EHRAgent Reproduction and Extension

PETER HO, University of Texas, Austin, USA

Even though EHRAgent based on gpt4 was proven to have better result than other solution by the work this report is based on, There are still many challenges in reproduction and optimization.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: EHR, Agent, LLM, Azure, OpenAI, Python, Debug, SQL, MIMIC

ACM Reference Format:

Peter Ho. 2024. EHRAgent Reproduction and Extension. 1, 1 (December 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

According to https://www.thebalancemoney.com/causes-of-rising-healthcare-costs-4064878, cost of healthcare has significantly increased over past decades and years, which can be a risk for health and well being of Americans impacting every age group. I believe there are many factors with many different solutions. One of the many is leveraging technology to help medical professionals to focus their time and energy on patients and/or research but not simple data retrieval or analysis.

LLM (Large Language Model) has gained signficant popularity over the past years. It can capture semantics of text, memorize past conversations,

2 Related Work

This a reproduction and alternation of part of the work - ["EHRAgent: Code Empowers Large Language Models for Complex Tabular Reasoning on Electronic Health Records"]https://arxiv.org/abs/2401.07128. EHRAgent is an LLM agent empowered with a code interface, to autonomously generate and execute code for complex clinical tasks within electronic health records (EHRs). The original project page is available at [this link](https://wshi83.github.io/EHR-Agent-page/).

@articleshi2024ehragent, title=Ehragent: Code empowers large language models for complex tabular reasoning on electronic health records, author=Shi, Wenqi and Xu, Ran and Zhuang, Yuchen and Yu, Yue and Zhang, Jieyu and Wu, Hang and Zhu, Yuanda and Ho, Joyce and Yang, Carl and Wang, May D, journal=EMNLP, year=2024

Text to SQL generation is another related project that the former utilizes its evelation logic with slight modification. Information of the project can be found at https://github.com/wangpinggl/TREQS/blob/master/README.md

@inproceedingswang2020text, title=Text-to-SQL Generation for Question Answering on Electronic Medical Records, author=Wang, Ping and Shi, Tian and Reddy, Chandan K, booktitle=Proceedings of The Web Conference 2020, pages=350–361, year=2020

Author's Contact Information: Peter Ho, peter-ho@live.com, University of Texas, Austin, Austin, Texas, USA.

Manuscript submitted to ACM

3 Methodology

3.1 EHR Agent Execution Flow

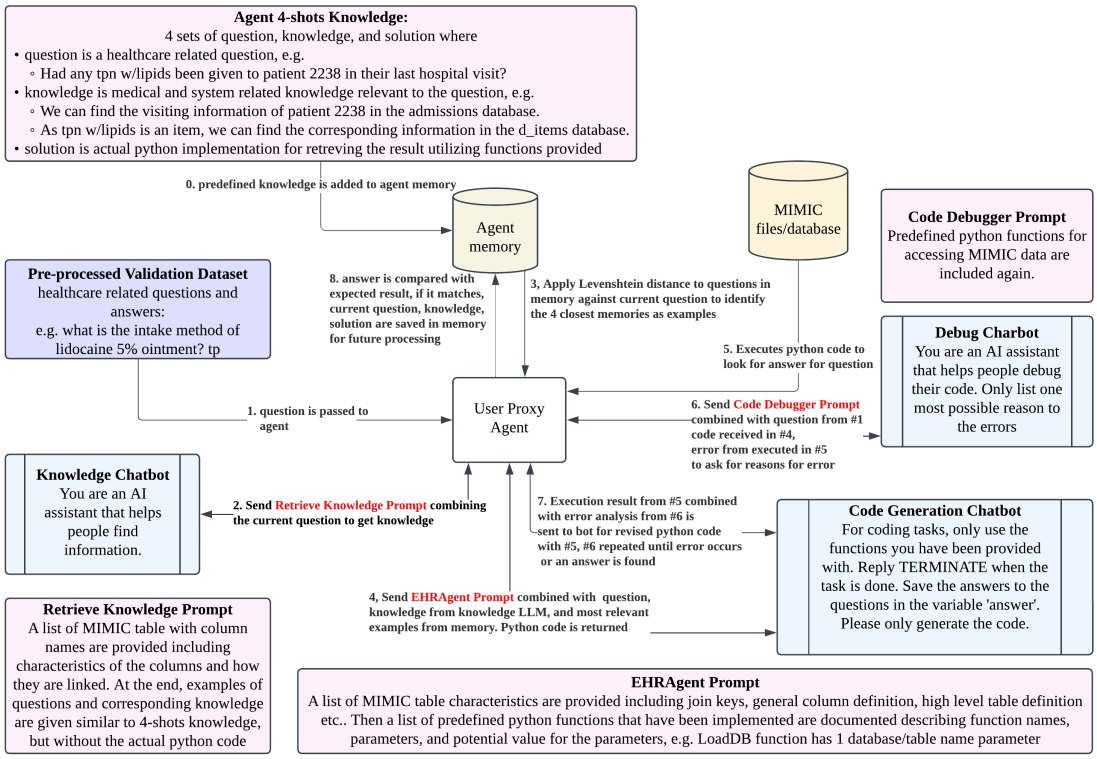

Fig. 1. EHR Agent execution flow

3.2 Execution Steps

(1) Predefined knowledge is intialized as memory, including question, knowledge, and associated solution.
(2) A question from validation data set is passed into UserProxyAgent
(3) During initialization process of main code generation chat, UserProxyAgent sends question to Knowledge chatbot to retrieve knowledge regarding the question applying chain of thought. Few shot learning is also applied as the prompt to the knowledge chatbot includes examples of questions and their corresponding knowledge.
(4) Retrieve 4 most similar examples by comparing current question with questions in memory according to levenshtein distance, applying few shot learning.
(5) Invoke code generation chatbot with table definition, function definition, knowledge retrieved from knowledge chatbot, examples from memory.

(6) Execute code retrieved from bot and identify answer or errors. IF errors were encountered, capture errors and provide potential hints in resolving them.
(7) When error was thrown in the last step, functions documentation, code executed and execution error are sent to debugger chatbot to identify potential fixes.
(8) Original error message and suggestions returned by debugger chatbot are packaged as response back to code generation cahtbot for it to review and regenerate a better response. As chat with code generation chatbot continues until error occured related to exceeding max token size or termination message was identified.
(9) The response is then compared with predefined answer, if the answer was correct, the question, knowledge and solution are all stored in memory for future reference.

3.3 Original functions provided for chatbot

(1) `Calculate`: passes argument to wolfram alpha for mathematical calculation.
(2) `LoadDB`: loads a MIMIC csv file into memory
(3) `FilterDB`: filters data returned froom LoadDB with predefined common operators similar to SQL WHERE clause, but with a little more functionality including min and max, which selects the row equals to the minimum or maximum of the whole file.
(4) `GetValue`: retrieves a string based on a comma delimited list of column names provided in table definition, with some operator supported including sum.
(5) `SQLInterpreter`: executes a given sql against sqlite3 database.
(6) `Calendar`: evaluates the date after the duration of time based on sqlite database DATE_ADD function applied to CURRENT_TIMESTAMP

3.4 Issues regarding original work

(1) `Cost`: The table cost was referenced in multiple places of the original implementation including examples, validaton questons and answers, but in MIMIC-III v1.4, cost is not a file available for download, nor in the demo dataset. Any question related to cost can't be answered, and solution based on the example referring to cost would result in a table not found error, which can't be recovered.
(2) `Python Package dependency`: In requirements.txt, autogen version 1.0.16 was referred, but 0.3.2 is the latest when this report was written. 0.3.3 was released afterwards, but it's still quite far from 1.0.16.
(3) `Sqlite3 calendar calculation`: After setting up sqlite3 by importing MIMIC-III v1.4 csv files, sqlite3 database DATE_ADD function doesn't use current date, but used 2000-01-01, so with input of -1 day, 1999-12-31 was returned.
(4) `LoadDB requirement`: Since LoadDB reads the whole table into memory, for larger table, out of memory exception was thrown. Due to re-reading the same csv or csv.gz files for every LoadDB command for various questions, disk IO can cause unnecessary delays when reading against the same data.
(5) `Validation set issue`: When trying to identify reasons for incorrect responses, SQL embedded in the validation data set was used to identify and verify given answers. There are answers that are incorrect based on data retrieved from MIMIC-III v1.4.
(6) `Execution challenge`: Since some code generated by chatbot can be quite inefficient, there are times that it takes more than hours to run, and sometimes it can be hard to tell whether how long it will take for the process to terminate.

3.5 Changes made corresponding to issues above

based on In the MIMIC-III database 24 questions and answers involved costs.

Terminating the whole evaluation process was needed

In addition to specifying the *template style* to be used in formatting your work, there are a number of *template parameters* which modify some part of the applied template style. A complete list of these parameters can be found in the *LaTeX User's Guide*.

Frequently-used parameters, or combinations of parameters, include:

- `anonymous,review`: Suitable for a "double-anonymous" conference submission. Anonymizes the work and includes line numbers. Use with the command to print the submission's unique ID on each page of the work.
- `authorversion`: Produces a version of the work suitable for posting by the author.
- `screen`: Produces colored hyperlinks.

This document uses the following string as the first command in the source file:

```
\documentclass[manuscript,screen,review]{acmart}
```

4 Modifications

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions, and the use of the `\vspace` command to manually adjust the vertical spacing between elements of your work — is not allowed.

Your document will be returned to you for revision if modifications are discovered.

5 Typefaces

The "`acmart`" document class requires the use of the "Libertine" typeface family. Your TeX installation should include this set of packages. Please do not substitute other typefaces. The "`lmodern`" and "`ltimes`" packages should not be used, as they will override the built-in typeface families.

6 Title Information

The title of your work should use capital letters appropriately - https://capitalizemytitle.com/ has useful rules for capitalization. Use the `title` command to define the title of your work. If your work has a subtitle, define it with the `subtitle` command. Do not insert line breaks in your title.

If your title is lengthy, you must define a short version to be used in the page headers, to prevent overlapping text. The `title` command has a "short title" parameter:

```
\title[short title]{full title}
```

7 Authors and Affiliations

Each author must be defined separately for accurate metadata identification. As an exception, multiple authors may share one affiliation. Authors' names should not be abbreviated; use full first names wherever possible. Include authors' e-mail addresses whenever possible.

Grouping authors' names or e-mail addresses, or providing an "e-mail alias," as shown below, is not acceptable:

```
\author{Brooke Aster, David Mehldau}
```

```
\email{dave,judy,steve@university.edu}
\email{firstname.lastname@phillips.org}
```

The `authornote` and `authornotemark` commands allow a note to apply to multiple authors — for example, if the first two authors of an article contributed equally to the work.

If your author list is lengthy, you must define a shortened version of the list of authors to be used in the page headers, to prevent overlapping text. The following command should be placed just after the last `\author{}` definition:

```
\renewcommand{\shortauthors}{McCartney, et al.}
```

Omitting this command will force the use of a concatenated list of all of the authors' names, which may result in overlapping text in the page headers.

The article template's documentation, available at https://www.acm.org/publications/proceedings-template, has a complete explanation of these commands and tips for their effective use.

Note that authors' addresses are mandatory for journal articles.

8 Rights Information

Authors of any work published by ACM will need to complete a rights form. Depending on the kind of work, and the rights management choice made by the author, this may be copyright transfer, permission, license, or an OA (open access) agreement.

Regardless of the rights management choice, the author will receive a copy of the completed rights form once it has been submitted. This form contains LaTeX commands that must be copied into the source document. When the document source is compiled, these commands and their parameters add formatted text to several areas of the final document:

- the "ACM Reference Format" text on the first page.
- the "rights management" text on the first page.
- the conference information in the page header(s).

Rights information is unique to the work; if you are preparing several works for an event, make sure to use the correct set of commands with each of the works.

The ACM Reference Format text is required for all articles over one page in length, and is optional for one-page articles (abstracts).

9 CCS Concepts and User-Defined Keywords

Two elements of the "acmart" document class provide powerful taxonomic tools for you to help readers find your work in an online search.

The ACM Computing Classification System — https://www.acm.org/publications/class-2012 — is a set of classifiers and concepts that describe the computing discipline. Authors can select entries from this classification system, via https://dl.acm.org/ccs/ccs.cfm, and generate the commands to be included in the LaTeX source.

User-defined keywords are a comma-separated list of words and phrases of the authors' choosing, providing a more flexible way of describing the research being presented.

CCS concepts and user-defined keywords are required for for all articles over two pages in length, and are optional for one- and two-page articles (or abstracts).

Table 1. Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Table 2. Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

10 Sectioning Commands

Your work should use standard LaTeX sectioning commands: `section`, `subsection`, `subsubsection`, and `paragraph`. They should be numbered; do not remove the numbering from the commands.

Simulating a sectioning command by setting the first word or words of a paragraph in boldface or italicized text is **not allowed.**

11 Tables

The "`acmart`" document class includes the "`booktabs`" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User's Guide.*

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

Always use midrule to separate table header rows from data rows, and use it only for this purpose. This enables assistive technologies to recognise table headers and support their users in navigating tables more easily.

12 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

12.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

12.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

13 Figures

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Your figures should contain a caption which describes the figure to the reader.

Figure captions are placed *below* the figure.

Every figure should also have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

A figure description must be unformatted plain text less than 2000 characters long (including spaces). **Figure descriptions should not repeat the figure caption – their purpose is to capture important information that is not already provided in the caption or the main text of the paper.** For figures that convey important and complex new information, a short text description may not be adequate. More complex alternative descriptions can be placed in an appendix and referenced in a short figure description. For example, provide a data table capturing the information in

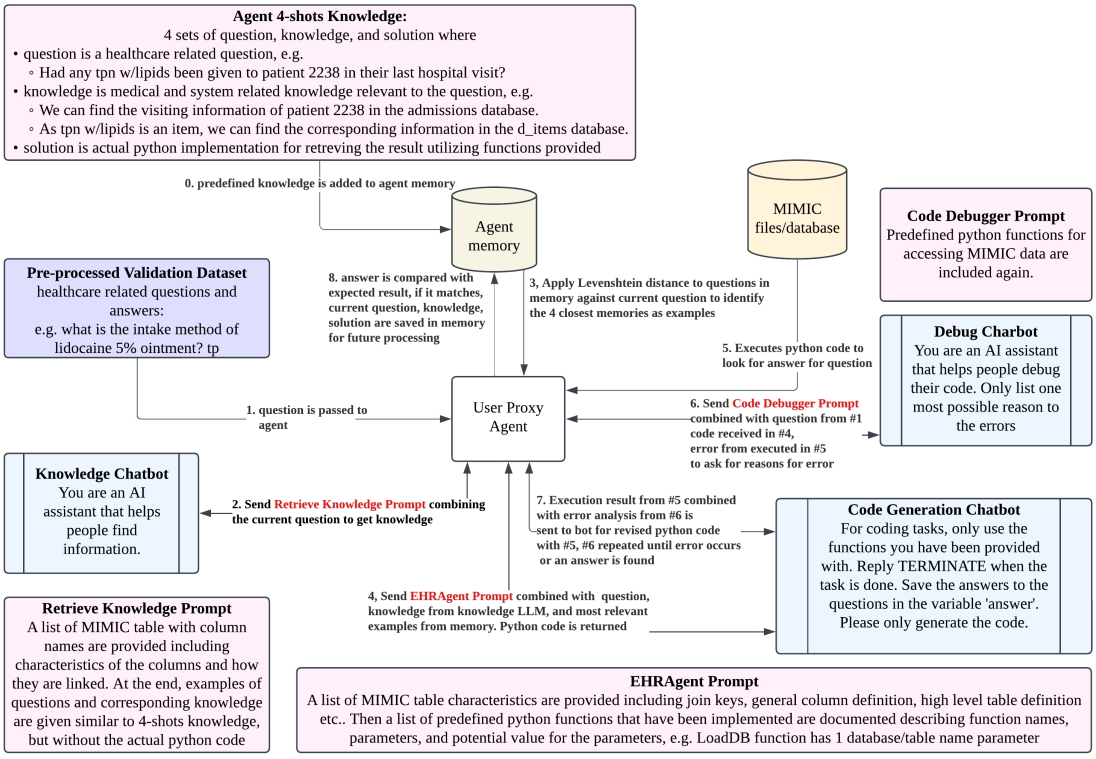

Fig. 2. 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

a bar chart, or a structured list representing a graph. For additional information regarding how best to write figure descriptions and why doing this is so important, please see https://www.acm.org/publications/taps/describing-figures/.

13.1 The "Teaser Figure"

A "teaser figure" is an image, or set of images in one figure, that are placed after all author and affiliation information, and before the body of the article, spanning the page. If you wish to have such a figure in your article, place the command immediately before the `\maketitle` command:

```
\begin{teaserfigure}
  \includegraphics[width=\textwidth]{sampleteaser}
  \caption{figure caption}
  \Description{figure description}
\end{teaserfigure}
```

14 Citations and Bibliographies

The use of BibTEX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTEX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before the command "`\begin{document}`") of your LATEX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only output the series if the volume number is given [**?**] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], a couple of articles in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**? ?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**], a couple of preprints [**? ?**], a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**? ? ?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], 'YYYYb'-test for prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Boris / Barbara Beeton: multi-volume works as books [**?**] and [**?**]. A couple of citations with DOIs: [**? ?**]. Online citations: [**? ? ?**]. Artifacts: [**?**] and [**?**].

15 Acknowledgments

Identification of funding sources and other support, and thanks to individuals and groups that assisted in the research and the preparation of the work should be included in an acknowledgment section, which is placed just before the reference section in your document.

This section has a special environment:

```
\begin{acks}
...
\end{acks}
```

so that the information contained therein can be more easily collected during the article metadata extraction phase, and to ensure consistency in the spelling of the section heading.

Authors should not prepare this section as a numbered or unnumbered `\section`; please use the "`acks`" environment.

16 Appendices

If your work needs an appendix, add it before the "`\end{document}`" command at the conclusion of your source document.

Start the appendix with the "`appendix`" command:

```
\appendix
```

and note that in the appendix, sections are lettered, not numbered. This document has two appendices, demonstrating the section and subsection identification method.

17 Multi-language papers

Papers may be written in languages other than English or include titles, subtitles, keywords and abstracts in different languages (as a rule, a paper in a language other than English should include an English title and an English abstract). Use `language=...` for every language used in the paper. The last language indicated is the main language of the paper. For example, a French paper with additional titles and abstracts in English and German may start with the following command

```
\documentclass[sigconf, language=english, language=german,
               language=french]{acmart}
```

The title, subtitle, keywords and abstract will be typeset in the main language of the paper. The commands `\translatedXXX`, `XXX` begin title, subtitle and keywords, can be used to set these elements in the other languages. The environment `translatedabstract` is used to set the translation of the abstract. These commands and environment have a mandatory first argument: the language of the second argument. See `sample-sigconf-i13n.tex` file for examples of their usage.

18 SIGCHI Extended Abstracts

The "`sigchi-a`" template style (available only in LaTeX and not in Word) produces a landscape-orientation formatted article, with a wide left margin. Three environments are available for use with the "`sigchi-a`" template style, and produce formatted output in the margin:

`sidebar`: Place formatted text in the margin.

`marginfigure`: Place a figure in the margin.

`margintable`: Place a table in the margin.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

A Research Methods

A.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

A.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

B Online Resources

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.

Received 1 February 2007; revised 12 March 2009; accepted 5 June 2009