が期待できる．以降の実験では，精度が最も高かった GPT-4o のみを対象として分析を進める．

針のカテゴリ毎の抽出精度　表 5 に針カテゴリごとの精度を示す．全カテゴリで高い Recall を示し，必要な情報を抽出することには成功している．一方で，Precision は針ごとに大きな差があり，特に「顧客の課題」カテゴリが最も低く 0.742 となった．この原因として，課題という概念の抽象度であるため，課題とは言い切れない情報も抽出されるケースが散見された．具体的には「成長には一番大切な部分なので，そこがどんなふうに活用できるかっていうのが気になりますかね.」のような事例が誤抽出されるケースを確認した．また，先行研究で用いられた「ピザ」カテゴリはほぼ全て抽出に成功しており，タスク難易度が低いことを確認した．

表 6　文脈長・偽針の数と Precision (GPT-4o)

		偽針数			
		0 本	1 本	2 本	3 本
文脈長	4k	0.984	0.976	0.838	0.800
	8k	0.981	0.937	0.784	0.747
	12k	0.905	0.887	0.766	0.732

表 7　針・偽針の数と Precision (GPT-4o)

		偽針数			
		0 本	1 本	2 本	3 本
針数	1 本	0.829	0.818	0.669	**0.598**
	2 本	0.981	0.952	0.767	0.754
	3 本	0.987	0.965	0.874	0.841

文脈長と抽出精度　表 6 に文脈長ごとの抽出精度を示す．文脈長が 4k でかつ偽針なし場合，Precision は 0.984 でほとんど誤抽出していない．しかし，文脈長が短い設定でも偽針が混入すると Precision は低下し，偽針 3 本の場合に 0.800 となった．これは文脈長が 3 倍である 12k で偽針なしの 0.905 に比べて 10 ポイント低く，偽針の存在が他の条件に比べて Precision へ強く影響することを示している．

針・偽針の数と抽出精度　表 7 に針・偽針数ごとの抽出精度を示す．偽針の数が多くなるほど Precision が低下し，誤抽出が増えることを確認した．特に，正解の針が 1 本で偽針が 3 本の時が最も低く，Precision は 0.598 となった．今回の実験では正解の針の数を複数設定したため，抽出時の指示は全て「XX を全てそのまま抽出してください」といった文言を使用した．そのため，正解が 1 本しかない場合もモデルは常に複数針を探す可能性があり，偽針を誤抽出する結果となった．実タスクにおいても正解数が事前にわからない場合は多く，解決すべき重要な課題である．

4　おわりに

本研究では，Needle in a Haystack LLM テストを拡張し，実運用環境により近い LLM 偽針混入テストを提案した．Recall のみでなく Precision を評価指標に加え，文脈に関連する針の使用や偽関係情報（偽針）の混入により，より現実に近い状況下での評価が可能となった．本手法は営業ドメインのみならず，他の実世界データセットへも適用可能であり，LLM による情報抽出タスクの信頼性と精度評価に新たな基準を提供しうる．今後は，他ドメインへの適用や評価結果を元に改善を回す方法を検討する．

参考文献

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. **arXiv:2303.18223**, 2023.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997**, 2023.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. **arXiv:2303.08774**, 2023.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. **arXiv:2312.11805**, 2023.

[5] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 157–173, 2024.

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, 2023.

[7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, Vol. 55, No. 12, pp. 1–38, 2023.

[8] Greg Kamradt. Needle in a Haystack - Pressure Testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

[9] LangChain. Multi Needle in a Haystack. https://blog.langchain.dev/multi-needle-in-a-haystack, 2024.

[10] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv:2403.05530**, 2024.

[11] Daniel Machlab and Rick Battle. LLM In-Context Recall is Prompt Dependent. **arXiv:2404.08865**, 2024.

A　抽出プロンプト

LLM を用いて情報抽出をする際のプロンプトを以下に示す．先行研究のプロンプトを元にしつつ，JSON の配列を出力するように修正した．また GPT4 と Gemini の設定により，JSON 形式での出力を保証した．

You are a helpful AI bot that answers questions for a user.
———-
{ 文脈 }
———-
Here is the user question:
———-
「{ クエリ }」を JSON の配列として全てそのまま抽出してください．
一つもない場合は JSON の extracted_texts を空で返してください．
Don't give information outside the document or repeat your findings.
Return in the following json format:
```
{"extracted_texts": ["text1", "text2"]}
```

B　文脈となる商談データ

文脈となるデータは，営業担当が特定のサービスを顧客に提案・販売する場面を想定したデモ商談動画で，ブリングアウトの音声認識エンジン[2)]を用いて文字起こしをしている．データの一部を以下に示す．

営業： こちらが弊社のサービスの全体概要ではございましたけども，いかがでしたでしょうか．
営業： ちょっと西村様と思い描いてるような課題感，今お話させていただいたところに対して，少しフィットしてるのか，ちょっとずれてるのかとか，その辺のご意見をぜひ忌憚のないちょっとご意見いただければと思っておりましてそうですね．
顧客： はいまず人材のなんか育成っていう観点で言うと，使える部分ありそうだなっていうふうに思ったのと営業のその業務効率化っていう点でも使えそうかなと．
営業： なるほどですね．
顧客： 今我々の一番の課題っていうものはもちろんその人材育成っていうところもあるんです．業務効率化もあるんですけれど．
顧客： なんかこのサービスでわかると，なんかより良いかなっていうのはちょっと思いましたけどね．
営業： 本当はそのどこの市場に刺さるかとかっていうところが今．
顧客： 一応少し前までは契約書って，契約書のデータベース作るサービスなんですけど，業界によって契約書位置づけというのがだいぶ変わるんですね．
顧客： もう注文書でやり取りしてるとこもあれば，一つ一つ契約書を取り交わしてるところもあって，あの見返す見返さないみたいなところも結構差がありまして，そういったためにその業界を絞ってアプローチしていくっていう戦略を前期から始めてまして．
営業： 前期から．
顧客： そうなんすよ．エンターテイメントとかコンテンツ業界，資産を中心にやってきて，今期は 6 月から新しいキーなんすけど，だからもう少しそこと類似したようなパターンがある業界に広げてアプローチしていけないかっていうことで建設業だったり，あとはシステム会社だったりとかそういう業界にちょっとターゲットを広げて，今ちょっと進めてるとこなんですよ．
営業： そういうことでございますね．
顧客： なのでそういうどこの業界だったらどういうふうに話すじゃなくて，コントラクト共通で伝えないといけないことはこれだねとか聞かないといけないことはこれだねみたいなことがわかって，それがさらにどの業界にアプローチしていけばいいかみたいなことまでわかると，なんかいいより良いかなっていうふうには思いました．
営業： ありがとうございます．なるほどですね．....

2） https://www.bringout.biz