Information Signals in Sponsored Search: Evidence from Google's BERT *

Poet Larsen
University of Southern California
poet.larsen@marshall.usc.edu

Davide Proserpio
University of Southern California
proserpi@marshall.usc.edu

September 27, 2024

*We thank Odilon Câmara, Anthony Dukes, Soohyun Kim, Nikhil Malik, Dina Mayzlin, Sean Melessa, Dinesh Puranam, Srinivas Tunuguntula, and participants of the 2022 Conference on Artificial Intelligence, Machine Learning, and Business Analytics, 2023 Stanford Quantitative Marketing Conference, and 2023 Marketing Science Conference for their helpful comments and suggestions. We thank the Institute for Outlier Research in Business and the Marketing Science Institute for generous financial support.

Abstract

We study how improvements to search engine interpretation algorithms and the information signals they generate affect sponsored search markets. We do so by analyzing changes in the number of bidders and cost-per-click (CPC) per query auction following Google's October 2019 rollout of Bidirectional Encoder Representations from Transformers (BERT). In aggregate, BERT increases the number of bidders but minimally changes CPC. Given that queries of different lengths have different amounts of information, we analyze how query length moderates the effect of BERT. Despite an increase in the number of bidders across all queries, CPC increases for shorter queries but declines as query length increases. We then develop a theoretical auction model to understand the mechanism driving these findings and make predictions about future algorithm updates. Our results offer insight into the economic impact of AI and Large Language Models (LLMs) on advertising markets and help advertisers prepare for future algorithm updates.

1 Introduction

Sponsored search continues to be a dominant advertising channel for firms to reach consumers. It also remains the primary source of revenue for many search engines. In 2023 alone, US search advertising revenue reached nearly $90 billion.[1] When a consumer submits a search query, a search engine faces a fundamental problem of interpreting the query in order to generate a response. Search engines such as Google and Bing increasingly rely on Natural Language Processing (NLP) algorithms to interpret consumer search queries and generate search intent signals. These signals help identify and rank relevant advertisers for sponsored search auction opportunities. At times, search engines update their interpretation algorithms, which can impact the quality of the search intent signal received by the search engine and subsequently influence the sponsored search auction market.

Changes to interpretation algorithms are becoming more prevalent due to recent advancements in NLP research. In June 2017, researchers at Google Brain introduced the transformer model architecture (Vaswani et al., 2017). Notable for its ability to dynamically accommodate contextual information through its "self-attention" mechanism, transformers significantly improved the ability to both interpret (encode) and generate (decode) text. This technology has since become the backbone of modern-day Large Language Models (LLMs) and has proclaimed a new era of programmatic interpretation capabilities.

In October 2018, Google researchers used the transformer architecture to build Bidirectional Encoder Representations from Transformers (BERT), one of the first LLMs. BERT was notable for its dramatic improvement at predictive tasks over previous state-of-the-art models (Devlin et al., 2018). In October 2019, Google replaced RankBrain with BERT as its main interpretation algorithm to parse and understand user-generated search queries.[2] We

[1] See: `https://www.iab.com/news/2023-u-s-digital-advertising-industry-hits-new-record-according-to-iabs-annual-internet-advertising-revenue-report/`

[2] See `https://blog.google/products/search/search-language-understanding-bert/` for the official BERT announcement. See `https://searchengineland.com/welcome-bert-google-artificial-intelligence-for-understanding-search-queries-323976` for industry announcement. See `https://twitter.com/searchliaison/status/1204152378292867074` for the announcement of international BERT stating that US English BERT was introduced in October 2019.

use this event to empirically study how new query interpretation algorithms and the new interpretation signals they generate impact downstream prices and the number of bidders in sponsored search auctions.

Advertisers are ranked based on ad ranks in sponsored search auctions. Ad ranks depend on submitted bids and relevancy scores. Relevancy scores measure the match quality between an advertiser and a search query and impact both auction eligibility and final paid prices.[3] Higher relevancy scores will boost ad ranks and increase an advertiser's likelihood of being matched to an auction and winning an ad slot. Higher scores also generally reward advertisers with lower final prices. Query interpretation algorithms provide search intent signals to the seller (Google) to help estimate advertiser relevancy scores. How might BERT's information affect relevancy scores and subsequently impact short-term auction cost-per-click (CPC) and the number of bidders present in the market (i.e., competition)?

BERT's new information could lead to more or less bidders per query auction. The computer science literature has documented BERT's ability to better understand sentence semantics (Tenney et al., 2019b; Rogers et al., 2021) and complex syntax (grammatical) structures (Devlin et al., 2018; Lin et al., 2019; Tenney et al., 2019a). On the one hand, this information could help Google better interpret search queries and identify more relevant advertisers, leading to more bidders per auction. On the other hand, the new information may help Google segment markets and remove irrelevant advertisers from auctions, leading to fewer bidders per auction.

How CPC changes is also ambiguous due to potential changes in the number of bidders competing and relevancy scores. A higher (lower) number of bidders should lead to higher (lower) prices, but increased relevancy scores may lead to lower CPC.

To study BERT's effect, we collect monthly average CPC and Competition Score (a measure of the number of bidders) data from SEMRush—a company that tracks search query performance—for a sample of roughly 12,000 search queries over two years (2018-

[3]See `https://support.google.com/google-ads/answer/1722122?hl=en` for discussion of Google ad rank and use of relevancy scores.

2020). We present two identification strategies to estimate the effect of BERT on CPC and Competition Score (CS).

The first exploits the panel nature of our data and employs a difference-in-differences (DD) approach akin to those employed in Eichenbaum et al. (2020), Bollinger et al. (2022), and Liaukonyte et al. (2022). Specifically, we compare changes in CPC and CS before and after the introduction of BERT, with a baseline of changes over the same months but in the year prior. This strategy exploits variation within queries and across time to identify BERT's effect.

The second identification strategy exploits inherent linguistic variation across queries. The idea behind this strategy is that linguistic properties (i.e., syntactic and semantic properties) should not affect our dependent variables (CPC and the number of bidders) except through their interaction with Google's query interpretation algorithm. When the interpretation algorithm changes (i.e., BERT is introduced), how the algorithm interprets these linguistic properties and subsequently affects our dependent variables will change. Under this assumption, we can measure and use these linguistic properties to identify how a change in the interpretation algorithm causes a shift in our dependent variables.

We estimate consistent effects across both strategies, with the second strategy providing more conservative estimates of the treatment effect.

Our study begins by analyzing aggregate changes to CPC and CS. Using our preferred specification (DD), we find that CS (i.e., the number of bidders) increases by roughly 1% after BERT's introduction. However, the aggregate increase in auction participants does not lead to significant changes in CPC. These results hint that prices may be experiencing heterogeneous changes.

The aggregate analysis does not capture the rich linguistic differences across search queries. In its release notes, Google states that it expects BERT to improve long query interpretation.[4] In other words, Google expects interpretation quality to improve as query

[4]Google stated it struggled with long queries in its patch notes post announcing BERT, which can be found here: `https://blog.google/products/search/search-language-understanding-bert/`.

length increases due to BERT's ability to understand complex language structures.

However, BERT also learns new information about query semantic relationships (Tenney et al., 2019b; Rogers et al., 2021). Take the short, broad query "socks". With BERT, Google may know that "socks" relates to more search terms, such as "shoes" or "sandals". For longer queries, such as "best-running shoe stores near me in LA", BERT may learn that these queries are generally more specific and related to fewer other queries. Changes to semantic understanding will occur for all queries and likely affect the pool of advertisers deemed relevant to a given query auction.

To focus on linguistic informational differences across queries, we study how query length moderates BERT's effect. Short queries tend to be semantically broad and maintain simple syntactic structures. We hypothesize that BERT will help Google expand short query auction markets due to increased semantic relatedness, leading to more bidders being allocated to the auction and higher CPC. Longer queries are generally more complex and specific (Anderson, 2006). BERT's ability to interpret complex syntactic information will significantly improve interpretation quality and decrease the semantic relatedness of longer queries due to their inherent specificity. We hypothesize that this will lead to fewer bidders and CPC, but more clicks to the Click-Through Rate (CTR) should increase.

Consistent with our hypotheses, we find that both CS and CPC increase for short queries (1.3% and 3.8%, respectively). However, as query length increases, we find that contrary to our expectations, CS *increases* while CPC *decreases*.

These findings create an interesting empirical puzzle. The uniform increase in bidders across query lengths suggests that BERT's new information helps Google identify more relevant advertisers for each query auction. In a standard auction theory model, this increase in the number of bidders should increase prices. Yet, CPC declines as query length increases.

We, therefore, develop a theoretical auction model that provides a potential explanation for our findings and offers a framework to extrapolate our empirical findings to future algorithm updates. The model has one seller and several buyers of different types in the market.

The seller uses an algorithm to generate signals about a query's type. The seller then uses the signals to estimate relevancy scores and pick auction participants. Chosen buyers are then ranked based on ad ranks, which depend on their estimated relevancy score and submitted bid. (Relevancy scores calculate the match quality between the advertiser and the query). The winning advertiser shows an ad, and a click is received depending on the match quality between the winning advertiser and the query.

The model has two key components. First, motivated by literature in computer science, psychology, neuroscience, and linguistics, we posit that language maintains a multi-dimensional type structure (Mnih and Hinton, 2008; Jäger and Rogers, 2012; Miyagawa et al., 2013; Coopmans et al., 2023). These dimensions define a query and advertiser's type and affect the query CTR probability. Second, previous empirical work on information disclosure in auction markets has focused on scenarios where buyers can use new information to self-select into preferred markets (Tadelis and Zettelmeyer, 2015; Cowgill and Dorobantu, 2020). In our setting, the seller (platform) uses the information signals to select its buyers in the same way Google matches advertisers to auctions. The model analysis primarily focuses on understanding how modifications to the seller's algorithm impact query prices.

To capture the type structure of a query, we define two dimensions of language in our model: topic and context. A query's topic relates to the categorization of the query (e.g., is the query about "shoes" or "insurance"), while context differentiates the query type within the topic (e.g., purchase intent or information acquisition intent). Consider the queries "best-running shoes for women", "stores near me to buy running shoes", and "where to purchase life insurance". The first two queries relate to the topic "shoes", while the last relates to "insurance". However, each query has types of contextual information that convey different search intents and goals. The last two queries express purchase intent, while the first one expresses information acquisition intent. The topic construct captures differentiation across queries, while context captures information that differentiates a query within topics. The seller's algorithm generates signals about each of these dimensions.

We observe several changes as the quality of the seller's algorithm improves. First, a better algorithm helps the seller understand a query's topic, which leads to better market organization and more bidders eligible for advertising space, increasing the average query price and the average number of bidders. Second, a better algorithm can identify the context dimension of a query. Learning context helps the platform estimate more precise relevancy scores and prioritize highly relevant advertisers in the auction. When a new algorithm generates a significantly better context signal, and advertiser-query context alignment matters to a query's CTR, average CPC may decline despite having more bidders.

Mapping to our empirical observations, our theory model suggests the following: 1) BERT's improved topical understanding leads to the increase in the average number of bidders and supply of ad auctions in our data, and 2) the simultaneous price shifts are caused by BERT's contextual interpretation gains and the variation in the importance of query-advertiser context alignment along query length (short vs. long). While we don't empirically observe CTR, we predict that CTR increases predominantly for longer queries.

Combined, our empirical findings and theoretical model help advertisers understand how improvements to query interpretation algorithms affect sponsored search markets. For advertisers, the linguistic structure of short, simple queries inherently limits the ability to differentiate bidders within the auctions, meaning future algorithms will cause these queries to become increasingly more competitive and costly. At the same time, future algorithms will reward advertisers with more precise relevancy score estimates in the long-query markets, leading to more relevant ads and greater bidder differentiation within auctions. Whether this translates to lower prices in the future depends on the new algorithm's relative contextual vs. topical signal improvements.

For marketing academics, our empirical findings contribute to the sponsored search literature and the growing literature on the economic impact of AI and LLMs. Additionally, our theoretical model offers testable predictions for future algorithm updates and improves our theoretical understanding of LLMs' benefits.

2 Related Work

Our paper relates to the growing literature on the economics of AI and LLMs, information disclosure in auction markets, and targeted advertising.

Sponsored Search Sponsored search continues to be a prevalent advertising channel for firms. It also remains an active area of research in both marketing and economics (Edelman et al., 2007; Ghose and Yang, 2009; Yang and Ghose, 2010; Rutz and Bucklin, 2011; Berman and Katona, 2013; Blake et al., 2015; Edelman and Lai, 2016; Simonov et al., 2018; Cowgill and Dorobantu, 2020). Motivated by the auction structure of sponsored search, Edelman et al. (2007) and Varian (2007) study equilibrium bidding strategies in generalized second price auctions. Empirical work has also analyzed ad effectiveness (Ghose and Yang, 2009; Blake et al., 2015), sponsored and organic complementarities (Yang and Ghose, 2010), keyword spillovers (Rutz and Bucklin, 2011), and advertising competition Simonov et al. (2018). One stream of research focuses on the downstream consequences of search engine platform design decisions, including the impact of a search engine's services on click behavior (Edelman and Lai, 2016), result page features (Gleason et al., 2023), and the interaction between search engine optimization (SEO) and sponsored links Berman and Katona (2013). We contribute to this literature by empirically and theoretically studying how improved search engine interpretation algorithms affect sponsored search markets.

Economic Impact of AI and LLMs Recent advancements in LLM technology have spurred academic interest in understanding the potential economic effect of these models. More recently, the development of ChatGPT has motivated researchers to study how generative LLMs impact areas such as labor markets (Eloundou et al., 2023; Zarifhonarvar, 2023), information markets such as Stack Overflow and Reddit (Burtch et al., 2023), and marketing practices (Kushwaha and Kar, 2021; Reisenbichler et al., 2022; Goli and Singh, 2024). We contribute to this growing literature by studying how LLMs used to generate better signals

about consumer search queries impact sponsored search auction markets.

Information Disclosure in Auction Markets Research on information disclosure in auction markets has primarily focused on settings where sellers can endogenously hide or reveal information to buyers. Ganuza (2004) and Board (2009) find that revealing information to buyers about object features when markets are thick generally leads to increasing profits. However, this may not occur in sparse markets due to what Board (2009) calls the *allocation effect*. The allocation effect occurs when information causes the rank ordering of bidder types to swap, leading to weakly decreasing prices.

Somewhat related to our setting, Cowgill and Dorobantu (2020) empirically studies how disclosing new information to advertisers in the sponsored search market affects prices, profits, and CTR. The authors find that information disclosure generally leads to thinner markets and lower prices but higher overall profits due to improved CTR. These market adjustments are due to advertisers improving their self-selection into preferred markets, leading to better query-advertiser matching.

Tadelis and Zettelmeyer (2015) also studies how information disclosure affects buyer selection into auction opportunities. In this paper, the authors find that information disclosure in the automobile resale market can help quality-differentiated buyers self-select into preferred auction opportunities, leading to higher market clearance rates and higher profits across all quality types.

When considering seller trade-offs between strategically revealing and hiding information, the theoretical and empirical literature has argued that there is a fundamental trade-off between keeping auctions dense and improving buyer pricing accuracy (Bergemann et al., 2021). Revealing information may help extract value from buyers (Tadelis and Zettelmeyer, 2015), but can also lead to thinner markets (Cowgill and Dorobantu, 2020).

Google faces the same theoretical trade-offs between keeping auctions dense and improving buyer pricing accuracy in our market setting. However, prior literature has focused

on settings where buyers receive information and subsequently self-select into markets to maximize their profits. In our context, the seller uses the information to pick buyers and maximize their profits. The incentive and market structure differences may lead to empirical results that differ from previous work.

Targeted Advertising Our paper relates to a stream of literature on matching and targeting technology improvements in advertising markets. Amaldoss et al. (2016) studies how improvements to sponsored search broad match technology affect market entry and seller profits. The authors find that better broad match *bidding* algorithms can lower market entry costs and induce greater auction participation, leading to prices.

Empirical and theoretical work has also documented the benefits and market effects of better-targeted advertising technology. Chandra (2009) empirically studies the newspaper market and finds that better targeting can lead to higher prices due to improved advertiser-audience alignment. Athey and Gans (2010) theoretically studies how targeting technology can lead to an increase in the supply of advertising opportunities, potentially putting downward pressure on prices. We contribute to this literature by studying how more informative search query intent signals impact matching in sponsored search markets.

3 Emprical Context and Data

3.1 Empirical Context

Sponsored Search Advertising When a consumer types a query into a search engine (e.g., "how to cook chicken"), a search engine must interpret the query and decide what domains (URLs) to present on the Search Engine Results Page (SERP). At the top of the SERP, search engines may offer sponsored search links. These links are bid for in a real-time auction before the SERP loads on the consumer's web browser.

An advertiser must create an ad campaign to begin bidding for sponsored search ad space.

A sponsored search campaign consists of seven main components: 1) The ad creatives (i.e., what the sponsored search ads look like), 2) The budget of the campaign, 3) The type of individuals the advertiser wants to target (e.g., target Chicago and Los Angeles residents), 4) The keywords the advertiser wants to target, 5) How much the advertiser is willing to bid for each keyword, 6) how to optimize bidding, and 7) The matching strategy. With these ingredients, an advertiser can begin bidding for sponsored search positions.

Sponsored search auctions follow a pay-per-click (PPC) model. Under PPC, advertisers pay only when a consumer clicks on an ad. Because search engines allow multiple sponsored search links to appear at the top of result pages, most rely on a Generalized Second-Price Auction (GSPA) to sell ad space and rank buyers (Edelman et al., 2007). Under a GSPA, advertisers are ranked by bids, and the top N winners for N ad slots show advertisements. If an advertisement receives a click, the advertiser pays the price needed to out-price the next highest bidder (usually by \$0.01).

In practice, advertisers are ranked with Ad Rank scores. Ad Rank scores depend on many factors, including the advertiser's bid and their ad relevancy score. The relevancy score measures how related an advertiser is to a particular query. Higher scores mean the advertiser is more relevant and predicted to receive a click. Ad relevancy matters because it impacts auction eligibility and CPC, with higher relevancy scores increasing the likelihood of auction participation and lowering final paid prices. [5] We assume BERT's new information will impact relevancy score estimation.

The organic rank results are below the sponsored links on the SERP. These are links that the search engine deems relevant to the given search query. Unlike sponsored links, advertisers do not purchase organic links. Instead, the search engine uses an internal ranking system to decide positions. A firm can appear in both sponsored and organic rank slots. For example, even if Nike bids (and wins) the top ad slot in the sponsored search position for the query "women's running shoes", it can still appear in the organic rank results.

[5] See `https://support.google.com/google-ads/answer/1722122?hl=en`.

BERT Bidirectional Encoder Representations from Transformers (BERT) is a large-scale neural network-based language model developed by Google in 2018. It is a pre-trained model capable of understanding the context and nuances of natural language text, making it a powerful tool for a wide range of NLP tasks (Devlin et al., 2018). Google introduced BERT in October of 2019.

BERT's novelty comes from its ability to understand the context in language objects (Devlin et al., 2018; Tenney et al., 2019b; Rogers et al., 2021).[6] Effectively capturing contextual information was a significant challenge for previous state-of-the-art NLP algorithms such as Word2Vec (Mikolov et al., 2013) and was considered a breakthrough in NLP research. With contextual knowledge, BERT can better understand language semantics, e.g., when "bank" indicates a financial institution or land alongside a river based on the other words in the sentence. It can also lead to a better understanding of complex syntactic language structures, such as adjectives or adverbs, and non-linear relationships between words. To capture this information, BERT generates a vector for each word based on the other words surrounding the focal word, i.e., its context, using the transformer-based self-attention mechanism (Vaswani et al., 2017). As such, the same word can have a different vector representation depending on its surrounding words.

3.2 Data

Sampling search queries We collect data from SEMRush, a leading provider of sponsored search rank and keyword data. To generate a sample of queries for the analyses presented in this paper, we relied on a survey that we administered to Amazon Mturkers. After selecting 32 different topics (see Appendix A for the list of topics), we asked survey respondents the following question: "Please write a search query related to the topic of '**Topic**' that you would search for on Google". 100 Amazon Mturk participants took the survey and each participant was asked about five randomly selected categories, giving us 500 responses and

[6] In simple terms, this means that BERT can differentiate when the word "mouse" refers to a computer mouse or rodent, depending on the words around "mouse".

roughly 16 responses per topic category.

Not all MTurk submissions were queries. To focus on submissions that were search queries, we removed 210 irrelevant answers. Irrelevant responses included answers where participants typed in specific URLs, website content, descriptions of how to type up a search query, and quotes from Google "Help Pages" describing how to search. Removing these responses left us with 290 search query answers.

Using the remaining 290 sampled queries, we turned to the broad match database at SEMRush.[7] For a given query, SEMRush returns a set of similar queries for which it has data for the current month.[8] We limited our match to queries with an average search volume of at least 500 from 2021-2022, generating a data set of roughly $120,000$ queries related to the original 290 queries. However, many of the queries contained explicit terms (e.g., adult content such as pornographic search queries) unsuitable for analysis (advertisers generally don't target adult content queries). To filter these out, we generated a list of adult content terms and fuzzy-matched each query with the list of terms. We removed the query from our data if there was a reasonable match between the query and any of the terms. After completing this process, we retained roughly $40,000$ queries.

CPC and Competition Score SEMRush provides two outcomes that are relevant to answer our research question: CPC and Competition Score. CPC measures the average price advertisers pay when an ad is clicked. Competition Score is a proprietary normalized score (between 0 and 1) that measures the relative number of advertisers bidding for ad space for the given query.[9] It is worth noting that SEMRush collects data for many queries, even if they have little to no advertising, meaning CPC and Competition Score can take on the value of zero.

[7] See: `https://www.semrush.com/features/keyword-research/`

[8] For example, for the query "shoes", the database will show monthly data for "shoes" as well as similar queries, such as "running shoes", "women's shoes", "best shoes for hiking", etc.

[9] From discussions with SEMRush, Competition Score is linearly comparable and can be viewed similarly to the Google Trends information commonly used in research to measure search volume. We can't estimate the precise number of bidders, but we can estimate relative changes.

We collected historical monthly information about CPC and Competition Score for each of these queries from January 2018 to February 2020. We then filtered out queries that did not have persistent historical information (more than a year of missing data) and those with limited historical search volumes (less than 100 average searches/month during the time frame).

Our final dataset of queries contains 11,949 unique queries and 284,146 monthly observations. In the top part of Table 1, we present overall summary statistics for our dependent variables, Competition Score and CPC, across all queries from January 2018 to February 2020.

Table 1: Summary statistics.

	Mean	St. Dev.	Min	Max	Median
			All queries		
CPC	1.25	3.343	0.00	584.73	0.54
Competition score	0.244	0.356	0.00	1.00	0.05
			Short queries		
CPC	0.911	1.918	0.00	64.110	0.400
Competition Score	0.248	0.364	0.00	1.00	0.050
			Medium queries		
CPC	1.339	3.947	0.00	584.730	0.610
Competition Score	0.278	0.378	0.00	1.00	0.060
			Long queries		
CPC	1.533	3.078	0.00	83.860	0.720
Competition Score	0.126	0.216	0.00	1.00	0.030

Table 1 show that, in the aggregate, 1) many queries are not competitive (low median and mean Competition Score) and 2) CPC is often relatively low, but the maximum price can be extremely high. We also break down queries by word length, an important moderator we focus on in our analysis. We group queries into three categories for ease of presenting results. A short query has two or fewer words, a medium query has three to five words, and a long query has six or more words. We present summary statistics by query length in the

bottom part of Table 1. The table shows that as query length increases, demand generally decreases while CPC increases. These results generally align with the trend that, as query length increases, the queries become more niche and valuable to advertisers but also harder to target effectively.

4 Empirical Strategy

We provide two identification strategies that rely on different underlying assumptions to identify BERT's effect on CPC and CS. This section will describe our preferred identification strategy. Section 5 presents the alternative identification strategy.

As a first attempt to measure the impact of BERT on CPC and CS, we exploit the panel nature of our data and compare changes in each query's CPC and CS before and after the introduction of BERT with changes in the same query's CPC and CS in the year before the introduction of BERT. In doing so, we implement a strategy akin to a difference-in-differences (DD) where the treated and control queries are the same, but outcomes are observed across years.[10] Our identification strategy is similar to those employed in recent papers, such as Eichenbaum et al. (2020), Bollinger et al. (2022), and Liaukonyte et al. (2022).

This strategy aims to use variation across time within queries to identify BERT. An advantage of this strategy is that selection into the treatment is not an issue since treated and control queries are the same. However, time-varying unobservables that vary at the year-month level may bias our results. For example, our results could be positively biased if some market factor motivated Google to launch BERT in October 2019. Alternatively, queries may not be comparable across years if a query correlates with year-month-specific unobservable events, e.g., large concerts or sporting events. Both issues are classic DD challenges akin to unit-time-specific events tampering with causal effect estimates.

[10] In classic DD settings, outcomes are observed over the same period, but treated and control units are different.

We operationalize this identification strategy using the following model:

$$Y_{ijt} = \beta_1 Treated_{ij} \times Post_t + \delta_i + \gamma_j + \psi_t + \epsilon_{ijt}, \tag{1}$$

where $Y_{i,j,t}$ is the outcome of interests, either $log(CPC)$ or $log(CS)$,[11] for query i, year group j and month t. $Treated_{ij}$ is a binary indicator that takes on the value of one when query i is observed during the treated year-group window (July 2019- February 2020), zero otherwise (July 2018-February 2019).[12] $Post_t$ is a dummy that takes on the value one if the month t is after the month Google introduced BERT (November-February), zero otherwise (July-October). For our balanced eight-month time window, the months of July–October have $Post_t = 0$, while November–February has $Post_t = 1$. We include query fixed effects (δ_i) to control for time-invariant query unobservables, year-group fixed effects (γ_j) to control for group-specific shocks impacting all queries (e.g., in 2019–20, demand is higher for search ads), and month-of-year fixed effects (ψ_t) to account for monthly shocks (e.g., holidays and advertiser monthly spend). We cluster standard errors at the query level to account for potential serial correlations in the dependent variable and estimate the model using OLS. The parameter of interest is β_1, representing the incremental impact of BERT on sponsored search auction outcomes of interest.

We estimate Equation 1 from July to February so that we have four pre-BERT months (July-October) and four post-BERT months (November-February).[13] In addition, we focus on two periods, July 2018–February 2019 (control period) and July 2019–February 2020 (treated period). In Appendix B, we show that results hold with additional control year groups, specifically, July 2016–February 2017 and July 2017–February 2018 months.

[11] We empirically use $ln(CPC + 1)$ and $ln(CS + 1)$ to account for 0 CPC and CS.

[12] We refer to j as a group rather than a year because our treated (control) units fall across multiple years. For example, the treated group includes the years 2019 and 2020 because the treated period spans July 2019 to February 2020.

[13] We limit the post-BERT period to February because we want to avoid picking up any COVID-19 effect, which started impacting the US in March 2020.

4.1 Identification Checks

The key assumption behind our DD identification strategy is that no unobserved time-variant, group-specific shocks correlate with the entry of BERT and auction market outcomes.

In our setting, we worry about changes in advertiser or consumer search behavior due to time-varying unobservable events unrelated to BERT but that correlate with BERT's release. It is possible that consumer search demand characteristics differ across years and that consumer behavior or advertiser behavior in, say, the holiday season in 2018 is different than that in 2019. If this is the case, the results we observe may be due to unobservable time-varying changes from consumers or advertisers. This concern is analogous to worrying about unit-specific time-varying unobservable events correlated with the treatment event in a DD across states or cities.

To reduce concerns about these issues with this identification strategy, we do two things. First, to alleviate demand-side consumer shocks, our sampling procedure removes queries that have intermittent or volatile search volume that may correlate with unobservable time-varying events. For example, a query about a political event that may see a spike in search in a particular month and then quickly disappear will not be part of our sample. Doing this restricts our analysis and results to primarily persistently searched queries in order to minimize exposure to volatile, time-varying search queries.

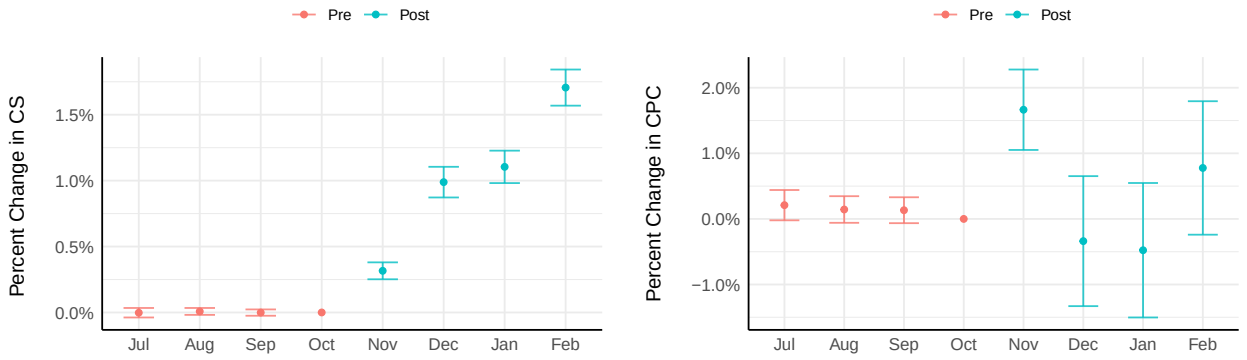
Second, as is common in DD analyses, we show that treated and control units behaved similarly before the introduction of BERT (parallel trends assumption) suggesting that the year prior to BERT is a good counterfactual. We do so by implementing an event study design.

To obtain event study parameter estimates, we estimate the following specifications:

$$Y_{ijt} = \beta_1 Treated_{ij} \times Month_t + \delta_i + \gamma_j + \psi_t + \epsilon_{ijt}, \tag{2}$$

Everything is as in Equation 1, but we replace our binary *Post* variable with eight monthly dummies that indicate the eight-month window from July to February. We estimate Equation 2 for both $log(CS)$ and $log(CPC)$, setting the baseline level for the monthly dummies to be October (the month of BERT's introduction).

Figure 1: Percent changes to CS and CPC across all queries. Plots visualize β_1 in Equation 2. Error bars represent 95% confidence intervals.



In the left panel of Figure 1, we present the event study parameter estimates for changes in $log(CS)$. Similarly, in the right panel of Figure 1, we present the event study for $log(CPC)$. In the months before BERT, we see no significant difference in marginal changes to $log(CPC)$ or $log(CS)$ between our control (2018-19) months and treated (2019-20) months. These results support the necessary parallel trends assumption. In addition, and foreshadowing our aggregate results, we observe a significant increase in CS in the post-BERT periods, while changes in CPC are much noisier and primarily null.

4.2 Results

Aggregate Effect of BERT We begin by analyzing aggregate changes to $log(CS)$ and $log(CPC)$ for all queries. To formalize the visual findings provided by the event study presented in Section 4.1, we offer the estimates of Equation 1 in Table 2. We find that CS increases by roughly 1% after the introduction time of BERT. At the same time, we find a positive but insignificant aggregate 0.3% increase in CPC. Thus, our aggregate market

results suggest that BERT is helping Google increase the average number of auction bidders, though this doesn't necessarily translate to higher CPC. When inspecting the event study plot for CPC in the right panel of Figure 1, we also see high standard error estimates post BERT, suggesting that there are potentially heterogeneous changes to CPC. We will come back to this point after discussing competitive vs. non-competitive queries.

Table 2: The effect of BERT on $log(CS)$ and $log(CPC)$.

	(1) log(CS)	(2) log(CPC)
Post × Treated	0.01***	0.003
	(0.0005)	(0.004)
Observations	191,161	191,161
R^2	0.962	0.728

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-group, month, and query fixed effects. Standard errors clustered at the query level are reported in parentheses.

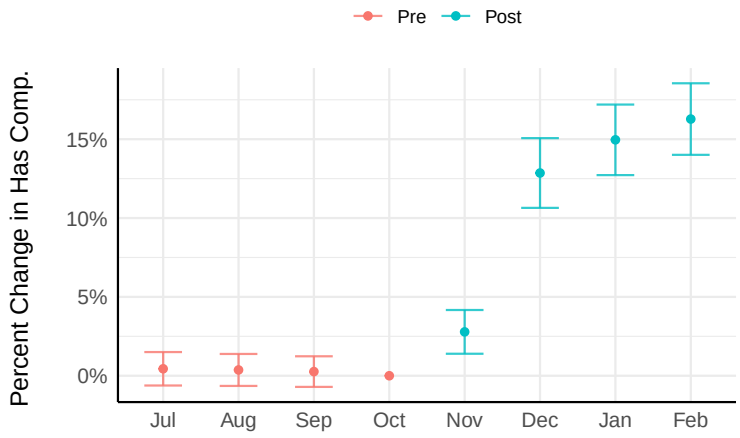
Competitive vs. Non-Competitive Queries One potential explanation for our empirical findings in Table 2 is that BERT could increase the supply of advertising space (Athey and Gans, 2010). Search engines generally avoid advertising on queries they don't understand (Broder et al., 2009). We partially observe this in our data. Roughly 20% of our queries maintained persistent 0's before October for both Competition Score and CPC in 2018 and 2019. If BERT helps Google understand more queries, causing Google to become more confident in advertising opportunities, overall ad space supply may increase, leading to higher aggregate number of bidders. (Auctions with zero CS now have non-zero CS). This may drive our aggregate increase in CS.

To test this hypothesis, we split our data into competitive and non-competitive queries. We deem a query non-competitive if the mode Competition Score during the July-October

period in 2018 and 2019 is zero. We find that roughly 20% of our queries are deemed non-competitive.

Taking the non-competitive queries, we then create a binary variable called $Has\ Competition_{ijt}$ that takes on the value one if query i's Competition Score is greater than zero for a given year-group j month t, zero otherwise. Using this variable, we can measure what proportion of non-competitive auctions become competitive in the post-BERT periods. We then estimate a linear probability model using Equations 1 and 2, using $Has\ Competition_{ijt}$ as our dependent variable. In Figure 2, we visualize the event study parameter estimates.

Figure 2: The effect of BERT on Has Competition for non-competitive queries. Error bars represent 95% confidence intervals.



Consistent with our expectations, we see evidence that BERT increases ad space supply. Table 3 suggests that BERT converts roughly 11.5% of non-competitive queries into competitive markets.

Before BERT, non-competitive queries generally maintained 0 CPC due to the lack of bidding advertisers. By making these auctions competitive, CPC naturally increases.[14] One potential explanation for our aggregate null CPC results could be that non-competitive

[14]As a sanity check, we estimated Equation 1 using $log(CPC)$ as the dependent variable for the queries deemed non-competitive before BERT. We find a positive and statistically significant coefficient (0.03 with $p = 0.0005$), suggesting that CPC did rise with the rise in the number of competitive auctions for the non-competitive queries following the introduction of BERT.

Table 3: The effect of BERT on Has Competition for non-competitive queries.

	(1)
Post × Treated	0.114***
	(0.009)
Observations	30,474
R^2	0.380

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-group, month, and query fixed effects. Standard errors clustered at the query level are reported in parentheses.

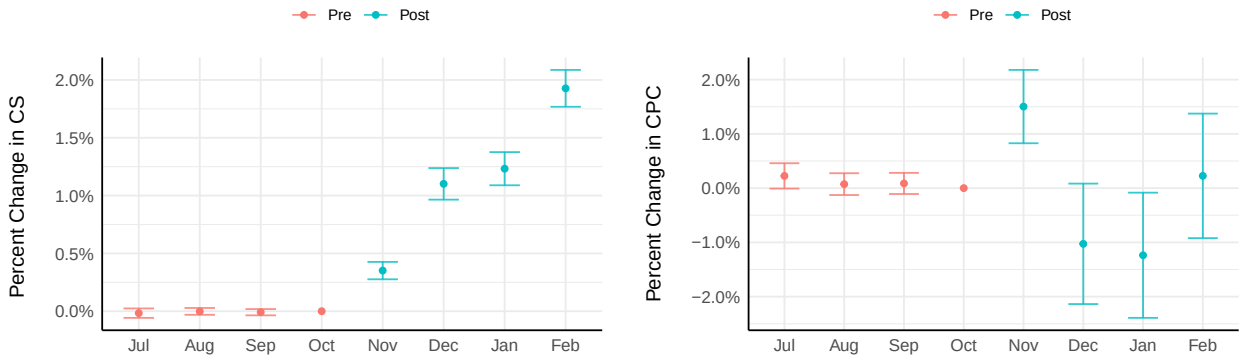
queries see prices increase while competitive queries see prices decrease, leading to null CPC changes.

We test this hypothesis by re-estimating Equations 1 and 2 using data for only the queries deemed competitive before BERT (the other 80%). These competitive queries are of primary interest because they allow us to test how BERT's new information impacts auction markets that existed before BERT. In Figure 3, we present event analysis estimates for changes in $log(CS)$ (left panel) and $log(CPC)$ (right panel) for these queries. We see evidence to support our parallel trends assumption and continue to see consistent results with those seen in Figure 1, suggesting that increasing the supply of advertising space is not the primary driver of our initial aggregate results (increase in the average number of bidders and null CPC).

In Table 4, we formalize our event analysis results. Consistent with our initial aggregate analysis, we find a statistically significant increase in CS (1.2% increase) and an insignificant 0.2% decrease in CPC.[15] Interestingly, this means BERT caused the average number of bidders to increase for existing auction markets, though the rise in the average number of bidders did not translate to a higher average CPC. We now turn our analysis to understanding

[15]As an alternative specification, we analyze how CS changes without being logged in Appendix C. We find consistent results.

Figure 3: Percent change in Competition Score and CPC for competitive queries. Error bars represent 95% confidence intervals.



why CPC has roughly zero average change.

Table 4: The effect of BERT on $log(CS)$ and $log(CPC)$ for competitive queries.

	(1) log(CS)	(2) log(CPC)
Post × Treated	0.012***	-0.002
	(0.0005)	(0.005)
Observations	160,687	160,687
R^2	0.961	0.715

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard errors clustered at the query level are reported in parentheses.

Heterogeneity Analysis by Query Length Computer Science research has found that BERT can identify more sophisticated language structures present in longer sentences and adequately account for contextual information within language objects (Jawahar et al., 2019; Goldberg, 2019; Tenney et al., 2019b). In other words, improvements to interpretation

quality likely increase with query length.[16]

In addition, BERT improves the semantic interpretation of search queries and the relations between queries. This knowledge will impact all queries, not just long, complicated queries. Take the short query "socks". With BERT, Google may learn that "socks" relates to other queries, such as "shoes" or "sandals" because each query falls within the "attire" topic category. This information will likely impact advertiser matching.

We hypothesize that keyword length moderates the effect of BERT due to inherent linguistic differences across queries. For short queries, we expect BERT to learn information that helps identify more semantic relationships to other queries because these queries are generally broader. This leads to more advertisers being deemed relevant to short query auctions, causing the average number of bidders and CPC to increase.

For longer queries, we expect BERT to better understand the complex linguistic information in these queries and subsequently improve interpretation quality. Because these queries tend to be more specific, Google will filter out irrelevant advertisers and segment long query markets into smaller markets, leading to fewer bidders and lower CPC.

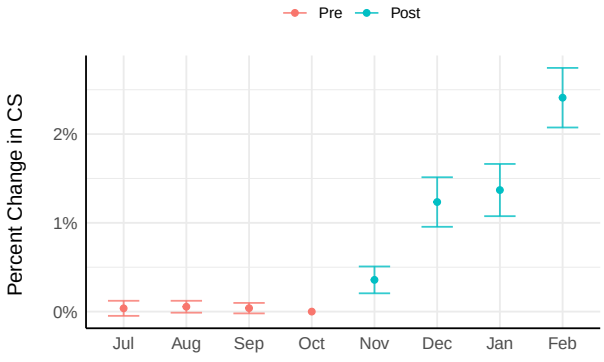
For our analysis and visualizations, we define short queries as queries with two or fewer words, medium queries with three to five words, and long queries with six or more words. Using these categories, we re-estimate Equation 1 by query length using only the queries deemed competitive before BERT. In Figure 4, we present event study plots for CS and CPC by query length. Note that, in all cases, we continue to find parallel pre-treatment trends.

In Tables 5 and 6, we present the estimates by query length for $log(CS)$ and $log(CPC)$, respectively.

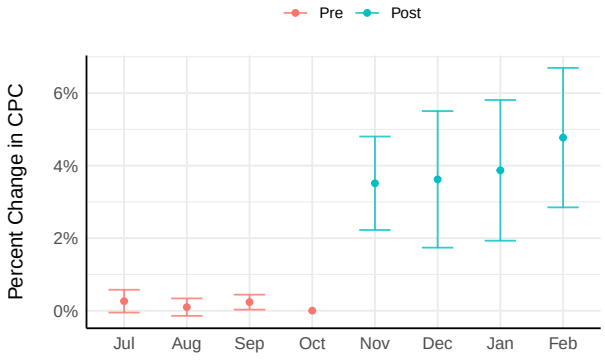
Consistent with our hypothesis, query length moderates how BERT affects existing auction markets. We observe a 3.8% increase in CPC and a 1.3% increase in CS for short

[16]In the blog announcing the roll-out of BERT, Google also stated that they expected BERT to predominantly impact the interpretation quality of longer, more complex queries. See Here: `https://blog.google/products/search/search-language-understanding-bert/`

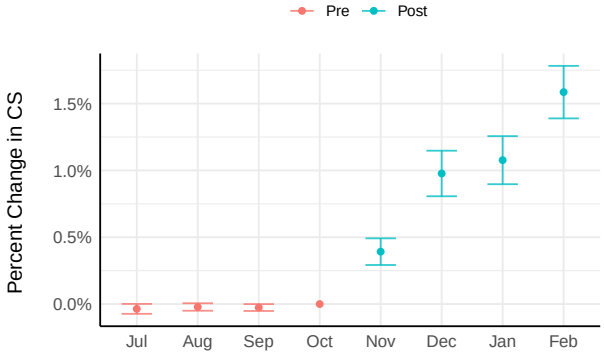
Figure 4: Percent change in Competition Score and CPC by query length. Error bars represent 95% confidence intervals.



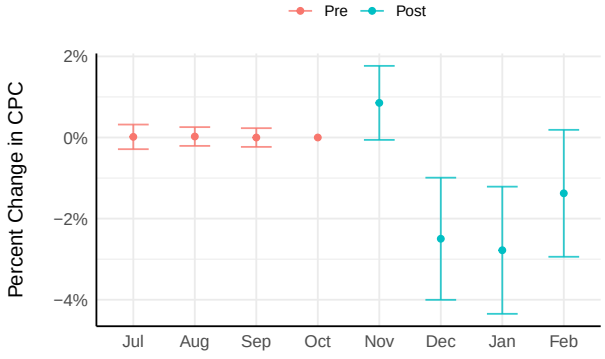
(a) Short query CS.



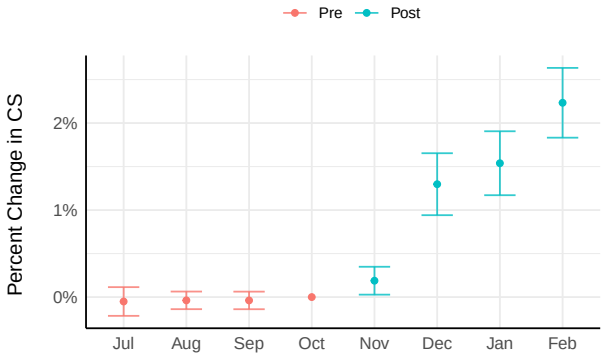
(b) Short query CPC.



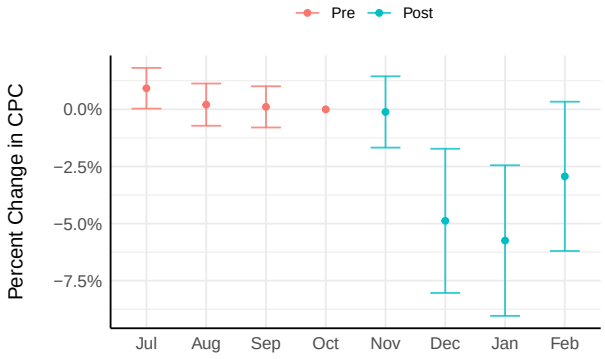
(c) Medium query CS.



(d) Medium query CPC.



(e) Long query CS.



(f) Long query CPC.

Table 5: The effect of BERT on $log(CS)$ by query length.

	(1) Short	(2) Medium	(3) Long
Post × Treated	0.013***	0.010***	0.013***
	(0.001)	(0.001)	(0.001)
Observations	47,912	88,888	23,887
R^2	0.951	0.971	0.920

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard errors clustered at the query level are reported in parentheses.

Table 6: The effect of BERT on $log(CPC)$ by query length.

	(1) Short	(2) Medium	(3) Long
Post × Treated	0.038***	-0.015**	−0.037**
	(0.008)	(0.006)	(0.013)
Observations	47,912	88,888	23,887
R^2	0.686	0.728	0.671

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard errors clustered at the query level are reported in parentheses.

queries. However, as query length increases, CPC quickly decreases while CS continues to increase. Medium queries see CPC decrease by 1.5% and CS increase by 1%. Similarly, long queries see prices drop by 3.7% despite CS rising by roughly 1.3%. The heterogeneous changes to CPC align with our hypotheses and explain why we estimated null average results. However, the average number of bidders didn't adjust as we expected and the reason why the rise in the average number of bidders does not translate to higher prices remains unclear. We address this question in Section 6, where we provide a theoretical explanation for our observed results. Before presenting the theoretical model, we discuss an alternative

identification strategy to reinforce our results' causal interpretation.

5 Alternative Identification Strategy

The DD identification strategy results just discussed will be biased if there exists a large time-varying confound correlated with the introduction of BERT. Given our findings, such a confound would need to be a significant event that uniformly drives up the number of bidders across queries and differentially affects CPC across query lengths. While such a confounder seems unlikely, as in any observational study, we cannot completely rule it out.

To strengthen the validity of our results, we present an alternative identification strategy that exploits the fact that BERT is likely to affect queries differently due to their inherent linguistic properties. In doing so, we compare queries more likely to be affected by BERT with queries less likely to be affected by BERT before and after the introduction of BERT. Under this identification strategy, a confound must be year-month specific *and* correlate with query linguistic properties.

5.1 Defining Query Metrics

The Computer Science literature has documented that BERT better understands complex syntactic language structures and semantic relationships between words and sentences (Devlin et al., 2018; Lin et al., 2019; Tenney et al., 2019a,b; Rogers et al., 2021). A complex query will benefit from BERT's ability to capture syntactic information, while all queries, including those with simple syntactic structures, will experience changes to semantic relationships and understanding.

Motivated by these observations, we create two measurement variables, Linguistic Complexity and Cosine, to capture query syntactic complexity and semantics. Under the assumption that the interpretation and use of these linguistic properties changed with the introduction of BERT, we can identify variation in our dependent variables caused by BERT's

implementation interacting with our measurements.

Linguistic Complexity BERT can better understand complex syntactic language structures (Lin et al., 2019; Tenney et al., 2019a). Therefore, BERT's introduction should change how Google handles complex syntactic language. Examples of complex syntactic information include query syntax tree structures (Lin et al., 2019), parts of speech (Tenney et al., 2019a), and sentence dependency features Tenney et al. (2019a); Liu et al. (2019). Hewitt and Manning (2019) finds that syntax trees, i.e., hierarchical characterizations of grammatical language structures, are embedded in BERT vector spaces and Tenney et al. (2019b) finds that Parts of Speech (POS) tags are also present in BERT vector spaces. These findings help us identify the information BERT will interact with and drive the design of our first measurement: Linguistic Complexity (LC).

For each query, we measure the depth of the syntax tree and count the unique number of Parts Of Speech (POS). LC is defined as a dummy variable and takes on a value of 1 if either a query's syntax tree depth is greater than the median depth in our dataset (median = 2) or the unique POS count is greater than the median county in our dataset (median = 2). Otherwise, it's 0.[17] LC captures the syntactic complexity of the query.

Cosine BERT understands semantics and how words and concepts *relate* to each other (Tenney et al., 2019b). It knows that socks relate to shoes, banks can relate to bodies of water or financial institutions, and computers can sometimes relate to mice. This semantic knowledge improves Google's ability to interpret and categorize search queries, ultimately affecting the relationships between queries. This latter observation is critical as changes to query relations likely impact Google's matching process and identification of relevant advertisers. Understanding that a query relates to more (fewer) queries will likely lead to more (fewer) advertisers being deemed relevant.These observations motivate the design of our second linguistic property: Cosine.

[17]In Appendix D.1, we present results using a continuous version of LC.

At a high level, we measure semantic changes (Cosine) using the difference in the number of queries related to a given query before and after BERT.

To define Cosine, we must first identify the primary interpretation algorithm used by Google before BERT (RankBrain). While we cannot be certain, RankBrain is likely Doc2Vec (D2V) or Word2Vec (W2V), the former being a more generalized form of W2V.[18]

For a query i, we generate a vector representation with BERT and D2V models. We then calculate the cosine similarity score between query i and all other queries in the dataset *within* a given model vector space. This generates two $N \times N$ cosine similarity matrices, one for the BERT vector space and one for the D2V vector space, where N is the number of queries in our dataset and matrix element i, j is the cosine similarity score between query i and query j.

We use these matrices to measure changes in query semantic relationships across D2V and BERT vector spaces. Specifically, we measure the set of queries that pass "relatedness" thresholds *within* vector spaces and then compare differences in these sets *across* vector spaces. We first calculate each cosine matrix's 25th, 50th, and 75th percentile cosine scores. These are our "relatedness" thresholds.[19] Then, for each query i, we identify the set of other queries with a cosine similarity score greater than or equal to the chosen threshold (i.e., 25, 50, or 75) in the respective vector spaces. For example, if the 75th threshold in the D2V space is 0.5 and the query "shoes" has a cosine similarity score of 0.8 with the query "socks" and 0.2 for the query "hat", then the "socks" query will be in the relevant set for the D2V space. If, in the BERT space, the 75th threshold is 0.8 and the cosine score between "shoes" and "hats" is 0.7, then "hat" would now appear in the relevancy set for "shoes" in the BERT space. This gives us two sets of relevant queries per query, one for each vector space.

[18]Google filed patents: `https://patents.google.com/patent/US9740680B1/en`. See also: `https://opensource.googleblog.com/2013/08/learning-meaning-behind-words.html` for technology before RankBrain that looks similar to Word2Vec, and the researchers for both Word2Vec and Doc2Vec were employed at Google.

[19]We vary the thresholds to generalize the measurement. We focus on thresholds because we reason that Google likely has some cut-off requirement determining whether an advertiser is relevant enough to a given query.

Then, we calculate the differences between the relevancy sets across BERT and D2V for a given query. We define "added" queries as those that did not appear in the D2V relevancy set but did appear in the BERT set. Alternatively, we define "removed" queries as those that appeared in the D2V set but did not appear in the BERT set. For a given query i, we sum up the total number of "added" and "removed" queries and take the ratio between the two sums. We define Cosine as this ratio.

A ratio value equal to one tells us that just as many queries were added as removed. Ratio values less than one indicate that more queries were removed than added, indicating that BERT believes the query is more specific and related to fewer other queries. Finally, ratio values greater than 1 tell us that more queries were added than removed, suggesting that BERT believes the query is related to more queries. Because the ratio of added to removed queries is skewed to the right (median = 1.094, mean = 175.605), we use the log of 1 + Cosine in our model.[20]

It is worth noting that we use this process to generate Cosine because the D2V and BERT vector spaces and their cosine scores are not directly comparable without making strong assumptions about what the dimensions of each model's vector space represents. D2V and BERT capture potentially different sets of information, warranting vector comparison and projection methods infeasible. Additionally, the scales of these vector spaces are potentially different, meaning we cannot directly compare query cosine scores across algorithms. Therefore, to effectively measure query semantic changes, we must adequately standardize measurements *within* vector spaces such that they can then be compared *across* vector spaces.

Average Query Metrics Table 7 and 8 present average Cosine and LC measures, respectively, for all queries ("All") and by keyword length. Table 8 also presents the proportion of queries with POS counts and syntax tree depths greater than the median. (A value of 1 means above the median for the particular measurement). We identify several patterns.

[20]In Appendix D.2, we show that our results are robust to how we define Cosine.

First, LC measures correlate with query length due to the presence of within-query contextual information, inherently making it longer. Second, short queries have, on average, a large Cosine score, while long queries have a low Cosine score. The average long query ratio between "added" and "removed" is 0.78, indicating that BERT finds these queries more specific than D2V.

Table 7: Average Log(Cosine).

Keyword Length	75th Ratio	50th Ratio	25th Ratio
All	1.761	1.324	1.049
Short	3.107	2.348	1.645
Medium	1.393	1.043	0.914
Long	0.650	0.481	0.451

These values are the average log(Cosine + 1), where Cosine is the ratio between added and removed sets.

Table 8: Average Linguistic Complexity.

Keyword Length	Linguistic Complexity	Tree Height	POS
All	0.613	0.566	0.400
Short	0.004	0.001	0.004
Medium	0.811	0.730	0.439
Long	1.000	0.998	0.986

5.2 Specification

Given the two linguistic metrics defined above, we estimate the following model:

$$Y_{i,t} = \beta_1 Post_t \times LC_i + \beta_2 Post_t \times log(Cos_i) + \beta_3 Post_t \times LC_i \times log(Cos_i) + \delta_i + \gamma_t + \epsilon_{i,t}, \quad (3)$$

where *Post* takes on a value of 1 for the months post-BERT (November to February), and 0 otherwise. LC_i is Language Complexity of query i and Cos_i is the the Cosine of query i. δ_i are query fixed effects and γ_t year-month fixed effects. β_1 captures the variation in Y that

is explained by the effect of LC after BERT gets introduced, β_2 captures the variation in Y explained by the effect of changes to Cosine after the introduction of BERT, and β_3 captures the interaction between LC and Cos after the introduction of BERT.

We include query and year-month fixed effects, δ_i and γ_t, respectively. We estimate Equation 3 using data from July 2019 to February 2020 and cluster standard errors at the query level. To replicate the findings in the previous section, we restrict this analysis to competitive queries, i.e., those queries with non-zero mode CS scores from July to October 2019.

Predicted Measurement Effects LC captures a query's syntactic complexity and whether it contains contextual information. BERT's ability to handle complex syntax structures will benefit complex queries, improving interpretation quality. Due to better interpretation, we expect Google to offer more advertisements on these complex queries, leading to higher CS and CPC.

Cosine captures changes in "relatedness" or semantic relationships between queries. Specific queries see small Cosine scores, while vague or general queries see larger ones due to BERT learning more semantic associations. We predict that an increase in Cosine will increase CS and CPC. Learning that a query is associated with more queries will lead to more advertisers being deemed relevant.

Linguistically complex queries contain contextual information and are generally more specific, meaning they can be affected by both LC and Cosine. For these queries, BERT can capture new information in the query *and* learn different semantic relationships. (BERT can only affect semantic relationships for simple queries due to their inherent lack of syntactic complexity). We hypothesize that Cosine will have a differentially negative impact, specifically for linguistically complex queries, due to BERT learning that these queries are specific and causing advertisers to get filtered out. Therefore, we estimate an interaction term between Cosine and LC. We expect this interaction term will be negative for CS and

CPC because it indicates BERT is filtering out irrelevant advertisers and increasing relevancy scores for those that do compete in the auctions.

Identification Assumptions This identification strategy relies on the assumptions that: (1) the linguistic properties we defined affect CPC and CS only through their interaction with the query interpretation system; (2) BERT interacts with our linguistic measurements differently than the previous algorithm; (3) queries with high and low values of Linguistic Complexity and Cosine are comparable; (4) there are no time-varying-query linguistic type-specific confounders.[21]

To partially validate these assumptions, we perform three tests. To test assumption three, in Appendix D.3, we use an event study-like model to show that we estimate null pre-trends. To account for consumer search behavior, in Appendix D.4, we show that results hold when we control search volume. Finally, in Section 5.4, we perform a placebo test that shows that our estimates are likely due to the introduction of BERT (partially validating assumption one and two).

5.3 Results

We present the estimates of Equation 3 using CS as dependent variable in Table 9. $Post_t \times LC_i$ is positive and significant, suggesting that when Linguistic Complexity is above the median, CS increases by $0.33-0.4\%$. These findings suggest that when syntactic information acquisition opportunities are significant, BERT's information increases Google's confidence that it correctly interprets the query and increases auction density. Second, consistent with our expectations, $Post_t \times Cos_i$ is also positive and significant. The coefficient estimates suggest that a 1% increase Cosine translates to about a 0.1% increase in CS. Finally, as expected, the interaction term between both LC and Cosine is negative. Linguistically

[21]For example, changes in search behavior that are query-type specific and unrelated to BERT can be problematic. Consider the case in which wealthier consumers begin submitting search queries that are more linguistically complex because of something other than BERT. If advertisers change their targeting strategies and increase their bids in response to this behavior, our results will be upward biased.

complex queries are generally more specific, and BERT learns this, leading to partial market filtering and some removal of advertisers.

Table 9: Alternative Strategy: $log(CS)$

	(1) 75th	(2) 50th	(3) 25th
Post × Linguistic Complexity	0.0040**	0.0033*	0.0033*
	(0.002)	(0.002)	(0.002)
Post × Log(Cosine)	0.0012**	0.0009*	0.0007
	(0.001)	(0.001)	(0.001)
Post × Linguistic Complexity × Log(Cosine)	−0.0015**	−0.0020**	−0.0030**
	(0.001)	(0.001)	(0.001)
Observations	63,474	63,474	63,474
R^2	0.9786	0.9785	0.9785

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data from July 2019 to Feb 2020.

To put these results into context, we can use the average keyword length values in Table 7 and Table 8 to bootstrap predicted changes to the average number of bidders. We find that aggregate CS increases by a conservative $0.1 - 0.3\%$. When we predict average values by keyword length, we find that, across lengths, CS is increasing by about $0.1 - 0.4\%$ (See Table 10). The uniform increase in CS is consistent with our empirical observations using the main DD setup. All estimates are more conservative, likely due to measurement errors in our treatment variables.

In Table 11, we present the results for $log(CPC)$. Consistent with our expectations, a 1% increase in Cosine positively increases CPC by about 0.6%. Inconsistent with our predictions, we find that $Post_t \times LC_i$ is negative and significant. Finally, the triple interaction is close to zero and not significant.

We again use the average values in Tables 7 and 8 to put these results into context. We

Table 10: Predicted $log(CS)$ estimates by model specification.

Keyword Length	75th	50th	25th
All	0.003***	0.002***	0.001***
Short	0.004***	0.002***	0.001***
Medium	0.003***	0.002***	0.001***
Long	0.004***	0.003***	0.002***

Note: 99% confidence intervals are estimated by bootstrapping predictions (1000 iterations). ***p<0.01.

present the estimated effect on CPC in Table 12. We find that the average CPC decreases by roughly 0.8%. When we break it down by keyword length, we find that CPC increases for short queries by roughly 1.5%, medium queries see CPC decrease by about 1.6%, and long queries see CPC decrease by approximately 2.5%. These average values are broadly consistent with our primary DD analysis, though we note that average CPC declines under this specification and is null in our main DD result. .

5.4 Placebo Test

Our identification rests on the assumption that the effects we observe are due to the linguistic variables interacting with a change in the interpretation algorithm (i.e., BERT). We support this assumption by performing a placebo test. We estimate 3 using the same period but a year before (July 2018 to February 2019) and create a placebo post-BERT variable that takes on a value of 1 for November through February, 0 otherwise. Since no known change exists that we suspect interacts with these measurements during this time frame, we expect to estimate insignificant results. In Tables 13 and 14 we present null results.

6 Modeling the Market

A standard auction model would assume that increasing the number of bidders would lead to weakly increasing prices. Better matching with horizontally differentiated buyers could

Table 11: Alternative Identification: $log(CPC)$

	(1) 75th	(2) 50th	(3) 25th
Post × Linguistic Complexity	−0.0270*	−0.0289**	−0.0307**
	(0.015)	(0.014)	(0.015)
Post × Log(Cosine)	0.0056*	0.0065*	0.0075
	(0.003)	(0.004)	(0.005)
Post × Linguistic Complexity × Log(Cosine)	−0.00003	0.0009	0.0006
	(0.005)	(0.006)	(0.009)
Observations	63,474	63,474	63,474
R^2	0.795	0.795	0.795

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data from July 2019 to Feb 2020.

also put upward pressure on prices (Tadelis and Zettelmeyer, 2015). Yet, in our context, we empirically observe prices sometimes increasing or decreasing despite all queries seeing an increase in the number of bidders. We propose that these results are driven by BERT generating more informative signals about multi-dimensional query types, allowing Google to better organize the market and differentiate bidders within certain auctions.

6.1 Model Overview

We first provide a high-level overview of the model before discussing the setup and results. After a consumer submits a search query, the platform (Google) uses an interpretation algorithm to generate a signal about the query search intent. A better algorithm generates a more informative signal for the seller. After receiving the signal, the platform estimates advertiser relevancy scores and picks buyers for the ad auction. Buyers submit bids if selected and ranked based on their Ad Ranks. The winner gets to then show an ad. If the ad is clicked on, the advertiser pays just enough to beat out the next highest bidder. We now

Table 12: Predicted $log(CPC)$ estimates by model specification.

Keyword Length	75th	50th	25th
All	-0.007^{***}	-0.008^{***}	-0.011^{***}
Short	0.017^{***}	0.015^{***}	0.012^{***}
Medium	-0.014^{***}	-0.016^{***}	-0.018^{***}
Long	-0.023^{***}	-0.025^{***}	-0.027^{***}

Note: 99% confidence intervals are estimated by bootstrapping predictions (1000 iterations). $^{***}p<0.01$.

discuss two model assumptions before describing the model.

6.2 Considerations

There are two novel components to our empirical setting. Motivated by psychology, neuroscience, linguistics, and computer science literature, we propose that language objects (queries) maintain a multi-dimensional type structure (Mnih and Hinton, 2008; Jäger and Rogers, 2012; Miyagawa et al., 2013; Coopmans et al., 2023). Specifically, we define a query by a topic and context. The topic establishes the category or focus of the query, while context is nuanced information within the query that differentiates the query's search intent within a given topic. Second, as previously mentioned, the platform (Google) gets to endogenously pick its buyers before running an auction. Therefore, advertisers must be allocated to an auction opportunity to compete for advertising space.

While not novel to our setting, it is essential to note that advertisers are horizontally differentiated (e.g., Nike and Geico are relevant to different queries), and Google does not advertise on all queries due to negative externality costs associated with showing potentially irrelevant advertisements (Broder et al., 2009). We now describe our model.

6.3 Setup

Query Our market contains a platform (Google) that sells query ad space to advertisers. Ad space for a query gets sold via a second-price auction (SPA). Queries are defined by a topic

Table 13: Alternative Strategy Placebo Test: $log(CS)$.

	(1) 75th	(2) 50th	(3) 25th
Post × Linguistic Complexity	0.001	0.001	0.0004
	(0.001)	(0.001)	(0.001)
Post × Log(Cosine)	0.0003	0.0003	0.0004
	(0.0002)	(0.0002)	(0.0004)
Post × Linguistic Complexity × Log(Cosine)	−0.0002	−0.0002	−0.00001
	(0.0002)	(0.0003)	(0.001)
Observations	63,468	63,468	63,468
R^2	0.997	0.997	0.997

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data from July 2018 to Feb 2019.

$t \in \{0, 1\}$, a context $q \in \{0, 1\}$, and a query length k. For a query Q_k of length $k \in \{s, m, l\}$, where s, m, and l indicate short, medium, and long queries, there are four possible query types ($< 0, 0 >$, $< 0, 1 >$, $< 1, 0 >$, and $< 1, 1 >$). t captures the underlying topic of the query (e.g., is it about "shoes" or "insurance") while q captures contextual information that differentiates the query type within a topic. Each query has an equally likely probability of being drawn ($\frac{1}{4}$).

To build intuition for the values of q and t, consider the queries "best-running shoes", "stores near me to buy running shoes", "Where to purchase life insurance", and "best insurance companies for retiring adults". The first two queries relate to "shoes" (t = "shoes"), while the latter two relate to "insurance" (t = "insurance"). However, each query has different types of contextual information that convey different search goals and needs. The first and last queries are focused on information acquisition, while the second and third are interested in making purchases. The different contexts convey different search intents within a given topic, which will impact the likelihood of clicking on different advertisers within the given

Table 14: Alternative Strategy Placebo Test: $log(CPC)$.

	(1) 75th	(2) 50th	(3) 25th
Post × Linguistic Complexity	0.001	0.002	0.001
	(0.002)	(0.002)	(0.002)
Post × Log(Cosine)	0.0004	0.0005	0.0004
	(0.001)	(0.001)	(0.001)
Post × Linguistic Complexity × Log(Cosine)	0.001	−0.0002	0.001
	(0.002)	(0.001)	(0.002)
Observations	63,468	63,468	63,468
R^2	0.979	0.979	0.979

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data from July 2018 to Feb 2019.

topic.

The importance of context varies along keyword lengths k. Short queries, such as "shoes", will still be defined with a topic and a context, but we assume the importance of the contextual factor matters less to the query's click-through rate (CTR). We discuss this point later in this section.

Advertiser An advertiser A_i is defined by a topic $\theta_i \in \{0, 1\}$ and context $z_i \in \{0, 1\}$. We will assume four bidders are in the market, one of each combination of topic and context ($< 0, 0 >, < 0, 1 >, < 1, 0 >$, and $< 1, 1 >$). θ and z directly map to t and q types for queries, respectively. For clarity and brevity, our primary analysis in Appendix E will focus on a single topic value $t = 1$ and two advertisers, one of each context type (0 and 1 types) for a particular query length k. Like context for queries, advertiser context captures heterogeneity *within* topics (i.e., is the "shoe" advertiser related to purchasing shoes, such as Nike's online store, or gathering information about shoes, such as a review website like Runblogger).

Advertisers maintain a private value $v_i \sim U[0, 1]$ for clicks. They do not compete in an

auction unless allocated to an auction opportunity. (More discussion on the selection process is below). Denote the winning advertiser by A_w, its type by θ_w, and its context by z_w. The CTR for a query Q_k depends on the match quality between the winning advertiser and the query. We model query CTR with Equation 4

$$CTR(A, Q) = \beta_{\theta,k} I_{\{\theta=t\}} + \beta_{z,k} I_{\{z=q\}} \tag{4}$$

where $\beta_{\theta,k} + \beta_{z,k} \leq 1$, $\beta_{\theta,k}, \beta_{z,k} > 0$, and $\beta_{\theta,k} \geq \beta_{z,k}$. The last assumption states that topic alignment matters most to a query's CTR. If the winning advertiser type is of the right type and context, the expected CTR probability is $\beta_{\theta,k} + \beta_{z,k}$. If the advertiser is the right topic (context) but not the right context (topic), CTR is $\beta_{\theta,k}$ ($\beta_{z,k}$). We allow $\beta_{\theta,k}$ and $\beta_{z,k}$ to vary along keyword lengths to capture differences in alignment importance.

Platform The platform endogenously picks its auction buyers (advertisers) and subsequently runs a pay-per-click (PPC) SPA for ad space. (We break ties with random selection). Denote the set of chosen advertisers by $\hat{A}$. We assume the platform sees advertiser types (θ) and advertiser context (z) for all advertisers before buyer selection. When a query Q_k arrives, the platform uses an algorithm $a \in \mathbb{R}$ to interpret it. For a query Q_k with components t and q, an algorithm a takes Q_k as input and outputs signal $\hat{Q}_k$ with components $\hat{t}$ and $\hat{q}$ to the platform.

For a given algorithm a and a query of length k, $Pr(\hat{t} = t) = \gamma_{t,k}(a)$ and $Pr(\hat{q} = q) = \gamma_{q,k}(a)$. For topics, with probability $\gamma_{t,k}(a)$ the platform learns the true value of t, and with probability $1 - \gamma_{t,k}(a)$ the platform receives an inconclusive signal. Similarly, with probability $\gamma_{q,k}(a)$, the platform learns the context, and with probability $1 - \gamma_{q,k}(a)$, the platform receives an inconclusive signal. We assume the inconclusive signal is equivalent to the platform's prior: $Pr(t = 1) = Pr(q = 1) = \frac{1}{2}$. Therefore, $\hat{t} \in \{0, \frac{1}{2}, 1\}$ and $\hat{q} \in \{0, \frac{1}{2}, 1\}$. We will assume that $\frac{\gamma_{t,k}}{\partial a} > 0$ and $\frac{\gamma_{q,k}}{\partial a} > 0$, i.e., a better algorithm always improves the probability of learning the true value of t and q. We assume advertisers and the platform

know $\gamma_{t,k}$ and $\gamma_{q,k}$.[22]

Ad Ranks determine the order of bidders. We model Ad Rank for an advertiser i as $Adr_i = b_i r_i(\hat{Q})$, where b_i is the submitted bid and $r_i(\hat{Q})$ is the relevancy score advertiser i receives from the platform given the information set $\hat{Q}_k$. The relevancy score takes on the functional form shown in Equation 5.

$$r_i(\hat{Q}) = \frac{Pr(\theta_i = t|\hat{t})\beta_{\theta,k} + Pr(z_i = q|\hat{q})\beta_{z,k}}{\beta_{\theta,k} + \beta_{z,k}} \tag{5}$$

We assume that the platform knows the CTR coefficients $(\beta_{\theta,k}, \beta_{z,k})$ for a query of length k and can correctly estimate relevancy scores given the available topic and context information set $\hat{Q}_k$. Importantly, when allocated to an auction, the advertiser does not know its relevancy score. In practice, advertisers do not get to see their relevancy score before bidding.

We present the platform's profit function for a given query Q and the set of chosen advertisers $\hat{A}$ in Equation 6.

$$\pi(\hat{A}, Q) = I_{\{Click\}} \frac{\tilde{Adr}}{r_w} - I_{\{\theta_w \neq t\}} C \tag{6}$$

In Equation 6, $\frac{\tilde{Adr}}{r_w}$ is the second highest Ad Rank in the auction scaled by the winning bidder's relevancy score, $I_{\{Click\}}$ indicates a click, $I_{\{\theta_w \neq t\}}$ indicates when there is a mismatch between the query topic and winning advertiser topic, and C is the negative externality cost associated with displaying an irrelevant advertisement to the consumer (i.e., wrong topic). $\frac{\tilde{Adr}}{r_w}$ is structured such that winning advertisers pay the minimum price they would need to bid to still win the position (Amaldoss et al., 2015). When the winning advertiser is the same type as the query, the search engine's expected profit before running the auction is $E[CTR_k]E[\frac{\tilde{Adr}}{r_w}]$. When the winning advertiser type does not align with the query type, the platform's profit is $E[CTR_k]E[\frac{\tilde{Adr}}{r_w}] - C$.

There are four possible information sets to observe for a given query Q_k. With probability

[22]This stems from the fact that Google announces the algorithm update and releases press articles that describe how new algorithms work.

$\gamma_{t,k}\gamma_{q,k}$ the search engine learns both the values of t and q ($\hat{t} = t, \hat{q} = q$). With probability $(1-\gamma_{q,k})\gamma_{t,k}$ the search engine learns the value of the topic t but not the context q ($\hat{t} = t, \hat{q} = \frac{1}{2}$). Finally, with probability $\gamma_{q,k}(1-\gamma_{t,k})$ the search engine learns the value of the context q but not the topic t ($\hat{t} = \frac{1}{2}, \hat{q} = q$) and with probability $(1-\gamma_{q,k})(1-\gamma_{t,k})$ the search engine receives an inconclusive signal ($\hat{t} = \frac{1}{2}, \hat{q} = \frac{1}{2}$).

After receiving the estimates $\hat{t}$ and $\hat{q}$, the platform picks which advertisers to add to the auction. Despite the presence of relevancy scores, conditional on being allocated to an auction, it is a weakly dominant strategy for advertisers to bid their valuations v_i. (Proof is in Appendix E). Advertisers do not pay unless the consumer clicks on their ad. We model the expected CPC for a query of length k with advertiser set $\hat{A}$ by $E[CPC_k(\hat{A})] = E[CTR_k] * E[\frac{\tilde{Adr}}{r_w}]$.

The order of the game is as follows. A query Q_k is randomly drawn by Nature, producing unobservable components q and t. A platform's algorithm a receives Q_k and estimates $\hat{q}$ and $\hat{t}$. The platform privately estimates relevancy scores, decides whether to run an auction with the given signal $\hat{Q}_k$, and then picks its buyers. If the seller runs an auction, chosen advertisers submit bids, the winning advertiser displays an ad, the consumer decides whether to click given the type alignment, and the platform receives a profit. An algorithm a affects profits, CPC, the average number of query bidders, and CTR by changing the seller's q and t signals.

6.4 Results

Our results will focus on auction outcomes at the query level and show how improving a can lead to, on average, more bidders, higher CTR, and increasing or decreasing CPC. We assume C is sufficiently large ($C > 1$) to simplify the analysis and primarily focus on price changes. All proofs are in Appendix E.

Proposition 1 (Platform Selection of Advertisers)**.** *Under sufficiently high cost C, it is a weakly dominant strategy for the platform only to run auctions when the query topic is*

known. The average number of bidders for query ad space is $2\gamma_{t,k}$ *and increases with* a*.*

Understanding a query's topic is critical to running auctions with relevant advertisers and increasing the average demand for individual queries. As long as the platform knows the topic, it can avoid the negative externality costs associated with showing an irrelevant ad and allocate relevant advertisers. A better algorithm leads to better topic identification and higher average demand. When the platform doesn't know the query topic, it is unwilling to run an auction because the negative externality costs of potentially showing an irrelevant ad outweigh the expected revenue. These externality costs explain why we do not see the platform (Google) run an auction for ad space in all instances or allocate all advertisers (buyers) to all potential auctions.

Corollary 1. *CTR increases with both topical and contextual information acquisition. Improvements to algorithm* a *lead to higher CTR. CTR increases faster as* $\beta_{z,k}$ *increases and potentially slower as* $\beta_{\theta,k}$ *increases.*

The result stems from the fact that better topic understanding leads to a higher likelihood of running an auction with relevant advertisers. Better contextual knowledge subsequently helps prioritize those advertisers who are more likely to receive a click. The extent to which CTR increases depends on how much understanding context and topic matters to the CTR likelihood. Since we expect long (l) query-advertiser context alignment to matter more for the CTR than short (s) query context ($\beta_{z,l} \geq \beta_{z,s}$), we predict that CTR should increase faster for long queries than short queries with BERT.

Corollary 2. *CPC decreases (increases) with improvements to* $\gamma_{q,k}$ *(*$\gamma_{t,k}$*).*

Holding all else equal, better identification of a query's topic leads to a higher likelihood of running an auction with relevant advertisers, leading to higher query-level average CPC. On the other hand, better contextual identification ($\gamma_{q,k}$) helps differentiate bidders *within* the auction, pushing the high CTR advertisers to the top and rewarding them for their relevancy by lowering their final bid costs. However, our interest is in understanding how an

algorithm change, which will impact both $\gamma_{q,k}$ and $\gamma_{t,k}$, impacts CPC. The following theorem shows under what conditions CPC increases (decreases).

Proposition 2. *Let $\gamma^{'}_{q,k}$ and $\gamma^{'}_{t,k}$ represent the partial derivatives of each γ with respect to a evaluated at a for queries of length k. Additionally, let $f(\beta_{\theta,k}, \beta_{z,k}) = \frac{2(\beta_{\theta,k}+2\beta_{z,k})(\beta_{\theta,k}+\beta_{z,k})}{\beta_{z,k}(3\beta_{\theta,k}+4\beta_{z,k})}$. For a given query Q with length k, when $\gamma_{q,k} + \gamma_{t,k}(\frac{\gamma^{'}_{q,k}}{\gamma^{'}_{t,k}}) > f(\beta_{\theta,k}, \beta_{z,k})$ CPC decreases, when $\gamma_{q,k} + \gamma_{t,k}(\frac{\gamma^{'}_{q,k}}{\gamma^{'}_{t,k}}) < f(\beta_{\theta,k}, \beta_{z,k})$, CPC increases, and when $\gamma_{q,k} + \gamma_{t,k}(\frac{\gamma^{'}_{q,k}}{\gamma^{'}_{t,k}}) = f(\beta_{\theta,k}, \beta_{z,k})$ there is no change to CPC.*

The proof is in Appendix E for a specific k. Starting with the left-hand side (LHS), we have two separate components: $\gamma_{q,k}$ and $\gamma_{t,k}(\frac{\gamma^{'}_{q,k}}{\gamma^{'}_{t,k}})$. Higher values of $\gamma_{q,k}$ evaluated at a lead to increasing LHS values and increase the potential for CPC to decrease. This is intuitive, as we know from Corollary 2 that CPC feels downward price pressure with increases in $\gamma_{q,k}$. The second component is the key differentiator that helps us understand when CPC will likely decrease or increase. The ratio $(\frac{\gamma^{'}_{q,k}}{\gamma^{'}_{t,k}})$ tells us there is an "arms race" between the rate of changes for context and topic signals in the γ parameter spaces. When there is a rapid change in the understanding of the context component *compared* to the topic component, the LHS has a greater chance of overcoming the threshold, and CPC decreases. But, if the topic signal rate of change is greater than or equal to context signal gains, or topic identification likelihood is low (small $\gamma_{t,k}$), the LHS is less likely to overcome the query-level threshold, and CPC is less likely to decrease with the introduction of a new algorithm. Even if the rate of change in identifying the context type is high for a given algorithm a, if the algorithm doesn't effectively identify the topic (low $\gamma_{t,k}$), then the "arms race" has little effect on the LHS.

Focusing on the right-hand side (RHS), the value $f(\beta_{\theta,k}, \beta_{z,k})$ for a given query Q acts as a threshold. As the importance of advertiser-query context alignment increases (higher $\beta_{z,k}$), the threshold function $f(\beta_\theta, \beta_z)$ decreases. Alternatively, as the importance of topical alignment increases ($\beta_{\theta,k}$), the threshold $f(\beta_\theta, \beta_z)$ increases.

CPC feels two forces: better matching due to better topic identification and better bid-

der differentiation due to better context identification. Topical understanding improves categorical matching, increasing the average number of bidders and prices for a given query. Simultaneously, better context signals help the platform differentiate the relevant buyers within the auction and accurately adjust relevancy scores. Contextual differentiation prioritizes advertisers who are more likely to receive a click and rewards them with lower prices.[23] This component softens the upward price increases caused by improved topical understanding and potentially enables CPC to decrease. Whether CPC decreases or increases with a new algorithm depends on the importance of contextual advertiser-query alignment and the relative gains made in identifying context. We visualize a numerical example at the end of Appendix E to illustrate the conditions for when prices increase or decrease for a particular algorithm across query lengths.

Mapping to Empirical Results We make the following assumptions to map our model to our empirical observations. Denote short (long) queries by subscript s (l). Since many short queries have little context, e.g., "running shoes", we assume that contextual alignment matters less for this type of query than a query like "best-running shoe stores near me in Los Angeles" where the search intent is more specific. Therefore, we assume that $\beta_{z,l} > \beta_{z,s}$ and that $\beta_{z,s} << \beta_{\theta,s}$. We will also assume that with the introduction of BERT, $\gamma^{'}_{q,l}$ is significantly larger than $\gamma^{'}_{t,l}$ and $\gamma^{'}_{t,s}$ is significantly larger than $\gamma^{'}_{q,s}$. Under these reasonable assumptions, long (short) queries can see CPC decrease (increase), demand will uniformly increase, and CTR will increase faster for long queries than short queries.

Summary Understanding context matters. Using a one-shot SPA with bidder relevancy scores, we propose that our empirical observations are driven by BERT generating better signals about the multi-dimensional type structure of language objects. Better context signals help differentiate bidders *within* auctions, while better topic signals help categorize and

[23]Talking to industry practitioners, the effects of the topic dimension in our model translates to what is called "ad sourcing" in industry, while the effects of the contextual dimension translate to what is called "ad ranking".

differentiate bidders across auctions. Only when advertiser-query contextual alignment matters to the query's CTR can within-auction differentiation lead to decreasing CPC post-new algorithm.

To better understand our model, consider a 2D X-Y plane going from -1 to 1 across all dimensions, where each quadrant represents a different topic. Advertisers are scattered throughout the plane. By learning the query topic, the seller knows which quadrant the query lives in, enabling the seller to pull all quadrant advertisers into the auction. However, the platform cannot differentiate the bidders within the topic quadrant without contextual information. When contextual information is known, the platform can accurately differentiate bidders by calculating distances (i.e., relevancy scores) between the query point and the advertisers in the quadrant. This differentiation can lead to lower CPC only when context signal gains are large, and the query's CTR heavily depends on contextual advertiser-query alignment.

Our model suggests that the importance of contextual alignment for CTR across query lengths dictates whether a new algorithm can cause prices to increase or decrease. Because short queries lack contextual signals within the query itself, they are doomed to increasingly competitive and costlier auctions. On the other hand, long query markets may see prices decline when a new algorithm is introduced due to the importance of the contextual present within the query and the ability to begin programmatically understanding it.

6.5 Limitations

Our model comes with limitations. First, we simplify the auction process by analyzing an SPA, not a GSPA or repeated auction format. Our model aims to understand how changes to an algorithm in a matchmaking auction market can affect the seller's information set and alter demand and prices. While our simplifying assumption limits our model's applicability to the real world, it lets us cleanly show how the information structure in language can potentially cause our empirical observations.

Second, we do not fully analyze potential market inefficiencies or the potential for market filtering. This is a byproduct of our assumptions that C is sufficiently large, that the platform knows advertiser types ahead of time, and that contextual and topical information are independent measurements. Future research can relax these assumptions and explore how better signals for the seller lead to market filtering and lower demand. (In the context of our alternative identification strategy, this extension would help the model map to the triple interaction terms estimated in Equation 3 for CS).

Our model assumes the platform (Google) drives our empirical observations. Given that we focus on a short, immediate time window, we think this is reasonable. However, it is possible that other mechanisms could drive our results. For example, advertisers could adjust their targeting behavior by entering the sponsored search market or altering their targeted keyword sets after BERT's introduction. Amaldoss et al. (2016) finds that improvements to broad match bidding algorithms can increase market entry and demand. Better interpretation algorithms could potentially have a similar effect. However, given our short time window, we do not think advertisers are quickly adjusting and driving our results. In display advertising, Alcobendas and Zeithammer (2021) find that it takes months for major advertisers to begin changing bidding strategies following a shift from a second to a first-price auction format. Event study results from Aridor et al. (2024) also find that advertisers appear to take several months to begin adjusting advertising budgets following shifts in data privacy regulations. If advertisers are slow to respond to such significant market changes, where the effects on profits are immediate and clear, it seems unlikely advertisers would rapidly respond to BERT, where the implications and direct market effects were not obvious.

Finally, our model does not consider potential interactions between the sponsored and organic search markets. We concede that this paper does not explore alternative explanations surrounding organic search. Future research can explore our model predictions and alternative theories surrounding organic search.

7 Conclusion

Advancements in NLP research have significantly improved search engine programmatic interpretation abilities. Yet, little is known about how changes to query interpretation algorithms potentially affect sponsored search markets. To help advertisers prepare for future algorithm changes, we study how Google's October 2019 rollout of BERT affected auction competition and prices. We find significant market changes: the average number of bidders uniformly increases across queries, while CPC heterogeneously adjusts across query lengths. These findings are supported by two identification strategies and several robustness checks.

We develop a theoretical auction model to conceptualize the market and extrapolate our empirical findings. In our model, a seller uses an algorithm to generate signals about query types. By imposing a multi-dimensional type structure (topic and context), we show how our empirical results could arise: topic information acquisition helps organize bidders and increase auction density, while contextual information helps estimate more precise relevancy scores and differentiate bidders within auctions. When advertiser-query context alignment matters to the query's CTR, and gains to contextual understanding are significant, prices may drop when a new algorithm is deployed.

Advertisers are frequently left in the dark when search engines like Google release significant platform updates. Our paper provides a rigorous empirical and theoretical study to help practitioners and academics better understand how changes to query interpretation algorithms impact sponsored search markets. It also contributes to the growing literature on the economic consequences of AI and LLMs in the search engine ecosystem.

References

Alcobendas, Miguel and Robert Zeithammer, "Adjustment of Bidding Strategies After a Switch to First-Price Rules," *Available at SSRN 4036006*, 2021.

Amaldoss, Wilfred, Kinshuk Jerath, and Amin Sayedi, "Keyword management costs and "broad match" in sponsored search advertising," *Marketing Science*, 2016, *35* (2), 259–274.

_ , Preyas S Desai, and Woochoel Shin, "Keyword search advertising and first-page bid estimates: A strategic analysis," *Management Science*, 2015, *61* (3), 507–519.

Anderson, Chris, *The long tail: Why the future of business is selling less of more*, Hachette UK, 2006.

Aridor, Guy, Yeon-Koo Che, Brett Hollenbeck, Maximilian Kaiser, and Daniel McCarthy, "Evaluating The Impact of Privacy Regulation on E-Commerce Firms: Evidence from Apples App Tracking Transparency," 2024.

Athey, Susan and Joshua S Gans, "The impact of targeting technology on advertising markets and media competition," *American Economic Review*, 2010, *100* (2), 608–613.

Bergemann, Dirk, Tibor Heumann, and Stephen Morris, "Selling Impressions: Efficiency vs. Competition," 2021.

Berman, Ron and Zsolt Katona, "The role of search engine optimization in search marketing," *Marketing Science*, 2013, *32* (4), 644–651.

Blake, Thomas, Chris Nosko, and Steven Tadelis, "Consumer heterogeneity and paid search effectiveness: A large-scale field experiment," *Econometrica*, 2015, *83* (1), 155–174.

Board, Simon, "Revealing information in auctions: the allocation effect," *Economic Theory*, 2009, *38* (1), 125–135.

Bollinger, Bryan, Eli Liebman, David Hammond, Erin Hobin, and Jocelyn Sacco, "Educational campaigns for product labels: Evidence from on-shelf nutritional labeling," *Journal of Marketing Research*, 2022, *59* (1), 153–172.

Broder, Andrei, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel, and Jeffrey Yuan, "Online expansion of rare queries for sponsored search," in "Proceedings of the 18th international conference on World wide web" 2009, pp. 511–520.

Burtch, Gordon, Dokyun Lee, and Zhichen Chen, "The consequences of generative ai for ugc and online community engagement," *Available at SSRN 4521754*, 2023.

Chandra, Ambarish, "Targeted advertising: The role of subscriber characteristics in media markets," *The Journal of Industrial Economics*, 2009, *57* (1), 58–84.

Coopmans, Cas W, Karthikeya Kaushik, and Andrea E Martin, "Hierarchical structure in language and action: A formal comparison.," *Psychological Review*, 2023, *130* (4), 935.

Cowgill, Bo and Cosmina Dorobantu, "Competition and Specificity in Market Design: Evidence from Geotargeted Advertising," *Columbia Business School Research Paper*, 2020, (101).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Edelman, Benjamin and Zhenyu Lai, "Design of search engine services: Channel interdependence in search engine results," *Journal of Marketing Research*, 2016, *53* (6), 881–900.

_ , Michael Ostrovsky, and Michael Schwarz, "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," *American economic review*, 2007, *97* (1), 242–259.

Eichenbaum, Martin S, Miguel Godinho de Matos, Francisco Lima, Sergio Rebelo, and Mathias Trabandt, "Expectations, Infections, and Economic Activity," Technical Report, National Bureau of Economic Research 2020.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130*, 2023.

Ganuza, Juan-José, "Ignorance promotes competition: an auction model with endogenous private valuations," *RAND Journal of Economics*, 2004, pp. 583–598.

Ghose, Anindya and Sha Yang, "An empirical analysis of search engine advertising: Sponsored search in electronic markets," *Management science*, 2009, *55* (10), 1605–1622.

Gleason, Jeffrey, Desheng Hu, Ronald E Robertson, and Christo Wilson, "Google the gatekeeper: How search components affect clicks and attention," in "Proceedings of the International AAAI Conference on Web and Social Media," Vol. 17 2023, pp. 245–256.

Goldberg, Yoav, "Assessing BERT's syntactic abilities," *arXiv preprint arXiv:1901.05287*, 2019.

Goli, Ali and Amandeep Singh, "Frontiers: Can Large Language Models Capture Human Preferences?," *Marketing Science*, 2024.

Hewitt, John and Christopher D Manning, "A structural probe for finding syntax in word representations," in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)" 2019, pp. 4129–4138.

Jäger, Gerhard and James Rogers, "Formal language theory: refining the Chomsky hierarchy," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2012, *367* (1598), 1956–1970.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah, "What does BERT learn about the structure of language?," in "ACL 2019-57th Annual Meeting of the Association for Computational Linguistics" 2019.

Kushwaha, Amit Kumar and Arpan Kumar Kar, "MarkBot–a language model-driven chatbot for interactive marketing in post-modern world," *Information systems frontiers*, 2021, pp. 1–18.

Liaukonyte, Jura, Anna Tuchman, and Xinrong Zhu, "Spilling the Beans on Political Consumerism: Do Social Media Boycotts and Buycotts Translate to Real Sales Impact?," *Available at SSRN 4006546*, 2022.

Lin, Yongjie, Yi Chern Tan, and Robert Frank, "Open sesame: Getting inside BERT's linguistic knowledge," *arXiv preprint arXiv:1906.01698*, 2019.

Liu, Nelson F, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith, "Linguistic knowledge and transferability of contextual representations," *arXiv preprint arXiv:1903.08855*, 2019.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

Miyagawa, Shigeru, Robert C Berwick, and Kazuo Okanoya, "The emergence of hierarchical structure in human language," *Frontiers in psychology*, 2013, *4*, 40804.

Mnih, Andriy and Geoffrey E Hinton, "A scalable hierarchical distributed language model," *Advances in neural information processing systems*, 2008, *21.*

Reisenbichler, Martin, Thomas Reutterer, David A Schweidel, and Daniel Dan, "Frontiers: Supporting content marketing with natural language generation," *Marketing Science*, 2022, *41* (3), 441–452.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, 2021, *8*, 842–866.

Rutz, Oliver J and Randolph E Bucklin, "From generic to branded: A model of spillover in paid search advertising," *Journal of Marketing Research*, 2011, *48* (1), 87–102.

Simonov, Andrey, Chris Nosko, and Justin M Rao, "Competition and crowd-out for brand keywords in sponsored search," *Marketing Science*, 2018, *37* (2), 200–215.

Tadelis, Steven and Florian Zettelmeyer, "Information disclosure as a matching mechanism: Theory and evidence from a field experiment," *American Economic Review*, 2015, *105* (2), 886–905.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick, "BERT rediscovers the classical NLP pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

_ , Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das et al., "What do you learn from context? probing for sentence structure in contextualized word representations," *arXiv preprint arXiv:1905.06316*, 2019.

Varian, Hal R, "Position auctions," *international Journal of industrial Organization*, 2007, *25* (6), 1163–1178.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017, *30.*

Yang, Sha and Anindya Ghose, "Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?," *Marketing science*, 2010, *29* (4), 602–623.

Zarifhonarvar, Ali, "Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence," *Journal of Electronic Business & Digital Economics*, 2023.

Appendix

A Sampling Procedure

Below is the list of search topics we asked for in our Amazon MTurk survey.

Insurance, Travel, Home Renovation, Cars, Restaurants, How-to's, Things To Do, Company Research, Product Research, Financial Resources, Financial Products, News, Politics, Animals, Books, Movies, Shows, Clothes, Electronics, Pest Control, Birthday Gifts, Kitchen Shopping, Moving Services, Social Media, Entertainment, Cooking, Transportation, Home Decor, Health, Medicine, Cosmetics.

Each participant was randomly asked about only five of these categories.

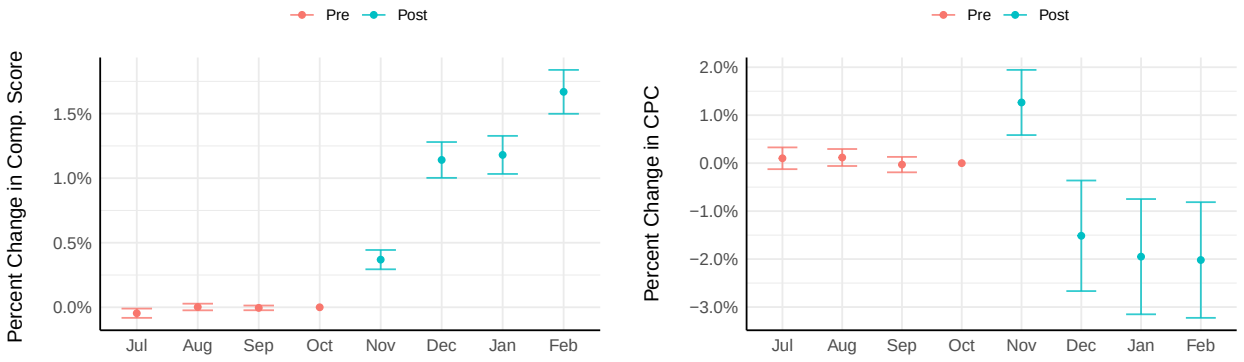
B DD: Additional Control Groups

Our primary DD analysis uses query data from 2018 to 2020, with July 2018 to February 2019 as control units for July 2019 to February 2020. The downside of using only one year-month set (2018-2019) as a control window is we don't know whether it accurately represents the behavior of the auction market during those months. (This is equivalent to worrying about using only two states in a traditional state-level DD). As a robustness check, we re-estimate Equations 1 and 2 using 2016-2017 and 2017-2018 July-February months as additional control groups.

We are primarily interested in replicating our results for competitive queries before BERT. To identify competitive queries with this additional data, we take the mode Competition Score across all July-October months and years (2016-2019, generating 16 observations per query) and deem a query non-competitive if the mode is 0. In Figure 5, we present event analysis market changes for $log(CS)$ and $log(CPC)$ for queries deemed competitive before BERT.

In Table 15, we present DD estimates for all competitive queries. Similar to Table 4,

Figure 5: $log(CS)$ and $log(CPC)$ for the sample of competitive queries before BERT and using all control years. Error bars represent 95% confidence intervals.



we find CS increases by roughly 1%. Interestingly, we see a statistically significant negative change in CPC, as is visually seen in the right panel of Figure 5, consistent with our alternative identification.

Table 15: The effect of BERT on $log(CS)$ and $log(CPC)$.

	(1)	(2)
	Log(CS)	Log(CPC)
Post × Treated	0.011***	-0.011**
	(0.001)	(0.005)
Observations	275,564	275,564
R^2	0.938	0.682

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data from 2016 to 2020.

Finally, we repeat our regressions by query length for $log(CS)$ and $log(CPC)$. Table 16 presents CS estimates by query length using all years available in our data. Table 17 shows changes to $log(CPC)$. The results are directionally consistent and similar in magnitude to

those estimated in Tables 5 and 6.

Table 16: The effect of BERT on $log(CS)$ by query length.

	(1)	(2)	(3)
	Short	Medium	Long
Post × Treated	0.012***	0.009***	0.017***
	(0.001)	(0.001)	(0.001)
Observations	83,765	154,872	36,927
R^2	0.917	0.956	0.905

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data from 2016 to 2020.

Table 17: The effect of BERT on $log(CPC)$ by query length.

	(1)	(2)	(3)
	Short	Medium	Long
Post × Treated	0.032***	−0.021***	−0.058***
	(0.009)	(0.006)	(0.013)
Observations	83,765	154,872	36,927
R^2	0.629	0.702	0.627

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: Estimated with competitive queries. Regressions include year-group, month, and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data from 2016 to 2020.

While these results do not validate our identification strategy, the fact that using additional control years leads to broadly similar results suggests we are reasonably controlling for within-unit seasonal patterns and that the 2018-2019 auction market outcomes are fairly representative of the average Google market in the absence of BERT.

C DD: Non-Logged Competition Score

We reported logged CS scores in the primary results section to present changes in relative (percentage) terms. In Table 18, we present non-logged CS estimates and find consistent results. When compared to the average CS in the months before BERT's introduction (July-September 2019), these estimates translate to a roughly 5-6% increase in CS.

Table 18: The Effect of BERT on CS for all queries and those that were competitive BERT BERT.

	(1)	(2)
	All Queries	Competitive Only
Post×Treated	0.014^{***}	0.016^{***}
	(0.001)	(0.001)
Observations	191,161	160,687
R^2	0.967	0.964
Note:		*p<0.1; **p<0.05; ***p<0.01

D Alternative Specification: Additional Robustness Checks

D.1 Continuous Linguistic Complexity Results

Our primary LC measure is binary, allowing for easy interpretation and presentation of results. We also model using a continuous form of LC. We create a binary indicator for whether a query's syntax tree is greater than the median and a binary indicator for whether it's POS count is greater than the median. We then take the sum of these two observations, meaning a query's continuous LC measure can have a value of 0, 1, or 2. 0 means both the POS count and syntax tree depth were below the dataset median, while 2 means both were above the median. Using this continuous form of LC, we present results for CS in Table 19 and results for CPC in Table 20.

Table 19: $log(CS)$ changes with Cosine and Continuous Linguistic Complexity.

	(1)	(2)	(3)
	75th	50th	25th
Post × Continuous LC	0.009***	0.007***	0.007***
	(0.002)	(0.002)	(0.002)
Post × Log(Cosine)	0.001***	0.001**	0.001
	(0.0005)	(0.001)	(0.001)
Post × Continuous LC × Log(Cosine)	−0.002***	−0.003***	−0.004***
	(0.001)	(0.001)	(0.001)
Observations	63,466	63,466	63,466
R^2	0.980	0.979	0.979

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

Table 20: $log(CPC)$ changes with Cosine and Continuous Linguistic Complexity.

	(1)	(2)	(3)
	75th	50th	25th
Post × Continuous LC	−0.037**	−0.040***	−0.042***
	(0.016)	(0.015)	(0.016)
Post × Log(Cosine)	0.005	0.006*	0.006
	(0.003)	(0.004)	(0.005)
Post × Continuous LC × Log(Cosine)	0.001	0.003	0.003
	(0.006)	(0.008)	(0.011)
Observations	63,466	63,466	63,466
R^2	0.795	0.795	0.795

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

D.2 Alternative Cosine Measure

We can standardize the cosine similarity scores within each vector space and compare differences across the vector spaces within a given query. Therefore, for a given query we can estimate where it falls in the BERT cosine similarity distribution and the D2V cosine similarity distribution and calculate the difference (change) when moving from D2V to BERT. This lets us analyze queries that, when moving from the D2V vector space to the BERT vector space, see a comparative increase (decrease) in the similarity between the query and all the other queries in our dataset. This measure can identify which queries are potentially more or less related to other queries and topics. We are theoretically interested in this measurement because we hypothesize that increased relatedness leads to an increase in the number of advertisers deemed "related", leading to an increase in bidders (CS). The downside of measuring these changes is that it doesn't capture shifts in the related queries like our first measurement. The average cosine similarity difference between the two normalized

counts is 0. At 0, there was no relative shift in the cosine similarity. Our main focus is to understand what happens when increasing cosine relatedness. To operationalize this, we create a binary variable called Normalized Cosine Diff that takes on the value of 1 if the query's relative change is greater than the 75th percentile (0.8532), 0 otherwise. This creates a binary variable that lets us compare queries that see a large increase in cosine similarity to those that see comparatively little change or a reduction in cosine similarity densities.

Table 21: $log(CS)$ and $log(CPC)$ changes with Cosine Quartile and Linguistic Complexity.

	(1)	(2)
	Log(CS)	Log(CPC)
Post × Linguistic Complexity	0.003*	−0.029**
	(0.002)	(0.013)
Post × Top Quartile Cosine Score	0.005*	0.031*
	(0.003)	(0.016)
Post × Linguistic Complexity × Top Quartile Cosine Score	−0.008**	0.004
	(0.003)	(0.024)
Observations	63,466	63,466
R^2	0.979	0.795

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

We also estimate Equation 3 using the previously described quartile indicator variable as an alternative cosine measure. Recall that the Top Quartile Cosine Score takes on a value of 1 if the relative change in the difference between the normalized BERT and normalized D2V cosine ranks is greater than the 75th percentile of differences, 0 otherwise. In other words, the Top Quartile Cosine Score takes on a value of 1 when a query sees a large increase in the relative similarity to it and other queries when shifting from the D2V vector space to the BERT vector space. This gives us a different way to measure increases or decreases in "relatedness" to other queries. We find directionally consistent results with those found

in the models shown in Tables 9 and 11. Under this model specification, average predicted values at the aggregate level find CS increasing by roughly 0.5%, and CPC decreases by roughly 0.9%. Short queries see CS increase by roughly 0.2% and CPC increase by roughly 1.5%, medium queries see CS increase by 0.6% and CPC decrease by roughly 1.7%. Finally, long queries see CS increase by roughly 0.6% while CPC decreases by roughly 2.7%. Again, all estimates are directionally consistent with our primary DiD specification but are more conservative.

Table 22: $log(CS)$ and $log(CPC)$ changes with Cosine Quartile and Continuous Linguistic Complexity.

	(1)	(2)
	Log(CS)	Log(CPC)
Post × Continuous LC	0.006***	−0.040***
	(0.002)	(0.014)
Post × Top Quartile Log(Cosine) Score	0.005**	0.025
	(0.002)	(0.015)
Post × Continuous LC × Top Quartile Log(Cosine) Score	−0.011***	0.020
	(0.004)	(0.030)
Observations	63,466	63,466
R^2	0.979	0.795

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

D.3 Event Study Parameter Analysis

For our alternative identification to hold, we need to assume that our measurements interact with BERT, that no changes before BERT correlate with these measurements, and that queries across measurements are comparable. To test this assumption, we re-estimate Equation 3 but replace *Post* with a monthly dummy variable, where October is the baseline. If

our measurements and assumptions are valid, we should see largely null effects in the months before BERT's introduction.

In Tables 23, we present event study parameter estimates for $log(CS)$ for our 75th Cosine threshold results. Similarly, in Table 24, we present the same event study results for $log(CPC)$. We generally find null results in the pre-period and significant estimates in the post-periods, supporting our identification assumptions.

Table 23: $log(CS)$ event study parameter estimates for Cosine 75th threshold.

Month	LC	Cosine	LC × Cosine
July	0.0001	−0.00002	−0.0001
	0.0003	0.00005	0.0001
August	−0.001**	−0.0001**	0.0002
	0.0002	0.00004	0.0001
September	−0.0002	−0.00005	0.0001
	0.0001	0.00003	0.00005
November	0.003*	0.0004	−0.001**
	0.001	0.0003	0.0004
December	0.005**	0.001*	−0.002*
	0.002	0.001	0.001
January	0.006**	0.001**	−0.002*
	0.002	0.001	0.001
February	0.002	0.002**	−0.002*
	0.003	0.001	0.001

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard Errors clustered at the query level are reported in parentheses.

Table 24: $log(CPC)$ event study parameter estimates for Cosine 75th threshold.

Month	LC	Cosine	LC × Cosine
July	0.0003	0.0003	−0.001
	(0.002)	(0.0003)	(0.001)
August	0.002	0.0004	−0.001
	(0.002)	(0.0003)	(0.001)
September	−0.001	−0.0003	−0.0002
	(0.001)	(0.0002)	(0.0005)
November	−0.019*	−0.001	0.003
	(0.011)	(0.002)	(0.004)
December	−0.029	0.008*	−0.002
	(0.018)	(0.004)	(0.006)
January	−0.034*	0.008**	0.00002
	(0.018)	(0.004)	(0.006)
February	−0.024	0.008**	−0.002
	(0.018)	(0.004)	(0.006)

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.
Note: All regressions include year-month and query fixed effects. Standard Errors clustered at the query level are reported in parentheses.

D.4 Controlling for Search Volume

One concern with our alternative identification strategy is that consumers could change their search behavior, potentially confounding the effect of our linguistic measurements and DVs. To control this, we re-estimate our alternative identification regression models controlling for the log of the query's search volume (SV). In Table 25, we present changes to CS, controlling for SV.

Similarly, in Table 26, we present percent changes to CPC controlling for SV. We find that estimates remain largely consistent with our primary specification.

Table 25: $log(CS)$ changes with Cosine, Linguistic Complexity, and SV.

	(1)	(2)	(3)
	75th	50th	25th
Log(SV)	0.0062***	0.0062***	0.0062***
	(0.0023)	(0.0023)	(0.0023)
Post × Linguistic Complexity	0.0041**	0.0035*	0.0035*
	(0.0020)	(0.0018)	(0.0020)
Post × Log(Cosine)	0.0012**	0.0009*	0.0007
	(0.0005)	(0.0006)	(0.0008)
Post × Linguistic Complexity × Log(Cosine)	−0.0015**	−0.0020**	−0.0030**
	(0.0007)	(0.0008)	(0.0012)
Observations	63,474	63,474	63,474
R^2	0.9786	0.9786	0.9786
Note:			*p<0.1; **p<0.05; ***p<0.01

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.

Note: All regressions include year-month and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

Table 26: $log(CS)$ changes with Cosine, Linguistic Complexity, and SV.

	(1)	(2)	(3)
	75th	50th	25th
Log(SV)	0.0129	0.0129	0.0128
	(0.0130)	(0.0130)	(0.0130)
Post × Linguistic Complexity	−0.0267*	−0.0286**	−0.0304**
	(0.0146)	(0.0138)	(0.0145)
Post × Log(Cosine)	0.0056*	0.0065*	0.0075
	(0.0032)	(0.0037)	(0.0052)
Post × Linguistic Complexity × Log(Cosine)	0.00002	0.0010	0.0006
	(0.0051)	(0.0064)	(0.0092)
Observations	63,474	63,474	63,474
R^2	0.7952	0.7952	0.7952
Note:			*p<0.1; **p<0.05; ***p<0.01

Significance Levels: *p<0.1; **p<0.05; ***p<0.01.

Note: All regressions include year-month and query fixed effects. Standard Errors clustered at the query level are reported in parentheses. Data is from July 2019 to February 2020.

E Model Proofs

Platform Reputation and Advertiser Selection The platform wishes to maximize Equation 7.

$$\pi = I_{\{Click\}}\frac{\tilde{Adr}}{r_w} - I_{\{\theta_w \neq t\}}C \tag{7}$$

While we incorporate the reputation cost C into our model, we focus our analysis on price shifts, not platform market inefficiency. To prove further results of the model, we will assume C is sufficiently large such that the platform does not allocate any advertiser to the auction opportunity when it does not know the query topic. We do not focus on this feature of the model but note the flexibility of the model for future research.

For those interested, we briefly describe how different thresholds of C affect auction choices. If $C \geq \frac{3}{5}(\beta_\theta + \beta_z)$, then the platform is unwilling to run an auction if the topic and context are not known (i.e., $\hat{q} = \hat{t} = \frac{1}{2}$). If the context is known but the topic is not known, the platform is unwilling to run an auction with only the bidders known to match the context if $C \geq \frac{2}{3}(\frac{1}{2}\beta_\theta + \beta_z)$. Finally, the platform is unwilling to run an auction with all bidders if the context is known and the topic is not known if

$$C \geq \frac{1}{12(\beta_\theta + \beta_z)^3}(\frac{7(\beta_z\beta_\theta^3)}{4} + \frac{7\beta_\theta^4}{8} + 8\beta_\theta(\frac{1}{2}\beta_\theta + \beta_z)^3 + 16\beta_z(\frac{1}{2}\beta_\theta + \beta_z)^3 - \frac{9\beta_\theta^2(\frac{1}{2}\beta_\theta + \beta_z)}{4} - 6\beta_z\beta_\theta^2(\frac{1}{2}\beta_\theta + \beta_z))$$

These values are calculated by comparing the inequality $C \geq E[CPC]$ across conditions. Note that the highest possible revenue the platform could ever receive from a particular auction is 1. (The topic is known, but the context is not, meaning relevancy scores disappear, two bidders bid exactly 1, and the tie randomly selects the better matching advertiser, leading to a CTR of 1.) If we assume $C > 1$, running an auction with advertisers is never profitable

when the topic is unknown. We restrict ourselves to only this scenario for all subsequent analyses.

E.1 Query Analysis

We will assume C is sufficiently large that the platform is unwilling to run an auction when the topic is unknown. We will also drop the subscript k to indicate query lengths. The model allows k to differentiate parameters across query lengths.

Number of Bidders Assuming C is sufficiently large (i.e., $C > 1$), the platform is never willing to allocate advertisers to the auction when the topic is unknown for fear of potentially showing an irrelevant advertisement and suffering the negative cost C. When the topic is known, the platform will allocate advertisers of that type to the auction. The topic is known with probability γ_t. When the topic is understood, 2 bidders of the particular topic are added to the auction. Therefore, the average number of bidders for a given query is $2\gamma_t(a)$.

Bidding Strategies Advertiser private valuations for a click are drawn $v_i \sim U[0,1]$. Conditional on being allocated to an auction, their objective is to maximize $E[v - \frac{\tilde{b}\tilde{r}}{r} | b^* r \geq \tilde{b}\tilde{r}]$, where b^* is the optimal bid, r is their relevancy score for the given auction, $\tilde{b}$ is the second highest bid value, and $\tilde{r}$ is the second highest relevancy score. ($\tilde{b}\tilde{r}$ is the second highest ad rank). Since they know they will only be allocated to relevant auctions, they aim to maximize the following:

$$\gamma_q \int_0^{\frac{b^* r}{\tilde{r}}} (v - \frac{\tilde{Adr}}{r}) f(\tilde{b}) d\tilde{b} + (1 - \gamma_q) \int_0^{b^*} (v - \tilde{b}) f(\tilde{b}) d\tilde{b}$$

Taking the derivative w.r.t b^* and rearranging, we find

$$(v - b^*)[\gamma_q f(\frac{b^* r}{\tilde{r}})) + (1 - \gamma_q) f(b^*)] = 0$$

We arrive to $b^* = v$ $\square$

Average CPC Recall that average CPC is equal to $Pr(Click)E[\frac{\tilde{Adr}}{r_w}]$. A bid is not paid unless there is a click. Bid costs are the second-highest Ad Rank normalized by the winning bidder's relevancy score.

When the platform knows the topic but not the context, all advertisers related to the topic are allocated to the auction, each with the same relevancy score of $\frac{\beta_\theta}{\beta_\theta+\beta_z}$. This simplifies the bids into a two-player SPA where valuations (bids) are drawn from $U[0,1]$. In expectation, the second highest bid will be $\frac{1}{3}$. CTR depends on which type of advertiser wins. With probability $\frac{1}{2}$ it is the advertiser with the right context and topic, and with probability $\frac{1}{2}$, it is the advertiser with the right topic but not the right context. This means expected CTR is $\frac{\beta_z}{2}+\beta_\theta$. Therefore, in expectation, expected CPC is $\frac{1}{3}(\frac{\beta_z}{2}+\beta_\theta)$. This occurs with probability $\gamma_t(1-\gamma_q)$.

When the platform knows the context and topic, all topical advertisers are allocated to the auction, but each type has a different relevancy score. Consider a query with context type H. The high context advertiser will have a relevancy score of $r_H = \frac{\beta_\theta+\beta_z}{\beta_\theta+\beta_z} = 1$ while the low context advertiser will have a relevancy score of $r_L = \frac{\beta_\theta}{\beta_\theta+\beta_z}$. Since advertiser bids are drawn from $U[0,1]$, the Ad Rank distribution for the type L advertiser is scaled by r_L to $U[0,r_L]$. We now calculate the expected CPC for this scenario.

$$(\beta_\theta+\beta_z)[(1-\frac{r_L}{r_H})\frac{r_L}{2r_H}]+(\beta_\theta+\beta_z)[\frac{r_L}{2r_H}\frac{r_L}{3r_H}]+\beta_\theta(\frac{1}{6})$$

Before simplifying, we describe how it is set up. With probability $1-\frac{r_L}{r_H}$, the H context advertiser has an Ad Rank that wins with certainty, leaving them to pay the expected $\frac{r_L}{2}*\frac{1}{r_H}$. The expected CTR when H wins is $(\beta_\theta+\beta_z)$. With probability $\frac{r_L}{r_H}$, the Ad Rank of the H context advertiser falls within the possible values of the L context advertiser. In this case, there is an equal chance for each bidder to win. Half the time, the winning bidder is the L context bidder; the other half is the H context bidder. When the H bidder wins, expected CTR is again $(\beta_\theta+\beta_z)$, when the L bidder wins its β_θ. When either bidder wins, the expected Ad Rank of the other bidder in this condition is $\frac{r_L}{3}$. When the H (L) bidder

wins, this value is scaled by r_H (r_L). Define $A = \frac{3r_L r_H - 2r_L^2}{6r_H^2}$. The expected CPC is

$$A(\beta_\theta + \beta_z) + \frac{\beta_\theta}{6}$$

Since $r_H = 1$ and $r_L = \frac{\beta_\theta}{\beta_\theta + \beta_z}$ in this setting, we can further simplify. The final expression is

$$\frac{\beta_\theta}{6}[3 - \frac{\beta_\theta}{\beta_\theta + \beta_z}]$$

We now take a weighted average of the various expected CPC values across cases to calculate the expected CPC for the query. Define $D = \frac{\beta_\theta}{6}[3 - \frac{\beta_\theta}{\beta_\theta + \beta_z}]$ and $M = \frac{1}{3}(\beta_\theta + 2\beta_z)$. The expected CPC for a q_H query is

$$E[CPC] = \gamma_t \gamma_q D + \gamma_t(1 - \gamma_q)M = \gamma_t \gamma_q L + \gamma_t M$$

Where $L = D - M$. L is equal to $\frac{\beta_\theta - 4\beta_z}{6} - \frac{\beta_\theta^2}{6(\beta_\theta + \beta_z)}$. L is less than 0 when $3\beta_\theta + 4\beta_z > 0$. Since β_z, $\beta_\theta > 0$, L is always negative.

Our focus is on understanding what happens when an algorithm change occurs. We first analyze changes to $E[CPC]$ w.r.t γ_q and γ_t and then look at how prices adjust w.r.t a, which will affect *both* γ_q and γ_t. First, the derivative w.r.t γ_t.

$$\frac{\partial}{\partial \gamma_t} \gamma_t \gamma_q L + \gamma_t M$$

$$\gamma_q(D - M) + M$$

$$\gamma_q D + M(1 - \gamma_q)$$

Since $\gamma_q \leq 1$, $M > 0$, and $D > 0$, the derivative is always positive. This means CPC

increases with topical information acquisition. Alternatively, the derivative of $E[CPC]$ w.r.t γ_q is decreasing.

$$\frac{\partial}{\partial \gamma_q} \gamma_t \gamma_q L + \gamma_t M$$

$$\gamma_t L$$

Since $L < 0$, this is always negative. This is because γ_q helps the platform prioritize the high CTR advertiser and rewards them with lower prices. We now wish to analyze how $E[CPC]$ changes w.r.t a, which impacts both γ_q and γ_t.

$$\frac{\partial}{\partial a} \gamma_t(a) \gamma_q(a) L + \gamma_t(a) M$$

$$L[\gamma_t^{'} \gamma_q + \gamma_q^{'} \gamma_t] + \gamma_t^{'} M$$

Where $\gamma_t^{'} > 0$ and $\gamma_q^{'} > 0$ represent the derivatives of each function w.r.t a. We are interested in knowing when CPC increases or decreases.

$$L[\gamma_t^{'} \gamma_q + \gamma_q^{'} \gamma_t] + \gamma_t^{'} M < 0$$

$$L[\gamma_q + \gamma_q^{'} \frac{\gamma_t}{\gamma_t^{'}}] + M < 0$$

Which equates to

$$\gamma_q + \gamma_q^{'} \frac{\gamma_t}{\gamma_t^{'}} > -\frac{M}{L}$$

Recall that $L < 0$, so $\frac{M}{L}$ is negative.

$$-\frac{M}{L} = \frac{2(\beta_\theta + 2\beta_z)(\beta_\theta + \beta_z)}{\beta_z(3\beta_\theta + 4\beta_z)}$$

Define $f(\beta_\theta, \beta_z) = \frac{2(\beta_\theta+2\beta_z)(\beta_\theta+\beta_z)}{\beta_z(3\beta_\theta+4\beta_z)}$. When

$$\gamma_q + \gamma_t \frac{\gamma_q^{'}}{\gamma_t^{'}} > \frac{2(\beta_\theta + 2\beta_z)(\beta_\theta + \beta_z)}{\beta_z(3\beta_\theta + 4\beta_z)}$$

$$\gamma_q + \gamma_t \frac{\gamma_q^{'}}{\gamma_t^{'}} > f(\beta_\theta, \beta_z)$$

The CPC decreases when the inequality flips, increases when the inequality flips, and remains the same at equality.

This inequality gives us significant insight into how CPC can adjust to a new algorithm. We first analyze the right-hand side. Taking the derivative of $f(\beta_\theta, \beta_z)$ w.r.t β_z tells us how the right-hand side constraint changes with β_z

$$\frac{\partial f(\beta_\theta, \beta_z)}{\partial \beta_z} = \frac{2\beta_\theta(3\beta_z^2 - 2\beta_\theta\beta_z - 3\beta_\theta^2)}{\beta_z^2(\beta_z + 2\beta_\theta)^2}$$

Similarly, taking the derivative w.r.t β_θ tells us how the right-hand side constraint changes with β_θ.

$$\frac{\partial f(\beta_\theta, \beta_z)}{\partial \beta_\theta} = \frac{2(3\beta_\theta^2 + 2\beta_\theta\beta_z - 3\beta_z^2)}{\beta_z(\beta_z + 2\beta_\theta)^2}$$

These results give us several useful insights. First, as β_z changes, we see that the threshold declines if $3\beta_z^2 - 2\beta_\theta\beta_z - 3\beta_\theta^2 < 0$, otherwise it increases. (Denominator is always > 0). Similarly, as β_θ increases, the threshold declines if $3\beta_\theta^2 + 2\beta_\theta\beta_z - 3\beta_z^2 < 0$, otherwise it increases. If we impose the assumption that $\beta_\theta \geq \beta_z$, meaning that contextual alignment is not as meaningful or equivalent to topical alignment, then we find that the ratio is always decreasing with β_z and increasing with β_θ.

We now consider the left-hand side (LHS). First, a higher value of γ_q causes the LHS to increase, potentially allowing it to overcome the RHS and decrease CPC. The second term, $\gamma_t \frac{\gamma_q^{'}}{\gamma_t^{'}}$, captures the "race" going on between topic and information acquisition and their

effects on CPC. When changes to topical information acquisition at an algorithmic point a are large, there is a lower likelihood of this ratio holding. However, if the gains to contextual knowledge are large and the contextual knowledge of the algorithm a is high, then there is a greater likelihood the LHS overcomes the RHS, causing CPC to decrease.

Mapping these results to our empirical results, we would say that long queries have a large β_z term, decreasing the RHS. BERT significantly improves contextual information acquisition, increasing the LHS, leading to this inequality holding and CPC decreasing.

CTR Consider the case where $q = 1$ and $t = 1$. Results are symmetric across. CTR is 0 when no auction is run. This occurs with probability $1 - \gamma_t$. With probability $\gamma_t(1 - \gamma_q)$, CTR is $\frac{1}{2}(\beta_\theta + \beta_z) + \frac{1}{2}\beta_\theta = \beta_\theta + \frac{1}{2}\beta_z$. Finally, with probability $\gamma_t\gamma_q$, expected CTR is equal to $\frac{r_L}{r_H}(\frac{1}{2}\beta_\theta + \frac{1}{2}(\beta_\theta + \beta_z)) + (1 - \frac{r_L}{r_H})(\beta_\theta + \beta_z) = \beta_z(1 - \frac{r_L}{2r_H}) + \beta_\theta$. Under the assumption that $r_H = 1$, $r_L = \frac{\beta_\theta}{\beta_\theta+\beta_z}$, this simplifies to $\beta_z(1 - \frac{\beta_\theta}{2(\beta_z+\beta_\theta)}) + \beta_\theta$.

Let $M = \beta_z(1 - \frac{\beta_\theta}{2(\beta_z+\beta_\theta)}) + \beta_\theta$ and $D = \beta_\theta + \frac{1}{2}\beta_z$. Expected CTR for the query is then $\gamma_t\gamma_q M + \gamma_t(1-\gamma_q)D$. Define $L = M - D$, we have $L\gamma_t\gamma_q + \gamma_t D$. $L \geq 0$ as long as $1 \leq \frac{\beta_\theta}{\beta_\theta+\beta_z}$, which always holds. The derivative of the expected CTR w.r.t a is $L(\gamma_q\gamma_t^{'} + \gamma_q^{'}\gamma_t) + D\gamma_t^{'}$, where $\gamma_q^{'}$ and $\gamma_t^{'}$ are the derivatives w.r.t a, and is always positive.

Taking the cross partial derivative w.r.t β_z equals $\frac{1}{2}(\gamma_t^{'}+(1-(\frac{\beta_\theta}{\beta_\theta+\beta_z})^2)(\gamma_t^{'}\gamma_q+\gamma_q^{'}\gamma_t))$. Thus, CTR gains increase as β_z increases in importance. The rate of change in CTR may increase or decrease with changes to β_θ. The cross partial derivative of the CTR function w.r.t a and then β_θ is $\gamma_t^{'} - \frac{1}{2}(\frac{\beta_z}{\beta_z+\beta_\theta})^2[\gamma_t^{'}\gamma_q + \gamma_q^{'}\gamma_t]$. This can be negative if $\frac{2}{(\frac{\beta_z}{\beta_z+\beta_\theta})^2} < (\gamma_q + \gamma_t\frac{\gamma_q^{'}}{\gamma_t^{'}})$ evaluated at a particular a. In this case, CTR changes w.r.t a still increase, but at a decreasing rate as β_θ increases.

Numerical Example Assume γ_t and γ_q follow sigmoid functions. Let $\gamma_t(a) = \frac{e^{ga+h}}{1+e^{ga+h}}$ and $\gamma_q(a) = \frac{e^a}{1+e^a}$. The derivative of γ_t w.r.t a is $\frac{ge^{ga+h}}{(1+e^{ga+h})^2}$ and the derivative of γ_q w.r.t. a is $\frac{e^a}{(1+e^a)^2}$. Our primary derivative outcome of interest is $\gamma_q + \gamma_t\frac{\gamma_q^{'}}{\gamma_t^{'}}$. Using those sigmoid functions, this evaluates to $\frac{e^a(2+e^a+e^{ga+h})}{(1+e^a)^2}$. Consider a scenario where, for long queries, $\beta_\theta = \frac{1}{2}$,

$\beta_z = \frac{1}{4}$, $g = 0.8$, and $h = 1.5$ and for short queries, $\beta_\theta = \frac{1}{2}$ and $\beta_z = \frac{1}{8}$. Assume also that BERT can be represented by $a = 1$. In Figure 6, we visualize the conditions for which prices will increase or decrease for short (left) vs. long (right) queries. We present the true data generating sigmoid curves for γ_t and γ_q as well as the numerical results for $\gamma_q + \gamma_t \frac{\gamma_q'}{\gamma_t}$ (called LHS). Additionally, we visualize the threshold function $f(\beta_\theta, \beta_z)$ for the different query types (red line). The dotted line represents BERT. Prices will decline when the blue line exceeds the red horizontal line (R1 region).

Figure 6: Visualization of price derivatives for long and short queries. Short query results are shown on the left, and long query results are on the right. The orange and green lines indicate the true underlying data generating curves for γ_q and γ_t. The blue line (LHS) represents $\gamma_q + \gamma_t \frac{\gamma_q'}{\gamma_t}$. The red dashed line indicates the threshold function $f(\beta_\theta, \beta_z)$. R1 indicates the region in which new algorithms will cause prices to decline for longer queries.