SAGE: Semantic Ambiguity Gate

Nihir Chadderwala*, Subrat Das, Chaitanya Vejendla,
Tasio Guevara, Somdeb Bhattacharjee, Atul Chaudhari

Amazon Web Services

Abstract

Large language model (LLM) agents deployed in healthcare and life sciences (HCLS) routinely receive queries that are semantically ambiguous—the same terms carry different meanings across clinical, regulatory, pharmacovigilance, data-standards, and research domains. Existing approaches address ambiguity post-hoc through output filtering or retrieval augmentation, but do not quantify it *before* the model responds. We introduce SAGE (Semantic Ambiguity Gate for Agentic Ecosystems), a client-side, pre-flight analysis framework that computes a semantic tension score (0–100) for each query using two parallel paths: (1) entity resolution with contextual domain inference and domain-distance-aware penalties against a curated HCLS concept registry, and (2) a context-aware NLP decision tree evaluating query structure. An adaptive weighting mechanism adjusts the relative influence of each path based on analysis conditions. The tension score is injected into the LLM prompt to control response behavior—high-tension queries trigger clarifying questions derived from concept analysis rather than model speculation. On a dataset of 30 real-world queries adapted from biomedical QA benchmarks (BioASQ [Tsatsaronis et al., 2015], PubMedQA [Jin et al., 2019]), clinical trial registries, and FDA guidance language, SAGE achieves 63.3% classification accuracy across three ambiguity levels and 100% domain inference accuracy, outperforming both entity-resolution-only (43.3%) and NLP-only (60.0%) baselines.

1 Introduction

LLM-powered agents are increasingly deployed across HCLS workflows—from clinical trial management and pharmacovigilance to regulatory submissions and biomedical research. These agents receive natural-language queries from diverse stakeholders: clinicians, data scientists, regulatory affairs specialists, and safety officers. A query such as *"Analyze patient adverse events"* is syntactically clear but semantically ambiguous: it could refer to protocol-defined treatment-emergent adverse events (clinical), MedDRA-coded safety records (data standards), signal detection in post-market surveillance (pharmacovigilance), or compliance reporting under ICH-E2A (regulatory).

Without a mechanism to detect this ambiguity *before* the model generates a response, the agent will select an interpretation—often confidently—without disclosing the alternatives it discarded. In regulated industries where precision directly impacts patient safety and regulatory standing, this behavior is unacceptable.

We propose SAGE (Semantic Ambiguity Gate for Agentic Ecosystems), a framework that:

1. Computes a *semantic tension score* entirely client-side, before any LLM call, using entity resolution and NLP structural analysis with *adaptive weighting*.

2. Uses *contextual domain inference* to propagate disambiguation signals from anchor terms and user-provided context layers to ambiguous entities, with *entity-count-aware* signal scaling.

3. Introduces a *domain similarity matrix* that distinguishes cross-domain queries (spanning semantically close domains) from genuinely

*Lead author.

ambiguous queries (spanning distant domains).

4. Employs *context-aware verb evaluation* that considers the specificity of a verb's direct object, not just the verb itself.

5. Derives *missing-context indicators and clarifying questions* directly from concept analysis, eliminating LLM-generated speculation about what information is needed.

6. Injects the score and concept-derived insights into the prompt to *control LLM response behavior* across three modes: clarify, partial answer, or full answer.

2 Related Work

Uncertainty quantification in LLMs. Prior work estimates model uncertainty through token-level entropy [Kadavath et al., 2022], semantic clustering of sampled outputs [Kuhn et al., 2023], or calibration techniques [Guo et al., 2017]. These approaches operate *after* generation and measure model confidence rather than input ambiguity. SAGE is complementary: it measures query-level semantic ambiguity before generation begins.

Retrieval-augmented generation (RAG). RAG systems retrieve relevant documents to ground LLM responses [Lewis et al., 2020]. While RAG reduces hallucination by providing context, it does not assess whether the query itself is sufficiently specified to select the correct documents. SAGE could serve as a pre-retrieval filter, ensuring queries are disambiguated before retrieval is attempted.

Prompt engineering and guardrails. Frameworks such as NeMo Guardrails [Rebedea et al., 2023] and constitutional AI [Bai et al., 2022] constrain model outputs. These operate on the response side. SAGE operates on the input side, gating whether the model should attempt a full response at all.

Clinical NLP and entity resolution. Medical entity recognition and linking systems such as MetaMap [Aronson and Lang, 2010] and SciSpacy [Neumann et al., 2019] resolve terms to ontology concepts. The Unified Medical Language System (UMLS) [Bodenreider, 2004] provides a comprehensive metathesaurus linking biomedical vocabularies. SAGE extends this idea by mapping entities to *operational domains* rather than ontology codes, and by using co-occurrence patterns to propagate disambiguation signals across entities within a single query.

Biomedical question answering. Benchmarks such as BioASQ [Tsatsaronis et al., 2015] and PubMedQA [Jin et al., 2019] evaluate biomedical QA systems but focus on answer correctness rather than query ambiguity. SAGE addresses the upstream problem: determining whether a query is sufficiently specified before attempting to answer it.

Query ambiguity detection. Prior work on web search query ambiguity [Song et al., 2009] classifies queries as clear, ambiguous, or broad using click-through data and query reformulations. SAGE adapts this concept to domain-specific professional queries where click data is unavailable, using structured domain knowledge instead.

3 Method

3.1 Architecture Overview

SAGE computes a composite tension score $T \in [0, 100]$ from two parallel analysis paths:

$$T = \lfloor (w_E \cdot S_E + w_N \cdot S_N) \times 100 \rceil \quad (1)$$

where $S_E \in [0, 1]$ is the entity resolution score, $S_N \in [0, 1]$ is the NLP decision tree score, and w_E, w_N are adaptive weights (Section 3.5). Default values are $w_E = 0.6$, $w_N = 0.4$.

3.2 Path 1: Entity Resolution with Contextual Inference

3.2.1 Concept Registry

We maintain a curated registry of ∼70 HCLS concepts. Each concept c has a set of aliases A_c (surface forms) and a set of domains D_c drawn from five HCLS domains: `clinical`, `regulatory`, `pharmacovigilance`, `datastandards`, and `research`. For

example, *"adverse events"* maps to four domains (clinical, pharmacovigilance, regulatory, datastandards), while *"signal detection"* maps to pharmacovigilance alone.

3.2.2 Entity Matching

Given a query q, we perform greedy longest-match extraction against all aliases, producing a set of matched entities $M = \{m_1, \ldots, m_k\}$.

3.2.3 Domain Vote Counting

For each domain d, we count how many matched entities include d:

$$V(d) = |\{m \in M : d \in D_m\}| \tag{2}$$

Votes are normalized to $[0, 1]$ by dividing by $\max_d V(d)$.

3.2.4 Anchor Identification

An *anchor* is an entity that resolves to exactly one domain: $|D_m| = 1$. Anchors provide strong directional signal. User-activated context layers are treated as additional anchors with elevated weight.

3.2.5 Confidence Propagation

For each multi-domain entity m with $|D_m| > 1$, we compute a confidence distribution $P_m(d)$ over its domains:

$$P_m(d) \propto \frac{1}{|D_m|} + \alpha \cdot \sigma \cdot \hat{V}(d) + \beta \cdot \sigma \cdot \frac{A(d)}{\sum_{d'} A(d')} \tag{3}$$

where $\hat{V}(d)$ is the normalized vote for domain d, $A(d)$ is the anchor count for domain d (including active contexts with weight 2), $\alpha = 0.3$ and $\beta = 0.4$ are boost parameters, and $\sigma = \min(1, |M|/5)$ is a *signal strength* factor that scales the vote and anchor boosts by the number of matched entities. Queries with more entities provide more reliable co-occurrence signal; a query with two entities has weaker evidence for domain inference than one with six. The distribution is normalized to sum to 1.

3.2.6 Resolution Status

Each entity is assigned a status based on its maximum confidence $P^*_m = \max_d P_m(d)$:

- **Resolved**: $|D_m| = 1$ (single domain)
- **Inferred**: $P^*_m \geq 0.6$ (contextually inferred)
- **Disambiguated**: exactly one active context matches
- **Likely**: $0.4 \leq P^*_m < 0.6$
- **Ambiguous**: multiple active contexts match, $P^*_m < 0.4$
- **Unresolved**: no active context covers the entity

3.2.7 Domain Similarity Matrix

A key insight from early evaluation is that *domain breadth is not equivalent to ambiguity.* A query spanning clinical and research (semantically close) is less ambiguous than one spanning clinical and data-standards (operationally distant). We define a symmetric similarity matrix $\text{Sim}(d_i, d_j) \in [0, 1]$ over the five domains, where values reflect semantic proximity (e.g., $\text{Sim}(\text{clinical}, \text{research}) = 0.7$, $\text{Sim}(\text{clinical}, \text{datastandards}) = 0.3$).

For each entity m, we compute the *domain spread*:

$$\text{spread}(m) = \frac{1}{\binom{|D_m|}{2}} \sum_{i<j} (1 - \text{Sim}(d_i, d_j)) \tag{4}$$

where the sum is over all pairs of domains in D_m. The spread ranges from 0 (all domains identical) to 1 (maximally distant).

3.2.8 Entity Resolution Score

The score S_E is the mean penalty across all matched entities, where penalties are scaled by both confidence and domain distance:

$$S_E = \frac{1}{|M|} \sum_{m \in M} \text{penalty}(m) \cdot \delta(m) \tag{5}$$

where $\delta(m) = 0.5 + 0.5 \cdot \text{spread}(m)$ is the distance factor. Entities spanning close domains (e.g., clinical + research, $\delta \approx 0.65$) receive substantially

lower penalties than those spanning distant domains (e.g., clinical + datastandards + pharmacovigilance, $\delta \approx 0.85$).

Base penalties range from 0.0 (resolved) to 0.8 (unresolved), with intermediate values modulated by $(1 - P_m^*)$ for inferred and likely statuses. If no entities are matched ($M = \emptyset$), $S_E = 0.5$ (a moderate default that defers to the NLP path rather than the previous conservative value of 0.85).

3.3 Path 2: NLP Decision Tree

The NLP path evaluates five structural dimensions of the query, each producing a sub-score $s_i \in [0, 1]$:

1. **Verb specificity** (25%): classifies the primary verb as specific (*calculate*, *list*), vague (*analyze*, *show*), or absent. Crucially, when a vague verb has a *specific direct object*—detected via domain-specific patterns (e.g., *"analyze the dose-response curve"*, *"check the Kaplan-Meier estimate"*)—the penalty is reduced from 0.8 to 0.35, reflecting that the object disambiguates the verb's intent.

2. **Scope markers** (25%): detects references to specific studies, datasets, time periods, filtering clauses, comparison groups (e.g., *responders vs. non-responders*), and population specifications (e.g., *intent-to-treat population*).

3. **Quantifiers** (20%): identifies numeric bounds, severity grades, temporal constraints, and trial phases.

4. **Output format** (15%): checks for explicit format specifications (*as CSV*, *in a table*).

5. **Query structure** (15%): evaluates length, compound intents, and sentence complexity.

$$S_N = \sum_{i=1}^{5} w_i \cdot s_i \tag{6}$$

3.4 Concept-Driven Insights

Rather than relying on the LLM to identify missing information, SAGE derives clarifying questions and missing-context indicators directly from the analysis:

- **Missing context**: domains referenced by matched entities but not activated by the user, with descriptions from the context definition registry.

- **Entity disambiguation questions**: for ambiguous or unresolved entities, questions listing the possible domain interpretations.

- **Structural questions**: from NLP gaps—unscoped queries prompt for study/time period, missing output format prompts for format preference, vague verbs prompt for specific actions.

These are injected into the LLM prompt as structured sections with explicit instructions to use them directly.

3.5 Adaptive Weighting

Rather than using fixed weights, SAGE adapts the entity/NLP balance based on analysis conditions:

- **No entities matched** ($|M| = 0$): $w_E = 0.3$, $w_N = 0.7$. The NLP path dominates since entity resolution has no signal.

- **Resolved but vague** ($S_E < 0.15$ and $S_N > 0.6$): $w_E = 0.35$, $w_N = 0.65$. Entities are domain-resolved but the query is structurally underspecified (e.g., *"What are the regulatory requirements?"*).

- **High NLP specificity** ($S_N < 0.3$): $w_E = 0.5$, $w_N = 0.5$. The query is well-formed; equal weighting prevents entity penalties from cross-domain terms from dominating.

- **Moderate NLP specificity** ($S_N < 0.45$): $w_E = 0.55$, $w_N = 0.45$.

- **Default**: $w_E = 0.6$, $w_N = 0.4$.

3.6 LLM Response Control

The tension score gates LLM behavior through prompt instructions:

- $T > 60$ (High): present concept-derived clarifying questions; do not attempt an answer.

- $30 < T \leq 60$ (Medium): provide a partial

answer using inferred interpretations; flag remaining ambiguities.

- $T \leq 30$ (Low): provide a complete answer using all available context.

4 Evaluation

4.1 Dataset

We evaluate SAGE on a dataset of 30 queries adapted from public biomedical QA benchmarks (BioASQ [Tsatsaronis et al., 2015], PubMedQA [Jin et al., 2019]), clinical trial registry patterns (ClinicalTrials.gov), FDA guidance document language, and pharmacovigilance reporting terminology. These represent realistic queries that HCLS professionals—clinicians, biostatisticians, regulatory affairs specialists, and safety officers—would pose to an agentic AI system.

The dataset comprises 10 high-ambiguity, 10 medium-ambiguity, and 10 low-ambiguity queries. High and medium queries have no active contexts (simulating a user who has not yet provided domain context); low-ambiguity queries have relevant contexts active. Each query is annotated with an expected ambiguity level, expected primary domain(s), active context configuration, and source attribution.

The concept registry contains ∼70 manually curated concepts covering standard HCLS terminology including hepatotoxicity, checkpoint inhibitors, Kaplan-Meier, DSUR, EudraVigilance, SUSAR, and define-XML.

4.2 Baselines

To isolate the contribution of each analysis path, we evaluate three configurations:

1. **Entity Resolution Only**: $T = \lfloor S_E \times 100 \rceil$ (no NLP analysis).
2. **NLP Structural Only**: $T = \lfloor S_N \times 100 \rceil$ (no entity resolution).
3. **SAGE Composite**: full system with adaptive weighting.

Table 1: Confusion matrix for SAGE composite (rows = expected, columns = predicted).

	High	**Medium**	**Low**
High	4	6	0
Medium	0	9	1
Low	0	4	6

Table 2: Per-class metrics for SAGE composite.

Class	**Prec.**	**Rec.**	**F1**
High	100.0%	40.0%	57.1%
Medium	47.4%	90.0%	62.1%
Low	85.7%	60.0%	70.6%
Accuracy		**63.3%**	

4.3 Results

4.3.1 Baseline Comparison

Table 3: Comparison of SAGE composite against single-path baselines.

Approach	**Acc.**	**High F1**	**Med F1**	**Low F1**
Entity Only	43.3%	46.2%	34.8%	50.0%
NLP Only	60.0%	71.4%	25.0%	75.0%
SAGE Composite	**63.3%**	57.1%	**62.1%**	70.6%

The composite system outperforms both single-path baselines in overall accuracy (Table 3). The entity-only baseline achieves only 43.3% accuracy, confirming that domain resolution alone is insufficient for ambiguity classification—it conflates multi-domain coverage with uncertainty. The NLP-only baseline achieves 60.0% accuracy with perfect high-class recall (100%) but very poor medium-class F1 (25.0%), because structural analysis cannot distinguish between queries that are vague across domains versus vague within a single domain.

The composite system's key advantage is balanced performance across all three classes: it achieves the highest medium-class F1 (62.1%) by a wide margin, demonstrating that combining domain knowledge with structural analysis captures ambiguity patterns that neither path detects alone.

4.3.2 Domain Inference

The contextual inference mechanism achieves **100% domain inference accuracy**: for all 30 queries, at least one inferred domain matched the human-annotated expected domain. The expanded concept registry and anchor propagation generalize well to real-world HCLS terminology.

4.3.3 Score Distribution

Table 4: Score distribution by expected ambiguity level.

Level	n	Min	Max	Mean	Std
High	10	46	67	57.1	7.6
Medium	10	23	56	43.7	9.4
Low	10	18	34	27.4	4.9

The three classes show clear mean separation (Table 4): high (57.1), medium (43.7), low (27.4). The high-medium overlap spans 10 score points (46–56), while the medium-low overlap spans 11 points (23–34). Notably, the domain-distance-aware penalties and adaptive weighting substantially reduced the high-class minimum from 28 (in an earlier version without these improvements) to 46, eliminating the most problematic overlap region.

4.3.4 Adaptive Weighting Analysis

The adaptive weighting mechanism activates for 12 of 30 queries:

- **Default** (18 queries): 72% accuracy.
- **High NLP specificity** (6 queries): 67% accuracy. These are well-formed cross-domain queries where equal weighting prevents entity penalties from dominating.
- **Moderate NLP specificity** (4 queries): 25% accuracy. These borderline cases remain challenging.
- **Resolved but vague** (2 queries): 50% accuracy. Queries with resolved entities but high structural vagueness.

4.3.5 Error Analysis

The 11 misclassified cases reveal three patterns:

High predicted as medium (6 cases). Queries like *"Analyze the biomarker data"* and *"Review the submission documents"* score in the 48–57 range. The entity resolution path assigns moderate penalties because the entities span only 2–3 domains with moderate distance, while the NLP path correctly identifies structural vagueness (scores 68–84). The default 60/40 weighting gives insufficient influence to the NLP signal. These cases would benefit from a "vague with moderate entity ambiguity" adaptive weight condition.

Low predicted as medium (4 cases). Well-specified queries such as *"Calculate the Kaplan-Meier estimate of progression-free survival for the intent-to-treat population in study ONCO-2024-001"* score 31 despite having a specific statistical method, endpoint, population, and study. The entity resolution path assigns residual penalties because terms like *"progression-free survival"* and *"Kaplan-Meier"* each span two domains (clinical + research). The domain-distance penalty reduces but does not eliminate this: $\text{Sim}(\text{clinical}, \text{research}) = 0.7$ yields $\delta = 0.65$, still producing a non-trivial penalty.

Medium predicted as low (1 case). *"Compare overall survival between the two treatment arms in the Phase III trial"* scores 23 because both paths rate it as well-specified—specific verb, specific endpoint, comparison groups, trial phase. The medium label reflects the missing study identifier, which neither path penalizes sufficiently.

5 Discussion

5.1 Pre-Flight vs. Post-Hoc Ambiguity Detection

Most existing approaches to LLM safety in HCLS operate on the output: filtering harmful content, checking factual consistency, or measuring generation uncertainty. SAGE operates on the *input*, measuring ambiguity before the model processes the query. This has several advantages: (1) it prevents the model from committing to an interpretation before ambiguity is surfaced, (2) it runs in constant time with no LLM inference cost, and (3) it produces deterministic, auditable scores suitable for regulated environments.

5.2 Domain Distance as a First-Class Signal

The domain similarity matrix addresses a fundamental conflation in naive multi-domain penalty schemes: treating all domain combinations as equally ambiguous. Our results show that queries spanning clinical and research (similarity 0.7) are systematically less ambiguous than those spanning clinical and data-standards (similarity 0.3). The distance factor δ reduces low-class misclassification by lowering penalties for semantically adjacent domain pairs, while preserving high penalties for genuinely distant combinations.

5.3 Value of Adaptive Weighting

The baseline comparison demonstrates that neither analysis path alone is sufficient. Entity resolution excels at domain identification (100% accuracy) but poorly classifies ambiguity levels (43.3%) because it conflates domain breadth with uncertainty. NLP structural analysis captures query specificity but cannot distinguish domain-specific from domain-general vagueness (medium F1 of only 25.0%). The composite system with adaptive weighting achieves the best balance, particularly for the medium-ambiguity class where both paths contribute complementary signal.

5.4 Concept-Driven vs. LLM-Generated Clarification

A key design decision is deriving clarifying questions from concept analysis rather than asking the LLM to generate them. LLM-generated questions may be plausible but disconnected from the actual ambiguity structure of the query. Concept-driven questions are grounded in the specific entities that failed to resolve and the specific domains that are missing—they are traceable, reproducible, and auditable.

5.5 Limitations

Registry coverage. The concept registry contains ∼70 manually curated concepts. While the expanded registry achieves 100% domain inference on real-world queries, terms outside the registry receive a moderate default score ($S_E = 0.5$), deferring to the NLP path. This is less conservative than the original design ($S_E = 0.85$) but may still produce false positives for novel but unambiguous terminology.

Domain similarity calibration. The similarity matrix values are manually assigned based on operational proximity. Empirical calibration using co-occurrence statistics from HCLS corpora could improve these values.

Evaluation scale. The evaluation dataset contains 30 queries with single-annotator labels. A larger dataset with multiple human annotators and inter-annotator agreement metrics (e.g., Cohen's κ) would strengthen the validity of the accuracy metrics.

Downstream impact. The evaluation measures classification accuracy of the tension score but does not assess whether injecting the score into LLM prompts actually improves response quality. A user study with domain experts comparing LLM responses with and without SAGE would quantify the end-to-end benefit.

Binary context model. Context layers are binary (active/inactive). A graded context model—where partial context provides partial disambiguation—could improve scoring granularity.

6 Future Work

Learned tension model. Fine-tuning a lightweight classifier (e.g., a small transformer) on a larger labeled HCLS query dataset could replace or augment the rule-based NLP tree, capturing semantic patterns that pattern matching misses.

Ontology-backed registry. Connecting the concept registry to standard biomedical ontologies (SNOMED CT, MedDRA, LOINC, RxNorm) would dramatically expand coverage and enable hierarchical reasoning about concept relationships.

Agentic feedback loop. Allowing the agent to autonomously retrieve missing context from connected knowledge sources—data catalogs, ontology services, regulatory databases—to reduce its own tension score before responding.

Tension-aware routing. Using the score to route

queries to specialized agent personas: high-tension queries to a disambiguation agent, medium-tension to a domain-specific expert, and low-tension directly to execution.

Multi-turn tension tracking. Extending the framework to track tension across conversation turns, measuring how each user response or context addition reduces ambiguity over time.

7 Conclusion

We presented SAGE, a pre-flight semantic ambiguity detection framework for LLM agents in healthcare and life sciences. By combining entity resolution with contextual domain inference, domain-distance-aware penalties, context-aware NLP analysis, and adaptive weighting, the system quantifies query ambiguity before any LLM inference occurs. On a dataset of 30 real-world HCLS queries, SAGE achieves 63.3% classification accuracy and 100% domain inference accuracy, outperforming entity-only (43.3%) and NLP-only (60.0%) baselines. The domain similarity matrix addresses the conflation of multi-domain coverage with genuine ambiguity, and the adaptive weighting mechanism ensures that structurally specific cross-domain queries are not penalized as ambiguous. SAGE demonstrates that measuring ambiguity at query time is both feasible and valuable for improving the safety and precision of agentic systems in regulated industries.

References

A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

Y. Bai, S. Kadavath, S. Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. PubMedQA: A dataset for biomedical research question answering. In *EMNLP*, pages 2567–2577, 2019.

S. Kadavath, T. Conerly, A. Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *BioNLP Workshop*, pages 319–327, 2019.

T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *WWW*, pages 1169–1170, 2009.

G. Tsatsaronis, G. Balikas, P. Malakasiotis, et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015.